Lexicography for IBM

Developing Norwegian Linguistic Resources in the 1980s

Jan Engh

Oslo University Library (Norway); jan.engh@ub.uio.no

Abstract: In 1984, IBM and the University of Oslo set up a joint project, probably the first project of its kind in Norway. Its aim was to develop Norwegian language resources for IBM application software – for PCs, midrange computers, and mainframes. The primary objective: to create a "base dictionary" module that would drive language sensitive functions. The technology was based on simple character sequence recognition; its great asset being high compaction and rapid access to correct data. The module was to be built on documented linguistic forms. The dictionary should cover the general part of the vocabulary, and a broad coverage module was created for Norwegian Bokmål. Later, one module for Nynorsk was developed as well. At that stage, however, the project had become a regular IBM project. In the following years, other linguistic functions were added and eventually, the result served as the foundation for a grammar and for machine translation. The project was terminated because of the corporate financial crisis of the late 1980s. Later, the dictionaries were transferred to the University of Oslo. They are now an integral part of the basic infrastructure for Norwegian academic computational linguistics.

Keywords: Lexicography, Norwegian language, natural language processing, IBM, application software

1. Introduction

In 1984, IBM's Advanced Office Systems Technology (AOST, Gaithersburg MA) launched an international corporate offensive to create language sensitive software after several years of development work for English (see [3] and [4]). As in the rest of IBM's Europe, Middle East, and Africa division, local managements in the Nordic countries were instructed to start development, if necessary in cooperation with the local universities.

Since IBM Norway had no staff with the necessary linguistic competence at the time, the natural choice was to turn to the University of Oslo for linguistic assistance. A joint project was set up with the specific aim of developing necessary resources for the Norwegian language and having them implemented and tested using relevant application software. Corresponding development was carried out for all the other major Nordic languages in the respective countries, although with great variation both as far as organisation and linguistic development etc. were concerned.

To my knowledge, this was IBM's first formalized software development contract with a Norwegian university. One University of Oslo research officer was assigned to the task as project leader (Jan Engh) that later involved several assistants. A steering committee was appointed, constituted by one IBM representative (Jan Hølen), and two representatives from the University of Oslo (Even Hovdhaugen and Jo Terje Ydstie) in addition to the project leader. The project took place on IBM premises and it was 100% funded by IBM, which also had the exclusive right to the research and development results.

2. The Project

The primary objective of the project, referred to internally as "the LEXIS project", was to create a linguistic component, a "base dictionary" module, for all natural language sensitive software. This module would function as an extensive "dictionary" for the analysis (recognition) of Norwegian word forms and for the generation both of alternatives to unrecognized possible words and of hyphenation points in word processing programs. Originally, the base dictionary module was intended for use in text-processing software only. Later, it was used for other types of application software as well and more components were added.

With a minimum of adaptation, the module was supposed to drive language sensitive functions in all of IBM's own application software for the embryonic PC market via midrange computers to the mainframes: Spelling checker, correct word form suggestions, and automatic hyphenation.

At this phase of development, all these functions were "unintelligent", based on simple character sequence recognition. The philosophy behind the base dictionary was that it would provide any program with documented information of the language in question. The coverage was extensive, and for instance rule-based hyphenation algorithms were to be used only for unrecognized character sequences. The great asset of the base dictionary concept was its compaction technology and the rapid access to the correct data, both extremely important factors at a time when the IBM XT (introduced 1983) came with a 256 kB memory working at a pace of 4.77 MHz.

The target group was *all* possible users. The dictionary component was not intended for office use only, but for school and everyday purposes as well. This was reflected in the coverage of the base dictionary. It should cover the general part of the vocabulary. IBM management had quite an optimistic view of how computers would penetrate into daily life.

3. The Base Dictionary

3.1 Vocabulary

The vocabulary of the base dictionary contained the core lexicon of Norwegian and as much more as practically possible. All and only the forms of the word

types (lexemes) were included except defective forms, e.g. plural forms of most abstract words, such as *hat* 'hatred'.

