Ivar Utne

Machine Aided Translation between the Two Norwegian Languages Norwegian-Bokmål and Norwegian-Nynorsk

Abstract

The article describes essential parts of a prototype system for machine aided translation from Norwegian-Bokmål to Norwegian-Nynorsk. The central parts of the system are a bilingual word list, inflection paradigms, phrases, and routines to deal with compound words. There are also syntactic and semantic rules, but they can be considered as preliminary. The article also includes a simple comparison of the two languages in question. The program is written in TurboPascal and runs on IBM PC compatible machines.

1 Introduction

The main goal of the project *Machine aided translation from Norwegian-Bokmål to Norwegian-Nynorsk* is to automatize translation between the two official Norwegian languages, both for proposing translated text, and for looking up words, phrases and text parts. There is also an important objective to gain knowledge about machine translation in general. The system will be based on comprehensive bilingual word lists and linguistic rules implemented as data for a computer program.

The development has so far partly been supported by the Norwegian Research Council for Science and the Humanities (NAVF) for 8 months in 1987–88. The author of this presentation has been responsible for the linguistic and computational work. Formal cooperators are also The Noregian Computing Centre for the Humanities and Department of Phonetics and Linguistics at the University of Bergen.

Proceedings of NODALIDA 1989, pages 250-259

1.1 The Soft- and Hardware

The translation program and other data development programs (especially for word lists) are written in TurboPascal without usage of graphics and special database procedures, and is compatible with versions 3.0, 4.0 and 5.0. The source code takes about 50 kb, and the compiled version about 34 kb without the data tables. The program is run on IBM compatible machines with the operating system DOS 4.0 and lower. Among the machines is a Toshiba T1000 with 512 kb RAM and one floppy disc with 720 kb.

2 The Two Official Norwegian Languages, General Remarks

The two official written Norwegian languages are the majority language Norwegian-Bokmål (NB), with an unofficial English name Dano-Norwegian, and the minority language Norwegian-Nynorsk (NN), with an unofficial English name New-Norwegian. These languages can, for foreigners, be regarded as relatively similar both in spelling, inflection, vocabulary and syntax, which will be exemplified below. To Norwegians these two languages are felt as different, both because of the linguistic differences and because of the political impact of the languages. The two languages have by political decision been stated as official Norwegian languages with equal status.

NN is based on Norwegian dialects, while NB is based on an older version of written Danish which has later been improved with elements from Norwegian dialects. There is nowadays no complete agreement as to which of the two written languages are nearest to Norwegian dialects. That depends on which linguistic categories are considered and their relative importance. There is also no complete agreement to what extent a written language shall be based on dialects in opposition to written tradition.

This language situation implies that there will exist two Norwegian languages in the future. And this will motivate translation tools between the two languages, especially from the majority language towards the miñority language.

3 The Two Official Norwegian Languages, Simplified Comparison

To give some impression of what the system has to work with, I will present a simplified comparison of the differences between the two languages, and consider some differences in spelling, inflection, suffixes, vocabulary, and syntax.

First it must be stated that the two Norwegian languages have more similar than different expressions, and that the differences are very often small and cover lots of expressions in a systematic manner (e.g. diphtong versus monophtong). This reflects a common history for the languages.

Norwegian-Bokmål				Comments
Mod. NB	Rad. NB	Rad. NN	Mod. NN	
spelling:				
grøt	graut		graut	(porridge)
høst			haust	(autumn)
løk	lauk	løk	lauk	(onion)
ren	rein		rein	(clean)
fler			fleir	(more)
sette		sette	setje	(put, set)
linje		linje	line	(line)
skap		skap	skåp	(cupboard)
hjem	heim		heim	(home)
inflection:				
gutter			gutar	(boys, m pl)
jenter			jenter	(girls, f pl)
epler			eple	(apples, n pl)
hus			hus	(houses, n pl)
problemer	problem		problem	(problems, n pl)
boken	boka	boka	[boki]	(the book, f sg def)
kastet	kasta		kasta	(threw, pret)
svømte			svømte	(svam, pret)
kommer		[kjemer]	kjem	(come, irreg pres)
skriver		[skriver]	skriv	(write, irreg pres)
suffixes:				
utdannelse,-ing			utdanning	(education)
kjærlighet			kjærleik	(love)
lærer			lærar	(teacher)
elektriker			elektrikar	(electrician)
vocabulary:				
tillatelse,løyve			løyve	(permission)
erfaring			røynsle	(experience)

Table 1: Spelling, inflection, suffixes and vocabulary.

