Norwegian Computational Lexicon (*NorKompLeks*)

Torbjørn Nordgård

Dept. of Linguistics
Norwegian University of Science and Technology (NTNU)
Trondheim
Norway

Background

There is no generally accessible lexicon for the Norwegian language. This situation prevents development of general language engineering applications for this language. In addition, linguists cannot use modern lexical tools in their research. Given this background, the lexicon project «Norsk komputasjonelt leksikon» (Norwegian Computational Lexicon), abbreviated as *NorKompLeks*, has the following goals:

1. Create a morphological lexicon for the written standard Bokmål
2. Create a morphological lexicon for the written standard Nynorsk
3. Provide phonological descriptions for both standards
4. Describe argument structures for relevant lexical items (verbs, prepositions, certain nouns)

«Bokmålsordboka» (65000 items) and «Nynorskordboka» (90000 items), both owned by Section of Lexicography at the University of Oslo, define the coverage of the two lexicons. These dictionaries contain the «official» lexicographic inventory of Norwegian, as defined by Norsk Språkråd (The Norwegian Language Council). In what follows the results of the project will be briefly described, and special attention is given to argument structure descriptions and how lexical descriptions can be converted into linguistic formalisms.

The project «Norsk komputasjonelt leksikon» (Norwegian Computational Lexicon) is funded by the Norwegian Research Council (NFR) and Telenor (The National Telematic Company).[1] The project period is from January 1996 through December 1998. The partners are

- Linguistics Department, NTNU, Trondheim
- «Dokumentasjonsprosjektet» (a national documentation project within the humanities), University of Oslo
- Computing Centre for the Humanities, University of Bergen
- Telenor

[1] Project assistants: Jardar Eggesbø Abrahamsen (Bokmål and Nynorsk morphology), Bodil Aurstad (Bokmål morphology and argument structure), Kristin Eide (Nynorsk morphology), Bente Moxness (Phonology), Eli Sætherø (Argument structure).

Morphology

The morphology of the written standard Bokmål (graphemic base forms and morphological paradigms) was finished in December 1996. The format of the lexicographic entries in Bokmålsordboka is

NB00*n*	Base Form
NB00*n*a	Inflectional paradigm(s)
NB00*n*b	Version number (in case of homonomy)
ARTNR	Identification label (an integer)
TR007	Lexicographic information (etymology, definitions, examples, etc.)

An example from the Bokmål source (irrelevant typographical codes included, i.e. $C, $B, @, etc.):

```
NB001  fremme
NB001a v24,v24a,v25
NB001b 2
ARTNR  17805
TR007
..OPP  #>$CII fremme@ v1
..ETY  (norr $Bfremja@, av $Bfram@)
..DEF  $C1@ hjelpe fram, øke, påskynde, stimulere
..UTR  $Bf- en sak, et formål !@
..UTR  $Btiltak som f-r kommunens økonomi !@
..UTR  $Bf- salget av, interessen for !@
..DEF  adj i pr pt:
..UTR  $Bvirke f-nde på@
..FOR  stimulerende
..DEF  $C2@ legge fram til behandling, ta opp, reise
..UTR  $Bf- et forslag !@
..UTR  $Bf- en sak for Stortinget, retten@
..DEF  $C3@ sette i verk, gjennomføre
```

In the more compact computational lexicon the morphological information is described, in Prolog, as

```
bm(fremme,[v24,v24a,v25],17805).
```