In accordance with the company's policy of observing official standards and maintaining political neutrality, the official orthography as laid down by *Norsk språkråd* (The Council for the Norwegian Language) was adopted.

3.2 Architecture

The base dictionary consisted of three levels:[1] One "ultrahigh frequency" wordlist of 204 word forms was contained in the set of words generated by a "high frequency" dictionary, which in turn was a proper subset of the word forms generated by the main dictionary. In the original English version, the ultrahigh frequency list was supposed to represent approximately 50% of the word forms in a general text. The high frequency dictionary, in turn, should cover 85%.

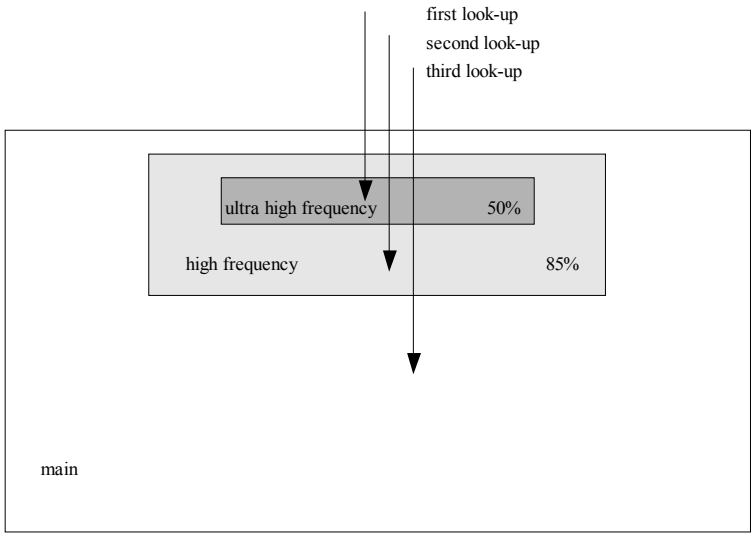

Figure 1. Base dictionary architecture

This general architecture for Norwegian was adopted without modifications, despite the fact that it was based on frequency data for English.

3.3 Linguistic Interlude

The frequency assumptions were not the only aspect of the base dictionary concept showing that it had been developed for the English language. There was, in fact, a clear correspondence between the technical solutions adopted and the

[1] A detailed documentation of the input files formats, the development project history and the ensuing products can be found in Engh 1991[5]. See also [6].

particular structural characteristics of English. Or, to put it differently: From a morphology point of view, English is, by coincidence, a "simple" language that happened to be adequately catered for by the current state-of-the-art technology.

A trivial point was the fact that the set of characters allowed in the input files was the one of English, A – Z. Since all other European languages use more characters, these had to be represented as double byte characters. Less trivial were the cases where the lack of English "simplicity" could be compensated for by quantity. English has only two noun forms, singular and plural[2] and a minimum of verbal forms.

singular	plural
car	*cars*

infinitive/present plural	*swear*
present singular	*swears*
past	*swore*
perfect	*sworn*

In linguistic terms, English has a poor morphology. Norwegian, on the other hand, has four noun forms and a few more verbal forms etc. Additionally, Norwegian orthography is characterized by a certain variability; that is, each "slot" in the paradigm may be occupied by more than one form, which means that for instance a number of nouns have many more than four forms.[3] Not only

singular		plural
indefinite	*bil* 'car'	*biler*
definite	*bilen*	*bilene*

but even

singular		plural
indefinite	*bok* 'book'	*bøker*
definite	*boka*, *boken*	*bøkene*
indefinite	*system* 'system'	*system*, *systemer*
definite	*systemet*	*systemene*, *systema*

[2] The *'s* genitive is a suffix that can be added to almost every noun, and it is correspondingly easy to analyse and generate. This also holds for its Norwegian parallel, *s*, which may even attach to participle forms.

