Jan Engh

Linguistic Normalisation in Language Industry

Some Normative and Descriptive Aspects of Dictionary Development*

Abstract

For commercial software with natural language functions, a high coverage is required. This implies that only extensive lexica and complete morphologies are of interest to the language industry. For many languages, lexical and morphological information has to be collected from traditional lexicographic files and printed dictionaries. However, such material may not provide adequate information - even if trivial defects such as misprintings and editorial inconsequences are left out of account. The present paper is an attempt to point out how basic information on any language drawn from traditional sources has to be controlled for normative correctness and descriptive adequacy, and how normalisation can only be defined relative to a given application. The presentation is based on the author's experience, and the examples are all Norwegian. Still, it is assumed to be of general nature, highlighting some very fundamental aspects of computational linguistics which are often neglected in practice, which "everybody" is aware of all the same, but very few - if anyone - has bothered to discuss in writing.

Introduction

A system reaching great analytic depths is of little use as long as it is based on a lexicon of one or two hundred words - both for scientific and, of course, for commercial reasons. This was the motive for a wide front linguistic development project for all the major languages of Western Europe, launched by IBM in the eighties. The result was a series of natural language functions with high coverage. One necessary condition for the development of such functions is an extensive lexicon and a complete morphology which can be used as the database from which words from one language are selected for specific types of software.

When IBM Norway started its activities within computational linguistics in 1984, no adequate machine-readable lexicographic material was

* I am grateful to Jørn-Otto Akø, Henning Bergenholtz, Dag Gundersen, Diana Santos, and Tor Ulset for having read and commented on a draft version of this paper, which was written before I resigned from IBM and the company decided to close down its linguistic development for Norwegian.

available for Norwegian, only old-fashioned files and printed dictionaries. Nor was any sufficient text corpus for Norwegian accessible which could have served as a basis for the compilation of a lexicon and a corresponding morphology. In practice, everything had to be made from scratch.

The point of departure was an internal frequency list based on a corpus of Norwegian business correspondence which had been established in the United States by American linguists and software engineers. This list was then screened and completed by the lexicographers: Lemma forms were entered, and missing inflected wordforms were supplied - indirectly drawing upon all other available resources.

The result of the work carried out since 1984, is a database containing an extensive documentation of the lexicon and the morphology of the two Norwegian written standards, Bokmål and Nynorsk - probably **the** most extensive[1] At present, it contains approximately 130,000 lemmata and 655 inflectional paradigms for Norwegian Bokmål and 111,000 lemmata and 576 paradigms for Norwegian Nynorsk.

> This makes it necessary to make a brief account of the rather peculiar linguistic situation in Norway. Spoken Norwegian is one language - with numerous different dialects. However, there are two ways of writing it, Bokmål and Nynorsk, each of them close to different groups of dialects. Although these variants are different with respect to a number of syntactic/stylistic features, in itself enough to legitimate the distinc-tion between them as two separate written languages, they are usually referred to as "målformer", i.e. written standards. The most important differences between them are of a lexical and, above all, of morphological nature. But the situation is more complicated still. Both Norwegian standards incorporate stems and inflected forms with different normative status: "Main orthographic variants", a subset compulsory for use in public sector and school textbooks, and "optional orthographic va-riants", encompassing only those additional forms which the students are allowed to use at school - but which are not authorised in textbooks. Typically, "optional variants" are either wordforms pertaining to an older state of the written standard or special for a particular group of dialects, or to the other written standard, Nynorsk or Bokmål respectively. This complexity accounts for quite a few of the paradigms mentioned above.

The development of a linguistic database has been a pioneering work - not only from a quantitative, but also from a qualitative point of view, as

---

[1] A presentation of the lexicographic activities carried out at IBM Norway can be found in Engh 1991 and 1992.

the objective was to establish a database characterised by both descriptive adequacy and normative correctness: Even though it is possible for the native language user to imagine "full" paradigms also including wordforms which are unacceptable or simply ungrammatical, such forms were not to be included. On the other hand, the content of the database was to be correct in the sense of 'conform to the official standard of Norwegian'.