The comparison is presented in two tables. Table 1 contains examples of spelling, inflection, suffixes and vocabulary. Table 2 contains examples of syntactic constructions.

The examples in Table 1 are grouped into four colums, two for each language. The leftmost column below NB and the rightmost column below NN contain the forms that are most different from the other language, and are usually called *moderate forms*. The rightmost column below NB and the leftmost column below NN represent expressions which are regarded as approaching forms, usually called *radical forms*. The square brackets marks forms which are allowed in writing except in documents from the central government and text books for the primary and secondary school, high school, and some other institutions. Two of *genitive:*

guttens bil
=> bilen til gutten
(the boy's car; NN-phrase word by word: the car belonging_to the boy)

lederens forslag
=> forslaget/framlegget frå/til leiaren
(the leader's proposal; NN-phrase word by word: the proposal from/belonging to the leader)

passive voice:

bøkene kastes
=> bøkene vert/blir kasta
(the books are thrown (away))

boka blir lest
=> boka vert/blir lesen
(the book is read; common gender (i.e. masc. and fem.) singular)

congruence (occurs in passive and in predicative):

skriftet blir lest
=> skriftet vert/blir lese
(the publication is read; neuter singular)

bøkene blir lest
=> bøkene vert/blir lesne
(the books are read; plural)

nominalization:

forsamlingen gjorde vedtak om nye skatteregler
=> forsamlinga vedtok nye skatteregler
(the meeting/assembly approved new tax rules;
NB-phrase word by word: the meeting/assembly did/passed a resolution on new tax rules)

fyrbøteren har behov for mer kull
=> forbøtaren treng meir kol
(the stoker/fireman needs more coal; NB-phrase word by word: the stoker/fireman has/is_in need for more coal)

Table 2: Syntactic phrases

the expressions in the table are similar for the two languages, e.g. the words for *girls* and *houses*. This means that feminine plural inflection is the same (with some exceptions), and that neuter plural in some cases is similar.

In Table 2 there are examples of genitive, passive voice of verbs, congruence and nominalization. Each of the examples contains: NB phrase, NN phrase (prefixed by an arrow: =>) and an English translation often followed by some grammatical information. When the word order in one of the Norwegian phrases differs from the ordinary English translation, I have added a word by word translation of the phrase in that language. See for instance Myking 1989 for an instructive comparison of the two languages.

A system for machine aided translation will be based on these facts. The most significant differences are found in spelling and inflection of words. As well, we have to perform a syntactic and semantic analysis (parsing) to identify the current form among homonyms (in source language which may imply different expressions in target language), to ensure congruence (which exist in NN; see examples above) and to ensure rewriting of syntactic constructions.

4 The Structure of the System

At present the system is a bilingual translation system with transfer routines, i.e. target language data is inserted in the result structures from the analysis of the source language. A more sophisticated solution, not implemented yet, could be an interlingua system, where the source text would be translated to a language independent representation before generation of the target text.

The system is based on syntax analysis, and a semantic analysis guiding the syntax analysis. The development is performed bottom-up, i.e. it has been of importance to establish a skeleton of a total system, and then refine the parts. Because of this, the linguistic models that are used have to be considered as preliminary.

The system consists of a computer program and a collection of data files. The rules and the reference data is stored in text files that can easily be edited. During the development phase parts of the linguistic rules are part of the program code because of preliminary design issues.

The data files consist of linguistic rules for controlling the translation process (i.e. analysis and generation) and reference data for the rules. For the time being there is one rule file, with syntax rules, and seven reference files, which are the bilingual word list, regular inflection patterns, irregular inflected words, phrases, word formation patterns (for compound words), semantic classification, and internal code conversion (which will not be explained further).

5 Syntax Rules

The syntax rules are binary unification rules with context constraints. The system is working bottom-up, i.e. grouping related words in larger and larger units until the result is complete sentences. At the time being the rules are not based on a specific linguistic model, and the rules must be regarded as a basis for establishing a skeleton of a syntax parser. Typical rules are `ADNOM + NOM -> NP`, `V + NP -> VP`, `NP + VP -> S`, `PREP + NOM -> PP`. The rules are at the time being based on the sequence of `SUBJECT + VERBAL + OBJECT` in main clauses, and existence of inversion in certain types of subordinate clauses. The syntax rules can to some extent handle relative clauses, conditional clauses and adverbial clauses.