Lexicographic information is removed, but can be obtained via the numeric identifier *17805*. *v24,v24a* and *v25* are morphological inflection codes which describe inflection patterns for the lexical item *fremme* («propose»). After expansion by these codes, we get the following paradigm, in accordance with Faarlund, Lie & Vannebo (1996):

```
nkl_ff(frem,[verb,imperativ,aktiv,hovedverb,17805,fremme]).
nkl_ff(fremma,[adj,mfn,pos,best,pl,17805,fremme]).
nkl_ff(fremma,[adj,mfn,pos,best,sg,17805,fremme]).
nkl_ff(fremma,[adj,mfn,pos,ubest,pl,17805,fremme]).
nkl_ff(fremma,[adj,mfn,pos,ubest,sg,17805,fremme]).
nkl_ff(fremma,[verb,perf_part,indikativ,aktiv,hovedverb,17805,fremme]).
```

```
nkl_ff(fremma,[verb,perf_part,indikativ,passiv,hovedverb,17805,fremme]).
nkl_ff(fremma,[verb,pret,indikativ,aktiv,hovedverb,17805,fremme]).
nkl_ff(fremme,[verb,infinitiv,indikativ,aktiv,hovedverb,17805,fremme]).
nkl_ff(fremmede,[adj,mfn,pos,best,pl,17805,fremme]).
nkl_ff(fremmede,[adj,mfn,pos,best,sg,17805,fremme]).
nkl_ff(fremmede,[adj,mfn,pos,ubest,pl,17805,fremme]).
nkl_ff(fremmende,[adj,mfn,pos,best,pl,17805,fremme]).
nkl_ff(fremmende,[adj,mfn,pos,best,sg,17805,fremme]).
nkl_ff(fremmende,[adj,mfn,pos,ubest,pl,17805,fremme]).
nkl_ff(fremmende,[adj,mfn,pos,ubest,sg,17805,fremme]).
nkl_ff(fremmende,[verb,pres_part,indikativ,aktiv,hovedverb,17805,fremme]).
nkl_ff(fremmer,[verb,presens,indikativ,aktiv,hovedverb,17805,fremme]).
nkl_ff(fremmes,[verb,presens,indikativ,passiv,hovedverb,17805,fremme]).
nkl_ff(fremmet,[adj,mfn,pos,ubest,sg,17805,fremme]).
nkl_ff(fremmet,[verb,perf_part,indikativ,aktiv,hovedverb,17805,fremme]).
nkl_ff(fremmet,[verb,perf_part,indikativ,passiv,hovedverb,17805,fremme]).
nkl_ff(fremmet,[verb,pret,indikativ,aktiv,hovedverb,17805,fremme]).
nkl_ff(fremmete,[adj,mfn,pos,best,pl,17805,fremme]).
nkl_ff(fremmete,[adj,mfn,pos,best,sg,17805,fremme]).
nkl_ff(fremmete,[adj,mfn,pos,ubest,pl,17805,fremme]).
```

(best=definite, ubest=indefinite,mfn=masculine,feminine and neuter, hovedverb= main verb, perf_part=past participle, pos=positive, pres=present tense, pret=pase tense, pres_part=present participle)

This expansion is designed so that it meets the requirements of a tagger being developed at the University of Oslo. The expansion can be defined differently, for instance with word class information only (the details of this process will be explained below):

```
nkl_ff(frem,[verb]).
nkl_ff(fremma,[adj]).
nkl_ff(fremma,[verb]).
nkl_ff(fremme,[verb]).
nkl_ff(fremmede,[adj]).
nkl_ff(fremmende,[adj]).
nkl_ff(fremmende,[verb]).
nkl_ff(fremmer,[verb]).
nkl_ff(fremmes,[verb]).
nkl_ff(fremmet,[adj]).
nkl_ff(fremmet,[verb]).
nkl_ff(fremmete,[adj]).
```

Full form expansion is controlled by the information connected to the inflection codes. Consider the code *v24* (from the set of paradigm codes for the verb *fremme*):

```
code(v24,
        [inf_v:>0,              % fremme      (infintive)
        imp_v:>[1:0,1:0],       % frem        (imperative)
        pres_v:>"r",            % fremmer     (present)
        pret_v:>"t",            % fremmet     (past)
        p_part_v:>"t",          % fremmet     (past participle)
        pr_part_v:>"nde",       % fremmende   (present participle)
        pass_part_v:>"t",       % fremmet     (passive participle form)
```

```
s_pass_v:>"s",          % fremmes    («s-passive»)
pos_ub_sg_v:>"t",       % fremmet    (adjective,positive,indefinite,singular)
pos_b_sg_v:>"de",       % fremmede   (adjective,positive,definite,singular)
pos_pl_v:>"de",         % fremmede   (adjective,positive,plural)
pos_v:>"nde"]).         % fremmende  (adjective,positive)
```