The Norwegian Language Council is the official normative authority for both Bokmål and Nynorsk. In principle, no single system developer or lexicographer can set a linguistic standard on his own. Thus, a natural part of the process of controlling the quality of the database content was to check the result of the "quantitative" collection of words against printed dictionaries reflecting the official standard for written Norwegian as laid down by the Council.

Although printed dictionaries represent a valuable source of data, especially for the selection of entry words, they turned out to be both unreliable and incomplete. In this context, it is important to keep in mind that technical defaults are not at issue: Mere misprintings, varying editorial practice when using codes and abbreviations in particular fields of the definition, exceptions from general rules taken for granted, omissions due to either a source - or target language bias, etc. Such defects are interesting in connection with automatic parsing of type setting versions of dictionaries as a step in the process of converting them to true databases. From a purely linguistic point of view, however, they are trivial. In practice, they have to be "cleaned" from **any** printed dictionary, and will not be discussed here. The subject of the present paper is the purely linguistic defects, those which need a deliberate normative or descriptive action on the part of the computational lexicographer. Such imperfections may be called the "gray zones" of the written language.

Due to its particular situation, Norwegian is predisposed to exhibit most normalisation problems that one might expect to encounter when working with any European language: Given the complexity of the situation and the rapid changes of the orthography during this century, the part of the language which has been normalised is probably smaller than in the case of the neighbour languages. The frequent changes also make it difficult to make the official standards known to the public in an adequate manner. Thus, an account of the particular experience of creating a linguistic database for Norwegian should be of general interest.

Normative Correctness

In principle, correct Norwegian is Norwegian conform to the decisions made by the Norwegian Language Council. For obvious reasons, the database had to contain the correct forms of each word. As a consequence, every inconsistency found in the printed material purporting to reflect the official standard or any extrapolation made on the basis of such material had, in principle, to be sanctioned by the Council.

No printed dictionary reflected the more **recent decisions on orthography** taken by the Council. Thus, it was necessary to check all data

- against annual reports etc. from the Council as far as old decisions were concerned
- directly (via mail or over telephone) with the staff of the Council for decisions made since the last annual report

However, the fundamental problem under such circumstances will always be to know what is incorrect, and, consequently, what data that ought to be checked.

Interpreting the current standard as it is actually represented in dictionaries and other relevant publications also inevitably implies certain problems. Obscure points, substantial editorial inconsequences etc. are revealed. At an earlier stage, when both main orthographic variants and optional variants were still included in our database[2], one problem of interpretation, for instance, occurred in the case of compound words: What could one correctly infer about "legal" optional orthographic variants of compounds containing a constituent with one secondary form or more? E.g. the verbs VINNE 'win' and OVERVINNE 'defeat', which consists of the preposition OVER and the verb VINNE (optional orthographic variants in brackets):

infinitive	*past form*
VINNE	*vant/[vann]*
OVERVINNE	*overvant/[*overvann]*

The only correct past form of OVERVINNE is "overvant", although one could have expected that a form "overvann" would also be legal. In this case, scanty dictionary entries together with the dispersion of other rele-

---

[2] In the beginning, also "optional orthographic variants" for Bokmål were included. In our last update, there are no more "optional variants" - mainly for reasons which have to do with economy: To facilitate the possible inclusion of more lemmata which are really different, and in response to a demand from the market.

vant information, rules and exception to rules, represented a practical problem.

There is a lot of words which have **not** (yet) been **formally standardised**, especially words of recent foreign origin. In such cases, a native user generally has rather vague intuitions about the inflection. What is, for instance, the plural definite form of the English loanwords MILE and ROYALTY? On request, The Norwegian Language Council decided for the following paradigm:

indefinite singular	*definite singular*	*indefinite plural*	*definite plural*
MILE	*milen*	*miles*	*milene*
ROYALTY	*royaltyen*	*royalties/royaltyer*	*royaltyene*

But not only new loanwords represent a problem. Also **unclear standardisation of words pertaining to the core vocabulary** of Norwegian was discovered in a number of cases. These had not yet been identified, let alone put in any dictionary, before the computer demanded exact information.