The context constraints concern both syntax and semantics. The syntactic constraints concern the existence of grammatical categories in the context, e.g. `SUBJECT`, `OBJECT`, `OBJECT2`, `THAT`-clauses and `INFINITIVES`. The semantic constraints concern what semantic features may characterize the grammatical categories (see above) which is tied up to a kernel word or phrase (usually a verb). It is quite a big project to work out a complete classification of semantic specification for all words. Therefore our principle is to start with classifying words which otherwise may cause wrong analyses and translations. The parser uses these specifications to reject analyses where the semantic features are in conflict. If one or none of the two phrases (words) have semantic specifications, a semantic check will not be performed. For further details, please see the description of the data file containing the semantic classification.

6 The Bilingual Word List

In the bilingual word list, a stem and an inflection code for both NB and NN are required. The follwing optional categories are present: context constraints (for the moment) for verbs (see syntax rules below) and semantic classification. The entries are stored in alphabetical order in a file with fixed record length to facilitate quick searches. The data is updated in a text file and converted to fixed file format.

The representation of stems are carried out according to a format that enables the program to search for the part of a word that is common to all forms, e.g. *mut* are common for *mutter* (indef. sing.) and *mutre* (indef. plur.), which are the NB word for the English word *nut*. Instances of double consonants which are single in inflected forms are represented with a hyphen in front of the second consonant, as for instance *-t*. And suffixes which are reduced in certain inflected forms are prefixed with an asterisk (*), as for instance **er*. According to this the entry for 'mutter' is *mut -t *er*.

7 Inflection

The system contains two data files consisting of inflection information, a table of regular inflection patterns and a list of irregular inflected words.

The presentation of the regular inflection paradigms has the following format for each pattern: pattern code + inflection expressions. For one of the inflection patterns for verbs, e.g. one of the two patterns for NB for the verb *kaste* (throw, cast), this means:

```
v6 e er a a
```

The reading of this is that according to pattern v6, these endings should be appended to the stem *kast* to produce the forms *kaste* (infinitive), *kaster* (present tense), *kasta* (past tense) and *kasta* (past perfect). There is a module in the program which uses this pattern information in combination with the stem form and its inflection code found in the bilingual word list. This routine will be exemplified in more detail below.

The presentation of irregular inflection is divided between the bilingual word list and a list containing the irregular forms in NN. In the bilingual word list the most relevant forms (four of nouns and verbs) of NB are listed according to the alphabetical sequence. Each of the NB forms are marked for a form category in the bilingual word list, e.g. present of verb or definite singular of a noun. In the list of irregular NN forms we usually find four forms organized according to a sequence, i.e. infinitive—present—preteritum—past perfect, that expresses what form it is.

8 Phrases

Phrases are stored in a text file to which the program has access. Words that usually occur in the phrase in only one form are presented in that form in the file. Words that may be inflected as part of the phrase are presented in inflected forms (at the time being in different entries, but this will soon be concentrated into one entry). A wild card (*) has been inserted in which an additional word can be inserted (this will be expanded to phrases), e.g. adverbials as for instance *not.* Each phrase (record) is prefixed with one of the words in the phrase. That word acts as a key to the phrase. This means that the keyword is also listed in the bilingual word list with a code that tells the parser to look into the phrase list. This keyword is usually chosen among words that is regarded as having low frequency in text, to prevent too many superflous searches. To the right of the phrase we find the NN synonym, prefixed with a dot ('.'). An entry may look like this (v0 is the inflection code for irregular verbs):

```
behov   ha v0 * behov for         .trenge v0
```

9 Word Formation, Compound Words

A module has been implemented for the analysis and translation of compound words. The main principle is that compound words that do not exist in the bilingual word list may be analysed as being composed of words in the list. The processing of word formation has two important types of restrictions.

One type of restriction is that words that are part of a compound word must usually be the so-called form 1, i.e. infinitive of verbs, indefinite singular of nouns and common gender (i.e. masculine and feminine) of adjectives. But the rightmost (last) part of the word may of course be inflected. At the time being these restrictions are included in the program code.