[3] And then there is the Bokmål/Nynorsk problem. Spoken Norwegian is one language with a number of dialects. However, there are two different ways of writing the language, Bokmål and Nynorsk. That is, Norwegian has two written standards. In the current setting, this means two separate base dictionaries etc. [19, pp. 53-57 and 98-104] describes the relationship between Bokmål and Nynorsk for those unfamiliar with the language situation in Norway.

or more, not to mention verb forms such as

infinitive	*sverge*, *sverje* 'swear (an oath)'
present	*sverger*, *sverjer*
past	*sverget*, *sverjet*, *sverga*, *sverja*, *svor*
perfect	*sverget*, *sverjet*, *sverga*, *sverja*, *svoret*
perfect plural/weak form	*svergete*, *svergede*, *svorne*

However, this can be compensated for by just adding word forms and making the dictionary bigger. In principle, this has the effect that the (English) frequency considerations behind the tripartite architecture of the base dictionary become somewhat distorted. In practice, however, the Norwegian ultra high frequency list represented no problem. As for the high frequency dictionary, it had to be based on a rough estimate, since adequate frequency data for Norwegian was not available and could not be produced within the limited timeframes of the first project. Still, relatively infrequent genitive word forms, for instance, were simply omitted from the paradigms of the *lexemes* selected based on frequency data from many sources.

To some extent, a greater quantity can even compensate for the, in theory, infinite number of compound words of Norwegian. In written English, compounds are, in general, sequences of separate words, whereas Norwegian compound words, in contrast, constitute single complex words with other possible words as their constituents. In the English dictionary, *red* and *wine* has two entries, which also cover *red wine*. The Norwegian dictionary needs three entries: *rød* 'red' *vin* 'wine', and *rødvin* 'red wine' with the same inflected forms as *vin*. One apparent solution is to introduce a rule to combine constituents, which, in fact, was an option in the base dictionary format. There are, however, two main reasons why one should not adopt such a solution. One relates to possible applications (see below); the other is of a direct linguistic nature.

In addition to the simple juxtaposition type mentioned above, Norwegian exhibits far more complicated systems in multiple compounds (see [11, pp. 71f]). For instance, the emerging S and the disappearing E linking the main constituents pertaining to various different classes of lexemes: On the one hand, *vinglass* 'wineglass', and *krystallvinglass* 'wineglass made of crystal', but *rødvin**s**glass* 'glass for red wine' with an *s*. On the other hand, *last**e**skip* 'cargo ship', and *diesellast**e**skip* 'diesel cargo ship', but *tørrlastskip* 'dry-cargo ship' without the *e*. No simple expansion of any dictionary can compensate fully for this type of occurrence.

3.4 Input Files

For both the main dictionary and the high frequency dictionary, the linguistic input files consisted of one stems file and several auxiliary files, of which the endings file was the most prominent (see [2]).

3.4.1 Stems File

From a linguistic point of view, a word is analyzed as a stem plus optional affixes (derivation or inflection). For instance, the Norwegian verb *bile* 'go by car; drive' contains the stem *bil* 'car' plus the infinitive suffix *e*. The noun *bile* 'axe', on the other hand, has the stem *bile*. Meaningfulness is a requirement. However, the base dictionary was *not* organized according to this linguistic principle. A stem in the stems file was a technical stem, which might or might not coincide with the linguistic stem. It had to be a valid word form, though, which meant that *bil* would represent both the noun *bil* and the verb and noun *bile* in the stems file. This had rather peculiar consequences for the analysis and inclusion of the Norwegian vocabulary, as will be shown below.