The expression *inf_v:>0* means that the *inf_v* (the infinitival form) is achieved if nothing is done to the lexical stem. The imperative is generated by deleting the two rightmost characters of the stem, the present tense requires that the string «r» is added to the stem, and so on. Thus, each generated form is tied to a morphosyntactic code (*inf_v, imp_v, ...*). These codes must be given an interpretation according to the requirements of the lexicographer, linguist or system designer who is going to use the lexicon. The symbol pres_v has the interpretation verb, present tense, indicative, active, main verb. In Prolog:

```
m_kode(pres_v,[verb,presens,indikativ,aktiv,hovedverb]).
```

Note that the list [verb,presens,indikativ,aktiv,hovedverb] is a sublist of [verb,presens,indikativ,aktiv,hovedverb,17805,fremme], i.e. the morphological information associated with the present form *fremmer* in the example above. By changing the interpretation of the code *pres_v* new lexical descriptions can be generated, for instance in accordance with TEI or EAGLES specifications. The same is true of all morphological codes in the lexicon. Thus, the codes is a method for describing the properties of morphological distinctions in Norwegian, and the descriptions can be adapted to various theories and formalisms, simply by changing the interpretation of the codes. A trivial example is the alternative where word class information is the only information, as in the example above:

```
m_kode(pres_v,[verb]).
```

The same coding system will be used for the Nynorsk material, and this work will be finished in the spring of 1998.[2]

Phonology

The project has completed phonological descriptions of approx. 65.000 Bokmål entries. The descriptions are written in SAMPA. Consider some examples (*begynne* = begin):

fremme	""frem@	V01	17805
begynne	b@"jyn@	v21	4974 .

Observe that stress and tones are marked: A single " means stress with toneme 1, and double "" signals stress with toneme 2.[3]

[2] The derivation of «verbal adjectives» is different in the Nynorsk lexicon where a verb licenes an adjective base form together with an adjective code. This pair serves as the basis for generating a normal adjective paradigm. The description of verbal adjectives in the Bokmål material will be changed to this format when time and resources are available, hopefully in the fall 1998.

The phonological descriptions are based on the most common pronunciation patterns in the South Eastern parts of Norway, where most Norwegians live. In controlling the phonological representations the developers use an automatic speech generation system developed by Telenor Research.
As in the morphology phonological paradigms will be generated, using the same description format. Thus, when the project is finished phonological representations will be available together with morphological forms, both for the base entries in the lexicon and in the expanded («full form») lexicon.

Phonological coding of nynorsk entries has started and will be finished in 1998. Generation and control of phonological paradigms will be accomplished in 1998.

Argument Structure in NorKompLeks

There is a well-known tension between linguistic sophistication in lexical descriptions and the time available for developing a lexicon of an adequate size. An important question in *NorKompLeks* is how to describe argument structure in a linguistically sound way, and at the same time being able to make these descriptions in a finite amount of time, i.e. 20 months work for one person.

It is very important for the reusability of a lexicon that theoretical adaptation can be achieved without rebuilding the entire lexicon. Consequently, the lexicon should be theory neutral (without being anti-theoretically). The project builds on previous lexicons for Norwegian: the *TROLL* lexicon[4] and *NorLex* verb lists (University of Bergen, 1992 - 1993).