What is, for instance, the supine form of the verb BRISTE 'burst'? *bristet,* the regular, "new" supine form of the verb, or *brustet,* close to the relict form, the adjective BRUSTEN, with its specialised sense of 'broken, dimmed'? The Norwegian Language Council decided that *bristet* is the correct supine form.

And what about the present participles of the verb BE/BEDE 'ask' in Nynorsk? *beande, bedande* or both?

> For those who do not belong to the happy few with a working knowledge of Norwegian: Verbs such as BE and BEDE have identical meaning, and are usually referred to as short and long versions of the same verb. There is a good historical reason for that, and from a semantic point of view, there is no difference between the two versions. (Although the verbs do have slightly different stylistic values.)

As can be seen in the paradigm below, there may be a number of "parallel" forms in each category. In some cases, e.g. 'infinitive' and 'present', these are variants with and without a *d*. However, *bedande* with a *d* is the unique present participle form.

infinitive	*be, bede, beda*
present	*ber, bed*
past	*bad*
supine	*bedd, bedt, bede, bedi*
past participle (singular neuter)	*bedt, bedt, bede, bedi*

past participle (singular masc./fem.)	*bedd, beden*
past participle (plural)	*bedde, bedne*
present participle	*bedande*
present participle mediopassive	*bedandes*
imperative	*be, bed*
infinitive mediopassive	*bedast*

Inflection of the verb BE/BEDE

The corresponding verbal noun is a similar case: BEDING is the only official form. A noun *BEING is inexistent, according to the official standard for Nynorsk.

At present, we consider that all our material is correct, in the sense that it does not contain wordforms that are considered to be incorrect according to the official standard for Norwegian.

Descriptive Adequacy

Let us now leave the normative domain, properly speaking, and take a look at what we may classify as descriptive phenomena. In practice, this means the question whether a given lemma has a complete paradigm or not. Each part of speech typically encompasses certain categories for which it is difficult to decide - for semantic reasons - whether they are complete or not in the case of a given lemma.

One clear example is the set of **participle forms inflected for agreement** when used in attribution. The set of transitive verbs and the set of verbs with past participle attributive forms are not identical. Eg. the transitive verb REKKE, as in *rekke en handa* 'offer somebody one's hand', which cannot appear in attributive position: **ei rukket hand*. In contrast, the intransitive verb GULNE '(turn) yellow' may very well occur in a similar phrase, cf. et gulnet blad 'a yellow leaf'. In practice, it is necessary to decide about the properties of each verb individually, and, as may be easily inferred, it is practically impossible to keep clear of normative considerations. One complicating factor is the negative attitude towards attributive use of past participles from a stylistic point of view.

Often, a question of norm also arises in a direct way, since there may be some confusion as to one or more particular attributive forms. For instance, very few native speakers of Norwegian have a firm intuition as to what are the correct attributive inflected forms of the past participles of BRISTE 'break' (cf. the discussion above) and SPREKKE 'split': *sprekt, sprukket*, or *sprukken*. The correct sets of wordforms are *bristet*

(neuter), bristet (feminine and masculine) and either *bristete* or *bristede* (plural) and *sprukket, sprukket*, and *sprukne,* respectively.

Number of nouns can be difficult. As for plural, one may regard the set of nouns as representing a continuum, whose extremes are considered by most native language users as having plural forms or not. FOT and GODFOT both represent such extremes. On the other hand, plurals such as *seneskjedebetennelser* and *bomuller* have some acceptance in the limited linguistic contexts of professional slang (doctors, textile engineers, etc.):

lemma form, singular	*plural*
FOT 'foot'	*føtter*
SENESKJEDEBETENNELSE 'tenosynovitis'	*(?)seneskjedebe tennelser*
FEBER 'fever'	*(?)febre*
BOMULL 'cotton'	*(?)bomuller*
GODHET 'goodness'	*?godheter*
MAT 'food'	*?mater*
LYKKE 'happiness'	*?lykker*
RØD 'red'	**røder*
FORDØYELSE 'digestion'	**fordøyelser*
GLEMMEBOK literally 'book of forgetting'[3]	**glemmebøker*
GODFOT literally 'good foot'[4]	**godføtter*