The stems file had the following format:

```
NORSK2    DICT     D1   V 150  Trunc=150 Size=33015 Line=7883 Col=1 Alt=0
====>
!...+...10....+...20....+...30.....+...50....+...60....+...70....+...80..
finn                          0   0NV.............O ØcING ØcV1 Øc11      07883
(---)
fins                          0   0JV.............B                      07898
finsk                         0   0J..............A ØJ11                 07899
Finske_bukt                   0   0N..............A ØF3B                 07900
fint_bygd                     0   0J..............O ØJ11                 07901
fint_føl_en_de                0   0J..............O ØJs                  07902
```

Figure 2. Sample of stems file

Every record contained information about one "stem" and other word forms derived from the stem and the endings indicated. The stems were written from the first column, in EBCDIC with non-English characters represented as described in an auxiliary file. "_" represented a hyphenation point. (In later versions, the possibility of preferred hyphenation points was introduced.) In the following columns, additional information was stored. Columns 41-44 contained information about confusable stems, while column 45 was reserved for grade level information in the US English version. Columns 46-53 contained part of speech information ("N" for 'noun', "V" for 'verb', "J" for 'adjective' etc.), and column 61 was reserved for a word compounding flag. From column 63 to the end of the record, optional information was entered about the word forms that could be derived based on the stem.

As for the compounding flags, they indicate whether a given word form could appear as the constituent of another – compound – word. Additionally, each ending associated with the stem carried such a flag, indicating the combinability of the derived word form. Based on this information, compound words not represented in the dictionary were supposed to be recognized.[4] There was a limited set of possible combinability values: "B" 'back or isolated', "O" 'offset

[4] Originally, this component was intended for both analysis and generation of compound words, e.g. even for the spelling aid window, see [13] and [14]. It was, however, completely unsuitable for the latter purpose, and was never used for generation in any software for Norwegian.

(i.e. cannot be a constituent of a compound word)', "F" 'front or middle', "A" 'anyplace', and four more. This was a simplistic, yet extremely powerful device – to the extent that it ought to be heavily restricted. Only a very limited use of the compounding flags was made for Norwegian. (With the extra precaution that the spelling checker of the application program would only consider words of considerable length as possible constituents.)

Column 63 etc. contained information about all the words derivable from the stem. Most of it in was provided in a shorthand writing as "implied endings", i.e. codes representing sets of endings. These implied endings were declared in the endings file.

3.4.2 Endings File

The endings of the endings file were technical suffixes, parallel to the technical stems of the stems file.

```
NORSK2   LEXISEND  A1  F 80  Trunc=80 Size=1494 Line=0 Col=1 Alt=0
====>
!...+....1....  +....2....+....3....+....4....+....5....+....6....+....7..

* * * Top of File * * *                                              00000
ØcJ1     e t ere est este s es eres estes                           00001
          >J AJ AJ AJ  J    J J  J    J                            00002
         >>O O  O  O   O    O O  O    O                            00003
ØcJ1s    e t ere est este es eres estes                             00004
          >J AJ AJ AJ  J    J  J    J                             00005
         >>O O  O  O   O    O  O    O                             00006
```

Figure 3. Sample of endings file

Every three-line "paragraph" constituted the record of an implied ending set. In the first columns, the name of the set was given with an initial Ø character. From column 10, the endings were listed. On the second line, the part of speech for each stem plus ending was given (a word *pen**e*** is an adjective, >J, *pen**t*** an adverb or an adjective, >AJ, etc.). For practical reasons, they were both marked as offset, >>O.

3.4.3 Additional Auxiliary Input Files

Several additional auxiliary files were to be created. Some were of a linguistic nature such as those containing the 60 most frequent endings, information about hyphenation of endings, alternative representation of sounds, and characters, while others had a strictly technical content.

3.5 No Duplicates

The unfolded result (i.e. all the stems and the word forms generated by means of the stems and the endings) had to be free from duplicates. That is, one word in the sense of 'sequence of characters' should be represented only once, regardless of

possible multiple meanings. This was important in order to save space, facilitating the compaction process and making recognition more efficient and precise. To implement the no duplicate requirement, a huge and complicated puzzle work was necessary. In fact, both the development of appropriate implied endings and the classification of the vocabulary by means of the result constituted a time-consuming activity. (See [5, pp. 12-18].)