Absolute theory independence is impossible (and undesirable). But theoretical «idiosyncrasies» and «hang-ups» must be avoided, for instance linguistic theories which insist that reference to grammatical functions is irrelevant in the lexicon, as in Government and Binding and Minimalist approaches, see e.g. Chomsky (1981, 1986) and Chomsky (1994). A criterion for success is that the lexicon can be used to generate argument descriptions in various theories, for instance Lexical-Functional Grammar (see e.g. Dalrymple, Kaplan, Maxwell & Zaenen 1995), Head-Driven Phrase Structure Grammar (see Pollard and Sag 1994) and GB-type argument descriptions (see the references above), by some limited adaptations.

The argument structure for a verb must contain information about the basic *construction types* that the verb can be engaged in. This information can be encoded at various levels of abstraction, and linguistic theories will typically tend to represent this information in a compact format which can be interpreted by theory internal mechanisms. A computational lexicon with reusability ambitions should take the opposite approach: Describe the construction types transparently, and show how these construction types can be interpreted by linguistic theories.

---

[3] The code V01 is an original morphological code from Bokmålsordboka. This code will be replaced by new morphological codes (v24,v24a,v25) and phonological codes (work in progress).

[4] See Hellan and Johnsen 1988.

Consider some intransitive construction types with examples:

- Intransitive verb with expletive subject
  Det regner (It rains)
- Intransitive verb with expletive subject and required adverbial
  Det kvakk i henne («It suddenly-surprised her», 'she was suddenly surprised by something')
- Intransitive verb with agentive subject
  Studentene tenker (The students think)
- Intransitive verb with experiencer subject
  Gutten fryser (The boy is freezing)

A construction type is characterised by its obligatory arguments. Each construction type is given a unique label. Thus, the construction type «Intransitive verb with expletive subject» is called *nullv* (meaning *nullverdig* or «zero valency»). The arguments of a construction type is described as a triple with information about syntactic function, thematic role and categorial realisation. The code *nullv* is interpreted as

```
arg_code(nullv,[arg1:su::norole::np]).
```

This simply means that a verb with this code can take a subject NP with no thematic role, i.e. an expletive subject. The code for intransitive verb with agentive subject is

```
arg_code(intrans1,[arg1:su::ag::np]).
```

That is, a construction with a subject NP which has the thematic role agent.

Passivization is a regular process in most languages, but certain verbs can not enter the passive construction. A well-known example in the syntactic literature is so-called ergative verbs, like *arrive*. In NorKompLeks passivization possibilities are tied to construction types. Therefore, intransitive verbs of the «arrive»-type has the label *intrans2* which is composed as

```
arg_code(intrans2,[arg1:su::th::np,--passiv]).
```

That is, a verb with a thematic subject. The tag «--passiv» signals that this construction type can't undergo a lexical passivization process, whereas the type *intrans1* does not have this restriction. The tag «--passive» is in a sense redundant, given that it is empirically true that verbs which take a thematic subject cannot be passivized, but this tag makes the codes more explicit and thus reduces the risk of wrong code assignments.

The passivization procedure can be defined as

If a verb has the code intrans1, it also has the possible realisation [arg1:su::norole::np] when it appears with passive morphology.
Consider two codes for transitive constructions:

- arg_code(trans1,[arg1:su::ag::np,arg2:obj::th::np]).
  Example: dekomponere (*decompose*)
- arg_code(trans2,[arg1:su::ag::np,arg2:obj::th::s1]).
  Example: dokumentere (*to document*, with clausal complement)

Note that each argument in the construction is represented as a triple, but the argument structure is conceived as a *list* of such triples.[5]

Control verbs require that the number of arguments is specified together with the controller of the infinitival subject. This is done as follows in NorKompLeks:

- arg_code(trans3,[arg1:su::ag::np,arg2:obj::th::inf, arg2:su=arg1:su]).
  Example: prøve (*try*)

The expression *arg2:su=arg1:su* says that the subject of the second argument is identical to the subject of the first argument, i.e. subject control.