And, complementary, of course, a continuum of nouns displaying singular forms or not may be constructed:

lemma form or plural	*singular*
"dager" 'days'	*dag*
PENGER 'money'	*(?)penge*
BESTEFORELDRE 'grand-parents'	*(?)besteforelder*
PRIMATER 'primates'	*?primat*
INNVOLLER 'intestines'	*?innvoll*
BOMPENGER 'toll'	**bompenge*
OPPTØYER 'riots'	**opptøy*
HVETEBRØDSDAGER 'honeymoon'	**hvetebrødsdag*

---

3 A part of the idiom GÅ I GLEMMEBOKA 'sink into oblivion'.

4 I.e. 'healthy foot'. A part of such idiomatic expressions as SKYTE MÅL MED GODFOTEN 'score a goal with the best of one's feet', i.e. 'score a goal, being in a terrific form'.

dag is a perfectly natural singular form corresponding to the plural *dager* (of DAG 'day'), while **hvetebrødsdag* (literally 'days of wheat bread') is clearly ungrammatical. *besteforelder* is a neologism which is more or less accepted today, while *primat* and *innvoll* are only considered to be grammatical by few native language users.

A third example is the **comparison** of adjectives. Between

SNILL 'kind'	*snillere*	*snillest*

and

FAGLIG 'professional', 'technical'	**fagligere*	**fagligst*

there is a multitude of adjectives whose comparison is more or less debatable.[5] For instance:

SVART 'black'	*(?)svartere*	*(?)svartest*
HIMMELBLÅ 'sky-blue'	*?himmelblåere*	*?himmelblåest*
KOMPLEKS 'complex'	*?kompleksere*	*?kompleksest*

There are also compounds where the first constituent already indicates 'the highest degree of' the property in question, thus blocking the possibility of comparison on semantic grounds.

HELSVART '(completely) black'	*?helsvartere*	*?helsvartest*
KJEMPEFLOTT 'excellent'	*?kjempeflottere*	*?kjempeflottest*
SMELLFEIT 'very fat'	*?smellfeitere*	*?smellfeitest*

However, completeness is not only a semantic problem This is shown by the number of deverbal adjectives with participle form for which there are no regular comparison forms, e.g. GLITRENDE 'glittering; brilliant', HVITSKIMRENDE 'shimmering of white', and GRØNNFARGET '(dyed) green', and also, for instance, by the genitive forms of the adjective. It is not at all clear whether certain adjectives may appear in genitive singular neuter positive, and genitive superlative strong form singular is a straightforward impossibility.

RIK 'wealthy'	*riks*	*rikes*	*?rikts*	*rikeres*	**rikests*	*rikestes*

Under all circumstances, we are here dealing with clearly marginal inflectional forms, but the cause underlying this defect, must have to do with phonotactic/graphotactic properties.[6]

---

[5] Here, I do not consider comparison by means of *mer,* 'more', and *mest,* 'most'. Periphrastic comparison is not relevant in the present context.

[6] Cf. a non-typical adjective such as NØYE 'thorough' (without the -T suffix) denoting 'neuter singular', which in fact does have a form genitive singular neuter positive, *nøyes*.

It will always be a matter of discretion whether one should include forms such as ?*rikts* and **rikests* above. Traditionally, the lexicographer simply does not take a stand at all in such cases, cf. the handling of RIK in the major monolingual dictionary for Norwegian Bokmål:

> **rik** al (norr *rikr* 'mektig') som eier mye, som har rikelig av noe *en r- og mektig mann* / *landet er r-t på vannkraft* / *r-e malmleier* / *ha en r- fantasi* / *bli en erfaring r-ere* ta lærdom av noe / god *en r- frukthøst* / mangeartet *en r- fauna* / *leve et r-t liv* få mye ut av livet **~dom** *-men, -mer* det å være rik, velstand, formue, overflod *vinne makt og r-* / *r- på ord*

The entry for RIK in Landrø and Wangensteen 1986

"a1" is a code indicating the inflection of the adjective in positive singular neuter and in positive plural. No information is provided for comparison.