4. Preconditions and Implementations

A complete development environment was provided by AOST for the developmental work, which was carried out on a 370 mainframe under VM/CMS. Initially, AOST even provided an electronic corpus, based on IBM business correspondence in Norway. Obviously, it was totally unbalanced and turned out to be of little value. Since no extensive machine-readable linguistic material for Norwegian was freely available at that time,[5] development continued by including all words found in accessible printed frequency material and words pertaining to all relevant lexical spheres (colors, parts of the body, kinship and construction terms etc.) were systematically entered. To extend the coverage further, the developers' private documents and all accessible company documents were regularly passed through the updated spelling checker to detect candidate words. Dictionaries were consulted, never copied. (It was a strict requirement that other authors' copyrights were not infringed upon.) Moreover, a considerable amount of information was required that could not be found in any dictionary such as defective forms. This was produced by the project group.

After successful test building, all source files were shipped to Gaithersburg for the final build and implementation in application software: *DisplayWrite* (later, even in lower end software such as *WritingAssistent* and in specialized composition software). Finally, the dictionaries had to be tested for each software release at IBM Norway. The project group established a special corpus of texts for this purpose.

5. History and Strategic Figures

The project started in the summer of 1984. The first phase ended one year later. The subsequent phases of the project were carried out as a regular IBM research and development project. Although the *formal* ties to the University of Oslo were severed, close informal ties were kept in view of recruitment. (24 linguists

[5] There were a few insignificant and scattered resources at the University of Oslo and Bergen in the early 1980s. However, they were inaccessible for an industry development project such as the one of IBM. Additionally, there was the machine-readable manuscript of *Bokmålsordboka* (a medium size monolingual dictionary) which was not yet finished in its first version [16],. However, *Bokmålsordboka* was inaccessible to the IBM project. Several years later, the right to use the electronic manuscript was acquired for IBM Norway internal use only. Still, it was never utilized for development purposes, only as test bed for a separate experimental linguistic database format, *WordSmith*, see [1].

worked for the group, part-time or full-time for shorter or longer periods. (See [12].)

At the end of the first phase, the base dictionary contained 30,972 stems; the number of unique word forms generated was 292,190 - much more than existed at the base of the linguistic functions of any competitor at that time. Yet, this dictionary size was still far from ideal for a language considerably more inflected than English, although a significantly higher number of lexemes than 30,972 were covered due to the great quantity of homonyms. (The limited use of the compound recognition device expanded the total number of word forms covered even more.) Only Norwegian Bokmål was catered for during the first phase of the project. There were five subsequent Bokmål releases in all. The last main dictionary (1989) contained 51,292 stems, generating 487,166 unique word forms.

A parallel development program for Nynorsk started in 1988. Due to linguistic factors, the Nynorsk resources had to be created almost from scratch. No simple conversion of linguistic data from one written standard to the other was possible. The second and final release for Nynorsk (1990) had a base dictionary of 92,787 stems, generating 360,680 unique word forms.[6]

6. The Need for an Implemented Morphology

In a subsequent phase, IBM wanted to develop new dictionary functions. Now, this was problematic, as the base dictionary was not properly organized from a linguistic point of view. That is, the information provided by the base dictionary did not indicate to which lexeme or lemma a given word form belonged. Cf. the case of *bil* 'car' and *bile* 'axe' above, which were recognized because of the very same "technical" stem. Thus, a genuine morphology and a corresponding lexicon had to be establish as a bridge between the word form recognition component of the base dictionary and, for instance, a synonym dictionary: Its main components were the lexicon input file and the inflection input file.

```
NOB_UTV  CLASS     A1  F 50  Trunc=50 Size=34 Line=0 Col=1 Alt=3
====>
!...+....1....  +....2....+....3....+....4....+....>

* * * Top of File * * *                                        00000
bygg                                  800                     00001
bygg                                  890                     00002
byggaks                               804                     00003
byggbrød                              800                     00004
byggdyrking                           900                     00005
bygge                                 031                     00006
```

Figure 4. Sample of lexicon input file

The structure of these files was very simple. In the lexicon input file, the lemma forms (the singular indefinite forms of nouns, the infinitive of verbs etc.)