Let us now turn to some examples from the argument structure descriptions:

```
w(dages,9088,[nullv]).            (dawning, «The day is dawning»)
w(dampe,9220,[trans1,intrans2]). (steam)
w(dandere,9237,[trans1]).         (shape, arrange)
w(dangle,9242,[intrans2,trans1]).(dangle, swing)
w(danse,9265,[intrans1,trans1]). (dance)
w(danske,9279,[intrans1]).        (speak Danish affectedly)
w(dirre,9285,[intrans2]).         (quiver, vibrate)
```

Note that a verb can enter into a set of construction types, as the verb *danse* (dance) which can be used transitively or as a standard intransitive verb.

Interpretation of Argument Structures in Grammatical Frameworks

The argument structure information in *NorKompLeks* satisfies the basic requirements of current generative theories. «Classical» LFG, i.e. the 1982 version, needs information about syntactic function in order to specify the argument structure portion of the lexical descriptions. More recent versions of LFG has argument linking information where syntactic functions and thematic roles are connected. NorKompLeks seems to have the required information for both versions of LFG. A preliminary version of a compiler which translates argument descriptions in *NorKompLeks* into LFG-82 descriptions has been developed. The Prolog session below illustrates the idea of translating *NorKompLeks* descriptions of verbs into LFG-82 format (the information is given in Prolog syntax). The program consults the lexical descriptions of a word form (e.g. *bable*), translates the code(s) into LFG-82 format, and returns the translation. The usual LFG notation is also given.

[5] Such a list is of special interest for subcategorization in HPSG.

27 ?- translate_word(bable,Trans,lfg82). *bable = babble*
Trans = [[up pred = bable(up subj)]]

LFG notation: (↑ PRED) = 'bable<(↑ SUBJ)>'

28 ?- translate_word(baktale,Trans,lfg82). *baktale = slander, speak ill of*
Trans = [[up pred = baktale(up subj, up obj)]]

LFG notation: (↑ PRED) = 'baktale<(↑ SUBJ), (↑ OBJ)>'

29 ?- translate_word(boble,Trans,lfg82). *boble = bubble*
Trans = [[up pred = boble(up subj)], [up pred = boble(up subj)]]

LFG notation: (↑ PRED) = 'boble<(↑ SUBJ)>'

30 ?- translate_word(beta,Trans,lfg82). *beta = impress, fascinate*
Trans = [[up pred = beta(up subj, up obj)]]

LFG notation: (↑ PRED) = 'beta<(↑ SUBJ), (↑ OBJ)>'

31 ?-translate_word(begynne,Trans,lfg82). *begynne = begin*
Trans = [[up pred = begynne(up subj)], [up pred = begynne(up subj, up obj)],
[up pred = begynne(up subj, up vcomp), up subj = (up vcomp, subj)]]

LFG notation: [(↑ PRED) = 'begynne<(↑ SUBJ)>']

[(↑ PRED) = 'begynne<(↑ SUBJ), (↑ OBJ)>']

[(↑ PRED) = 'begynne<(↑ SUBJ), (↑ VCOMP)>'
(↑ SUBJ) = (↑ VCOMP SUBJ)]

The last version of the verb *begynne* (begin) has a control equation which states that the matrix subject is the same as the subject of the vcomp. Observe that translation 29 in the Prolog session gives two identical argument structures, which means that the lexicon makes a distinction that this LFG compiler doesn't.

Lexical descriptions in HPSG need information about categorial realisation and ordering among the arguments. The relevant ordering is implicitly encoded (arg1 is the subject, arg2 is the object, etc.). The semantics of lexical items HPSG is not provided by *NorKompLeks*, simply because this semantic information is more or less internal to HPSG.