However, when one has to deal with the kind of pedant, which the computer in fact is, one is compelled to decide. Which in fact has been done. With due margin for linguistic creativity, only clearly unacceptable or ungrammatical forms (indicated by means of interrogation marks and asterisks without parentheses above) have been excluded. As a result, our documentation is probably the **most accurate and extensive** that currently exists for Norwegian.

Why then decide? Why not simply let the dictionary contain all conceivable inflected forms of a given lemma as long as it is not clearly incorrect? The tacit conventions assuring representation economy in traditional printed dictionaries are usually not misinterpreted - at least not by native users of the language in question. And, as the above discussion clearly demonstrates, every choice and decision about a given wordform is also a normative act, thus blurring the ideal, theoretical distinction between the normative and the descriptive domains of linguistics.

However, there are several reasons for being precise on this point in the perspective of commercial software development. Space and storage is the least important of them. Of more importance is the fact that application software will also be intended for non-native users. And, no application actually needs overgeneration of this sort, while there are those for which a precise list of inflected forms is of great value. One such application is spelling aid. In case a wordform such as *rikests* should appear as the proposed correct form of for instance the misspelling "rikestr" (intended correct form: *rikeste*), it would probably undermine the user's confidence in this software function - and he or she is liable not to use it

again. Overgeneration is also bad for analysis grammars, which are - or rather will be - basic to a lot of linguistic functions. Superfluous wordforms may present the parser with an extra number of homonyms as the possible source of multiple parses and consequent lack of precision. E.g. *røder*, a dubious form of the noun RØD, but the present form of the verb RØDE 'talk, chat'. Similar examples are *støyer* of STØY 'noise'/STØYE 'make noise', *lufter* of LUFT 'air'/LUFTE 'air', and mater of MAT 'food'/MATE 'feed'. In case a form ?*mater* (of MAT) is included in the lexicon, a sentence such as De *mater ender* will have at least two parses:

$De_{NP}\ mater_{VP}\ ender_{NP}$. og $(De\ mater)_{NP}\ ender_{VP}$.

On the other hand, it is of great importance that all legal forms of a word are represented. This is in fact the reason why it will always be difficult to base oneself completely on actual, documented language use, i.e. some sort of a text corpus. When compiling a machine-readable dictionary one cannot simply list the lexemes (and their respective inflected forms) or the wordforms represented for each lexeme. No matter how infrequent one wordform might be - in a given corpus and in all conceivable actual use of the language in question - it may always be felt as a severe deficiency when it is missing in some sort of application. The not so often used variants *huser* (of HUS 'house', indefinite plural, parallel to *hus*), *lesinga* of LESING 'reading', definite singular, parallel to *lesingen*), and *veit* (of VITE 'know, past form, parallel to *vet*) of Norwegian Bokmål may serve as examples of inflected forms. On the level of lexemes, words denoting certain parts of the body are good examples of infrequent words which still pertain to the kernel vocabulary of a language.

In fact, a text corpus has to be very big and very well composed in terms of representativity in order not to contain far more disastrous lacunas. One example is the frequency list based on a corpus of Norwegian business correspondence mentioned above. It did not contain the adjective *rød* 'red' nor the personal pronoun *jeg* '**I**'.

Still, actual texts - and dictionaries - have to be consulted in order not to leave out important lexemes, since no lexicographer has the complete oversight of the entire vocabulary of a language. Sublanguage material has to be scanned to ensure the inclusion of for instance essential professional terms etc.

Normalisation Relative to Application

So far, it has been taken for granted that a machine-readable dictionary as the base for the development of software functions has to contain only correct wordforms, and that it has to be as extensive as possible - leaving out no frequent correct forms. However, this is by no means obvious: It depends on the use one is going to make of the dictionary.