[6] The low number of unique word forms compared to the number of stems is mainly due to the fact that only proper names have a genitive form in official Nynorsk orthography, not nouns in general as in Bokmål.

were listed with a code referring to an entry in a paradigm, which was given in the inflection input file. For example, *bygg* 'building; construction site' is inflected according to the paradigm Ø800, while *bygg* 'barley' is inflected according to Ø890, the verb *bygge* 'build; construct' according to Ø031 etc.

In its fifth and last commercial release, the Bokmål morphology contained 65,128 lemmata and 705 paradigms (1989). In the version ready when the entire project was terminated (1991), this had been expanded to a total of 121,577 lemmata. As for the Nynorsk files, the second and last release (1990) contained 110,412 lemmata and 576 paradigms.

```
NOB6     TABLE     A1  F 50  Trunc=50 Size=7540 Line=782 Col=1 Alt=0
====>
!...+....1....  +....2....+....3....+....4....+....>
*neutr fullst N1 <tak>                                              00781
Ø800 Ø    NORNN                                                     00782
01                                                                  00783
02 et                                                               00784
03                                                                  00785
04 a,ene                                                            00786
05 s                                                                00787
06 ets                                                              00788
07 s                                                                00788
08 as,enes                                                          00788
```

Figure 5. Sample of inflection input file

The development software and the test-building environment for the morphology were developed in cooperation by AOST, the Centro Científico de IBM (Madrid), and by the Norwegian lexicography group during the first quarter of 1986. This sub-project was carried out via VNET, IBM's own network system in the 1980s. VNET made it possible to run continuous working sessions between persons in different locations, in this case Bethesda (MA, USA), Madrid, and Oslo. In general, VNET was extensively used during the entire project period. There was contact between all the linguistic groups of Europe and the Middle East and the US laboratories more or less on a daily basis with a two-way flow of technical and linguistic information.

7. New Dictionary Functions

The implemented morphology made it possible to create a "morphology window", where the declension or conjugation of a given word form in a text could be displayed. This feature had a great educational potential, since the end user had the possibility to swap between linguistic modules (Norwegian, English, German, etc.). More importantly, the morphology paved the way for a synonyms function. Two extensive synonym dictionaries were created from scratch by the IBM Norway linguist group, one for Bokmål and one for Nynorsk.

The latter edition was the largest, containing approximately 25,000 entries, corresponding to a mid-size printed dictionary. The objective of the synonym dictionaries was to help the end user to write better Norwegian. There were two reasons why the synonym dictionaries had to be developed: The existing (printed)

dictionaries had been edited mainly in view of solving crossword puzzles, i.e. not to help the users improve their writing. In addition, unlike the case with other language communities, IBM could not purchase them.

```
NYNORSK  IBM_SYN   A1  V 80  Trunc=80 Size=267 Line=91 Col=1 Alt=0
====>
!...+....1....+....2....+....3....+....4....+....5....+....6....+....7..
akta:j  gjæv, god, høgvørd.                                          00089
akte:v                                                               00090
    ta omsyn til, leggje vekt på, anse, ense, merkje seg,            00091
     vere merksam på;                                                00092
    ha stor vørnad for respektere, ære, heidre, vørde, tykkje om,    00093
     synast om;                                                      00094
    verdsetje, vurdere, mæte, skatte, estimere;                      00095
    vilje, emne på, rekne med å, tenkje, tenkje på, ha i sinne,      00096
     intendere.                                                      00097
    <C seg>                                                          00098
        vare seg;                                                    00099
        etle seg;                                                    00100
        vilje;                                                       00101
```