Government and Binding approaches to syntax is not widespread in computational linguistics, but GB grammars (and Minimalist approaches) are popular among syntacticians. Because the *NorKompLeks* lexicon is meant to be useful also for syntacticians who are not NLP practitioners, GB translations of lexical descriptions has some interest. Lexical descriptions in GB require information about thematic roles, categorial information and whether an argument is «external»

or «internal». The former two types of information is encoded directly, but the «external» / «internal» dichotomy is present indirectly. Agentive subjects are «external», but thematic subjects are «internal». Thus, the code *intrans1* has one external argument in GB terms, and *intrans2* has one internal argument. Some examples of the translation program:

```
22 ?- translate_word(bable,Trans,gb).
   Trans = [bable lex_gb verb args [ag : np - 'E']]
```

The expression *bable lex_gb verb args [ag : np - 'E']* means that *bable* is a verb with one external agent argument which is realized as NP.

```
23 ?- translate_word(baktale,Trans,gb).
   Trans = [baktale lex_gb verb args [ag : np - 'E', th : np]]
```

Here the verb baktale has two arguments, an external agent NP and an internal thematic NP.

```
24 ?- translate_word(boble,Trans,gb).
   Trans = [boble lex_gb verb args [],  boble lex_gb verb args [th : np]]
```

The empty list signals a verb with no theta-marked arguments (the first version of *boble*). The second version of the verb has one argument which is internal, which is the theoretical description of an ergative verb.

The verb *beta* has a thematic external argument and an internal argument with the role experiencer:

```
25 ?- translate_word(beta,Trans,gb).
   Trans = [beta lex_gb verb args [th : np - 'E', exp : np]]
```

The actual translation from NorKompLeks descriptions to GB descriptions is governed by Prolog predicates like

```
arg_translation(gb,[arg1 : su :: ag :: np],  [ag : np - 'E']).
```

and

```
arg_translation(gb,[arg1 : su :: ROLE :: np, -- passiv],  [ROLE : np]).
```

The third arguments, i.e. [ag : np - 'E'] and [ROLE : np] are GB-translations of NorKompLeks codes. Recall that [arg1 : su :: ag :: np] is the interpretation of the code *intrans1*, and the verb *bable* has this code. ROLE is a Prolog variable which has the effect of transferring the NorKompLeks theta role onto the GB description. LFG translations is accomplished by the same method, but with other interpretations, of course:

```
arg_translation(lfg,[arg1:su::_ARG1ROLE::np], up pred = [up subj]).
```

Note that the thematic role is of no interest here, and it is made invisible by the anonymous Prolog variable _ARG1ROLE.

Argument Information and Morphological Information

The verb

w(debutere,9434,[intrans1]). *(make one's debut)*

has an identifier 9434 which is a «pointer» into the stem lexicon. The stem lexicon has information about morphological properties (conjugation classes and spelling). Generating full forms with morphosyntactic (see the morphological section above) and argument structure information is a rather trivial matter, but the details are of course dependent upon theoretical and formal considerations.

Conclusions

The basic ingredients of the morphological, phonological and syntactic-semantic properties of the *NorKompLeks* lexicon has been described. Most of the work is finished. Today the lexicon is used at the University of Oslo in a rule-based POS-tagger, and it will be used in the Norwegian part of the ongoing EU-project SCARRIE. TEI interpretations of the morphological and syntactic-semantic information will be made later this year.

References

Hellan, L., Johnsen, L. and Pitz, A. : *TROLL* (The Trondheim Linguistic Lexicon Project). Ms, Dept. of Linguistics, NTNU, Trondheim.

Bresnan, J., ed. (1982): *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, Mass.

Bresnan, J., and R. Kaplan (1982): Introduction. In Bresnan (1982).

Chomsky, N. (1981): *Lectures on Government and Binding*. Foris, Dordrecht.

Chomsky, N. (1986): *Knowledge of Language. Its Nature, Origin and Use*. Praeger, New York.

Chomsky, N. (1995): *The Minimalist Program*. MIT Press, Cambridge, Mass.

Dalrymple, M., R.M. Kaplan, J.T. Maxwell, A. Zaenen (1995): *Formal Issues in Lexical-Functional Grammar*. CSLI Publications, Stanford, CA.

Faarlund, J.T., S. Lie and K.I. Vannebo (1997): *Norsk Referansegrammatikk*. Universitetsforlaget, Oslo.

Pollard, C. and I. Sag: *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago.