Encoding Danish Verbs in the PAROLE Model

Costanza Navarretta
Center for Sprogteknologi
Njalsgade 80
2300 Copenhagen S
DK - Denmark
costanza@cst.ku.dk

Abstract

This paper describes the work of specifying and encoding an NLP lexicon for Danish as part of a current EU project concerned with the construction of linguistic resources for European languages. In particular we outline how the common European PAROLE model has been adapted to the syntax of Danish verbs and we present the first results from the analysis and encoding of approx. 3000[1] general language verbs.

1 Introduction

The work presented in this paper is part of the Danish lexicon contribution to the current EU-funded MLAP project LE-PAROLE the aim of which is to construct linguistic resources (large corpora and electronic lexica) for EU languages. The main objective of the lexical part of the project is to build monolingual electronic lexica for the European languages following a common model, the PAROLE model. The strategy is to reuse as many as possible existing linguistic resources and construct reusable lexica, i.e. lexica which can be used by different applications and systems. The PAROLE lexica are encoded in SGML format and are based on the corpora collected by the PAROLE corpus groups. However, in this paper we will concentrate on the lexical aspects of the project.

We first sketch out the PAROLE model and we explain the main guidelines followed in constructing the Danish lexicon[2]. Then we describe how the syntax for Danish verbs has been encoded in the common model and we discuss the problems the model presents. Finally we make some conclusions and remarks on the PAROLE model and lexicon on the basis of the entries encoded up til now.

[1] 3000 syntactic units corresponding to approx. 2900 verbal lemmas.

[2] Other members of the Danish lexicon project are Anna Braasch, project responsible, and Nicolai Hartvig.

2 The PAROLE Model

The output of two previous European projects, EAGLES[3] and GENELEX[4], were taken as a starting point for constructing the PAROLE lexica. The GENELEX-EAGLES model was adapted so that it could account for the linguistic phenomena of all the involved languages[5] and the result is the so-called PAROLE model. It defines a generic model for reusable lexica, i.e. it does not follow a single linguistic theory and it is designed so that lexica conformant to it contain the basic information needed by different NLP applications. All the language groups are using the same tool constructed by the French partner in the GENELEX project and adapted to the new model. The tool contains also devices to convert SGML lexica, conformant to the PAROLE DTD, to and from the internal database and a browser.

The model[6] distinguishes three descriptive levels, morphosyntax, syntax and semantics. Each level requires elementary and obligatory types of information and permits more refined optional types too. At the morphology level the main entities are Simple Morphological Units (**MuS**[7]) containing basic information on orthography and inflection. One or more Syntactic Units (**SynU**) are linked to each **MuS**. **SynUs** are the principal entities of the syntactic level and contain information about the syntactic behaviour of lexical units. At the semantic level the main entities are the Semantic Units (**SemU**).

A lexical unit is thus described by a morphological unit and the associated syntactic and semantic units.

[3] The EAGLES project focused on the identification of the minimal observable facts in morphosyntax and syntax common to different generalizations of different theories and devices of different systems in order to achieve a common set of consensual and broadly agreed specifications (Calzolari & Zampolli 94). The specifications for morphosyntax and syntax can be found in (Monachini & Calzolari 96) and (Sanfilippo *et al.* 96) respectively.

[4] The GENELEX project had as aim to develop and implement a model of generic lexicons for NLP (GENELEX Consortium 93).

[5] The following languages are currently costructing PAROLE lexica: Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish, Swedish.

[6] For details the reader is referred to (Calzolari ed. 1996).

[7] The acronym derives from the original French name.

The current project only deals with the encoding of the morphological and the syntactic levels. Thus at present a lexical unit corresponds to a syntactic unit and its associated morphological unit.

At the end of the project each lexicon will contain 20,000 **MuSs** and the **SynUs** related to them.

2.1 The Syntactic level

The main information at the syntactic level is subcategorization, including functions of subcategorized for complements and (optionally) thematic roles, characteristics of the lexical unit when associated with a specific subcategorization frame, control, diathesis alternations, linear order constraints.