Different wordlists have to be created for different applications on the basis of the general dictionary. As the input for a spelling checker, for instance, a list containing only **correct** wordforms is necessary. But for special purpose spelling checkers, e.g. for dyslectics, a limited choice of correct wordforms is optimal. And certain correct wordforms may also have to be omitted from any spelling checker. E.g. IN 'fashionable, trendy', a relatively recent loanword in Norwegian with English origin, since it is identical to one frequent misspelling of INN, the far more frequent adverb 'in'.[7] On the other hand, a linguistic component in a text cri- tiquing system also needs a separate list of incorrect wordforms (both incorrect stems and inflectional forms), while a search program may need one single list with both correct and incorrect wordforms which are supposed to be frequently used in the potential search data.

Such incorrect wordforms may belong to either earlier stages of language standardisation or to regional or social dialects. This is of particular importance in the case of Norwegian, since it has been subject to a number of orthographic changes in this century - far more radical than those of other Western European languages. As an example showing the different requirements of application programs, consider the preposition ETTER 'after'. *etter* is the correct form, and consequently the only candidate as input to a spelling checker. On the other hand, *efter,* an obsolete form in Bokmål, must also be included in the input to a search system - also as a constituent of compound words. If not, one fails to recognise both *etter* and *efter* in older or non-standard data. One example of a widely used non-standard form with a certain regional basis, is *sanda* (definite form singular of SAND 'sand'), which is of feminine gender in

---

[7] Current commercially available spelling checkers are unintelligent. I.e., not even a fragment of an analytic grammar is implemented. The technical solutions may vary considerably, but, basically, they are all based on the very same principle: Sequences of characters in a text are matched with sequences in some sort of dictionary. This means that also a correct wordform at an incorrect place in the sentence will be recognised as correct.

South-Eastern dialects (cf. the -A suffix) in opposition to the "official" masculine gender, *sanden* (with an -EN suffix). The dictionary of search functions may also contain words from more than one language. In the case of Norwegian, it will be useful to include a number of Nynorsk words in a Bokmål search program (e.g. LØYVE 'permission', *kjømda* 'the near future', and STØNAD 'support') and vice versa (e.g. BE-HANDLE 'handle', FORPLIKTELSE 'obligation', and OVERVINNE 'defeat' in addition to VINNE OVER). The inclusion of a few English words, e.g. KNOW-HOW, TEAM, and LEASING will also be natural.

Conclusion

Three points of general interest can be inferred from the above discussion on correctness, adequacy, and selection in dictionary development:

Normalisation is relative to the type of software which the dictionary is designed for. Some applications need lists of correct words only, others not. Some need extensive dictionaries, others specially tailored ones.

In language industry, **normative and descriptive considerations are inseparable**: Descriptive adequacy and economy are only achieved if the developer takes a stand as to linguistic correctness.

Thus, language industry also means **new challenges for the traditional lexicographer**. In academic computational linguistic research, there is a natural concentration on formalisation and test systems with a limit-ed and frequently inaccurate linguistic basis. In language industry, tradi-tional lexicography is needed, but at the same time, the traditional lexicographer has to face new requirements for precision.

References

Akø, J.-O. (1992): Gråsoner i norske ordbøker. In: *Fjeld*.

Engh, J. (1992): IBM Norges database for moderne norsk. In: *Fjeld*.

Engh, J. (1992): Språkforskning i IBM Norge. *Norskrift* 72. Enlarged English version: Engh, J.: IBM Norway's Database for Present-day Norwegian. Unpublished paper (IBM Norway). Kolbotn 1991.

Fjeld, R.V.(ed.) (1992): *Nordiske studier i leksikografi. Rapport fra Konferanse om leksikografi i Norden 28.-31. mai 1991*. (Proceedings from the First Nordic Conference on Lexicography, Oslo 28-31 May 1991). Oslo.

Landrø, M.I. and Wangensteen, B. (1986): *Bokmålsordboka*. Oslo.