Figure 6. Sample of synonyms input file

Additionally, rules for algorithmic hyphenation were implemented and a variation representation prototype was developed, exploiting the vacant grade level indicator for US English. It indicated to which level within a given written standards of Norwegian a word belonged, e.g. "radical" vs. "moderate" Bokmål. However, this particular feature was never implemented in any product.

8. Linguistic Challenges

During the technical development, the project group spent much time and effort clarifying the linguistic norm – not at all a trivial matter as far as the Norwegian language is concerned. One complicating factor was the inherent variability that characterizes Norwegian in contrast to most other languages, although the main cause was undoubtedly the surprisingly incomplete and incoherent standardization of the language in general. In innumerable cases, the *Norsk språkråd* had to be consulted – also for the benefit of Norwegian normative grammar. See [8]

9. Further Development and Market Considerations

The twin morphologies were later used as a basis for further "intelligent" linguistic functions (grammar and style critiquing) and stand-alone software (machine translation). That is another story (documented in [10]). Neither did materialize as products, due to the financial crisis that shook IBM in the late 1980s. At that time, entire development areas were eliminated, no matter their quality or state of progression. In the case of the linguistics development, one reason for its termination may have been that IBM's own word processing software, the *DisplayWrite* products, which never sold well, despite their comprehensive dictionary features. IBM sales representatives never understood

this asset, and more importantly: They had no special incentive to promote them. The project failed to finance itself during the development period, and by the end of the 1980s, the days of very long-term investments in American software industry were gone.

10. Concluding Remarks

The entire IBM linguistic effort was a broad front offensive. At a time when other companies and academia contented themselves with creating linguistic toy systems that did not scale up, IBM went for full-scale development, covering language in general. At a regular pace, development advanced from the lexicographic basics to state-of-the-art computational grammar. (See[15]) Moreover, equally important, every language of at least the size of Icelandic saw almost parallel development. This was, in fact, the first worldwide investment by a private company in the area of multilingual natural language processing.

When IBM finally quit linguistic development for Norwegian, after almost 8 years of intense work (see [7]), AOST's successor sold the penultimate version of the lexicon and the morphology to a publisher, while IBM Norway transferred the most recent files to the University of Oslo for a symbolic sum. Today, they are part of the basic infrastructure of academic computational linguistics in Norway.[7]

Acknowledgments

I extend a special thanks to Diana Santos who read the draft and to Stig Johansson and Per Vestbøstad for clarification of facts.

References

[1] Baustad, Jostein: 1992, "Automatisk analyse av maskinleselige ordbøker til bruk i en orddatabase" 'Automatic analysis of machine readable dictionaries for the creation of a word database'. In Fjeld, Ruth Vadtvedt (ed.): *Nordiske studier i leksikografi. Rapport fra Nordisk konferanse i leksikografi i Oslo, mai 1991*. (*Skrifter utgitt av Nordisk forening for leksikografi* 1) Oslo: Nordisk forening for leksikografi, 423-431.

[2] Casajuana, Ramón: 1989, LEXIS input files. A comprehensive description. TESTBLD. A user guide. Unpublished paper, Centro Científico de IBM, Madrid.

[3] Convis, Danny B., Glickman, David, and Rosenbaum, Walter S.: 1982, "Alpha content match prescan method for automatic spelling error correction". United States Patent 4,328,561.

[4] Convis, Danny B., Glickman, David, and Rosenbaum, Walter S.: 1983, "Instantaneous alpha content prescan method for automatic spelling error correction". United States Patent 4,355,371.

[7] See [18], [17], [9], and the following pointers (1 December 2008): http://www.dokpro.uio.no /ordboksoek.html; http://www.hf.uio.no/tekstlab/innsyn/norsk.html; http://www.edd.uio.no /prosjekt/ordbanken/

[5] Engh, Jan: 1991, *IBM's Norwegian Lexicon Projects 1984-91*. Unpublished report, IBM Norge. Kolbotn.
[6] Engh, Jan: 1992a, "Leksikografi i IBM Norge" 'Lexicography at IBM Norway'. In Fjeld, Ruth Vadtvedt (ed.): *Nordiske studier i leksikografi. Rapport fra Nordisk konferanse i leksikografi i Oslo, mai 1991*. (*Skrifter utgitt av Nordisk forening for leksikografi* 1) Oslo: Nordisk forening for leksikografi, 409-422.
[7] Engh, Jan: 1992b, "Språkforskning i IBM Norge" 'Linguistic research at IBM Norway'. [Paper read at *Møte om norsk språk* (*MONS*) IV, Oslo 15.-17.11.1991] Printed in *NORSKRIFT* 72, 16-36.
[8] Engh, Jan: 1993, "Linguistic normalisation in language industry: Some normative and descriptive aspects of dictionary development". *Hermes. Journal of linguistics* 1, 53-64.
[9] Engh, Jan: 1994a, *IBMmorf*. *Bruksanvisning for IBMs leksikon og morfologi for moderne norsk* 'Manual for the use of IBM's lexicon and morphology of Modern Norwegian'. Dokumentasjonsprosjektet, Universitetet i Oslo [Available at http://folk.uio.no/janengh /IBMmorf.htm]
[10] Engh, Jan: 1994b, *Developing Grammar at IBM Norway 1988-91*. Unpublished report. Oslo.
[11] Engh, Jan: 2001, "Bindebokstaver" 'Binding morphemes', In Gundersen, Dag, Jan Engh, and Ruth E. Vatvedt Fjeld: *Språkvett. Skriveregler, grammatikk og språklige råd fra a til å*. Oslo: Kunnskapsforlaget, 67-72.
[12] Engh, Jan: [s.a.]: "Natural language processing at IBM Norway". Avaliable at http://folk.uio.no/janengh/IBMnorsk.htm
[13] Frisch, Rudolf and Antonio Zamora: 1988a, "Method for verifying spelling of compound words". United States Patent 4777617 [available at http://www.freepatentsonline.com /4777617.html (7 April 2007)].
[14] Frisch, Rudolf and Antonio Zamora: 1988b, "Spelling assistance for compound words". *IBM Journal of research and development* 32/2, 195-200.
[15] Jensen, Karen, George Heidorn, and Steve Richardson (eds.): *Natural Language Processing: The PLNLP Approach*. Hingham (Mass.): Kluwer 1992.
[16] Landrø, Marit Ingebjørg and Boye Wangensteen et al.: 1986, *Bokmålsordboka*: definisjons- og rettskrivningsordbok 'A dictionary of Norwegian Bokmål: Definitions and orthography'. Bergen: Universitetsforlaget.
[17] Ore, Christian-Emil: [1999], "Metaordboken - et rammeverk for Norsk Ordbok?" 'The metadictionary – a framework for Norsk Ordbok?' [Paper read at the conference *Leksikografi i Norden*, Göteborg 1999] Available at http://www.edd.uio.no/artiklar /leksikografi/artikkel_Goeteborg.html.
[18] Ore, Christian-Emil and Nina Kristiansen (eds.) [s.a.]: *Sluttrapport 1992-1997* 'Final report 1992-1997'. Dokumentasjonsprosjektet, Universitetet i Oslo. Available at http:// www.dokpro.uio.no/sluttrapp.pdf.
[19] Vikør, Lars: 2001, *The Nordic languages*. *Their status and interrelations*. [3rd edition] (*Nordic Language Secretariat. Publication* 14) Oslo: Novus.