The other type of restriction concerns what part of speech the different word parts can come from. An adjective and a noun may be combined, like *rødvin* (red wine). But we cannot combine a noun and an article, like *gutten* (the boy) read as *gutt* + *en* (boy + a) and will be translated to *guttein.* In the system these types of restrictions are written into a data file where patterns express illegal sequences of parts of speech.

10 Semantic Classification

As mentioned above, a module for semantic classification to guide the parser has been included. The system is implemented as a thesaurus in a tree structure. The raw data is written into a text file with expression for the subordinate concept in the left column and the expression for the superior concept in the right. From these data the program builds the thesaurus structure.

The parser utilizes this information to decide if two words or phrases can be unified according to the semantic classification. This is of importance when the two words/phrases have semantic representation at different levels in such a way that one of the representations is subordinated to the other. That means that they are compatible.

11 The Morphological Analysis for Single Stem Words

The main component of the word analysis is recognition of words in the bilingual word list and the related inflection endings in the inflection tables. In the analysis of compound words this information is combined in repeated word-recognition during the whole compound word.

The general strategy for word recognition works from right to left. The program first searches in the word list for the whole word. If it is not found, a new search is performed with the rightmost character deleted. The search process is repeated with deletion of the rightmost character until there will be a match or there will be no search string left. If the input word is *kaster* the program will search for

kaster,
kaste, and then
kast,

which is a verb stem in the list. Kast has two different inflection codes, v6 and v7.

The program will expand the stem with endings from both the inflection paradigms in order to find identity between the complete input word and one or more inflected forms of the word list entry. The paradigms are:

v6	e	er	a	a
v7	e	er	et	et

According to this the stem *kast* and inflection endings produce these word forms:

v6:	kaste	kaster	kasta	kasta
v7:	kaste	kaster	kastet	kastet

The second word forms, i.e. the present tense in both the paradigms match. That means that the program has identified *kaster* as present tense of the stem kast.

12 The Morphological Analysis of Compound Words

The analysis of compound words follows the same main strategy as for single stem words. The main difference is repeated searches for stems, endings and joining characters. If the input word is *lærerhøyskolestudenter* (Teachers' Training College Students, read stem by stem as: 'teacher + high + shool + student + s') the program will delete the rightmost letters until it matches *lær*. That means that searches will be done for the following character sequences:

lærerhøyskolestudenter

lærerhøyskolestudente

lærerhøyskolestudent

.

lærerh

lærer

lære

lær

The matched stems in the word list are the noun stem *lær* with the suffix *-er* and the two masculine inflection codes m3 and m4, and the verb stem *lær* with the inflection code v1. The program first expands these two stems with the three paradigms, like this:

m3:	lærer	læreren	lærere	lærerne
m4:	lærer	læreren	lærerer	lærerne
v1:	lære	lærer	lærte	lært

None of these word forms match the whole (compound) word *lærerhøyskolestudenter*. Then we have to activate the module for analysis of compound words. As stated above, according to our rules, words which are part of compound words can only occur in form 1. In this case there are three candidates that matches the beginning of the compound word, of which two word forms are unique: *lærer* and *lære*. This means that the rest of the word is either *høyskolestudenter* or *rhøyskolestudenter*. The program tries both strings. It fails for the last one, but will succeed for the first. The identified wordforms as parts of the compound word are *høy*, *skole* and *studenter*. The program will also propose the noun *student* and the copula verb *er* instead of studenter. This solution will be excluded because of the rules which imply that part of speech sequences are not allowed in compound words.

This word will be translated to NN *lærarhøgskulestudentar*.

13 Parsing Strategy

The parsing strategy is bottom-up. It is based on binary rules like those listed under "Syntax rules". The rules consists of two conjoining (a pair) syntactic labels and restrictions for their syntactic features (gender, definitness, number, tense etc). The parsing module per June 1989 is a preliminary one. The details of the syntactical and semantical description and procedures will be radically revised and accomodated to a methodology based on a Lexical Functional Grammar (LFG).

References

Myking, Johan. 1989: Term Harmonization, 'Selective Purism' and Language Autonomy. An "intra-national" case. To be printed in the proceedings from The 7th European Symposium on LSP, Budapest, 21.–26. August 1989.

Strømgaten 53
N-5007 BERGEN
Norway