Subcategorization is the basic information for major word classes. It is described by at least one **Description** associated to one **SynU**. One **Description** can be linked to many **SynUs**. A **Description** is composed by a **Self**, the actual lexical entry, and by a **Construction** where complements are described by **Positions** directly surrounding the lexical entry. Each **Position** is defined as obligatory or optional.

Positions are realized by terminal or non-terminal categories. Typed-features can be added to the basic category information. A **Position** is characterized by its distribution and its function.

Characteristics of a lexical unit are encoded in the **Self** by means of features. The **Self** may or may not be the syntactic head of the actual **Description**.

Syntactic alternations (e.g. dative shift alternation) are handled by a link between two **Descriptions** and their **Positions**.

During the project, the different language groups have been harmonizing their work so that similar phenomena are treated in the same way. However, differences remain due to the variety of the languages (and family of languages), to the different linguistic traditions and to the variety of the linguistic resources which each group has reused.

3 The Danish Lexicon

The 20,000 **MuSs** (lemmas) of the Danish lexicon are based on the Official Danish Spelling Dictionary "Retskrivningsordbogen" (Dansk Sprognævn 86). The linguistic resources which could be (re)used at the syntactic level are the EUROTRA Danish Lexicon (Jørgensen & Kirkchmeier-Andersen 91), the LSGRAM Danish lexicon (Music & Navarretta 96)[8] and the first version of the Odense Valency Dictionary (Schøsler & Van Durme 96).

The lexicon is a general language dictionary. The 20,000 **MuSs** cover besides all closed word classes, general language words chosen according to their frequency count, based on the general language corpus of the Danish Dictionary (DDO)[9].

The Danish PAROLE morphology[10] is strictly inflectional and does not account for the syntactic function of the lemmas. Thus, as an example, at the morphological level modal, auxiliary and main readings of a verb are not distinguished. These distinctions are added to the lexicon at the syntactic level in the form of language specific features.

4 Syntax for Danish Verbs

Although none of our available linguistic resources, the EUROTRA Danish Lexicon, the LSGRAM Danish lexicon and the first version of the Odense Valency Dictionary, could be directly converted to the PAROLE lexicon, some interesting data such as case-marking prepositions and simple control could be reused for many of the 500 verbal lemmas. The subcategorization frames in EUROTRA, although simpler than those definable in PAROLE have also been a valuable starting point for the encoding of **Constructions**.

4.1 Syntactic Functions

The PAROLE model defines a closed list of syntactic functions. Some are language specific, others are common to all the languages. The syntactic functions used for Danish verbal subcategorization frames are the following: SUBJECT, OBJECT, INDIRECT OBJECT, OBLIQUE, ADVERBIAL, SUBJECT PREDICATE, OBJECT PREDICATE, REALSUBJ, CLAUSECOMPL, PREPDEPENDENT. SUBJECT, OBJECT, INDIRECT OBJECT, SUBJECT PREDICATE and OBJECT PREDICATE do not need any special remark. OBLIQUE and ADVERBIAL distinguish the function of other complements, mainly but not exclusively prepositional phrases. The former, obliques, are usually prepositional phrases introduced by case marking prepositions and correspond to the following thematic roles: THEME, AGENT, SECOND PARTICIPANT, INSTRUMENT. The latter, adverbials, are often adverbs or prepositional phrases, which (in many cases) can be replaced by an adverb or introduced by weakly bound prepositions. They can be assigned the following thematic roles: SOURCE, GOAL, LOCATION, MEASURE. Adverbials include measure positions which formally are objects, such as

> *Han vejer halvfjerds kilo*
> (He weighs seventy kilos).

The function CLAUSECOMPL is assigned to infinitives subcategorized for by modal verbs, and to subject clauses which are introduced by the expletive *det* (it), e.g.

[8] Actually the LSGRAM Danish lexicon was for the main part semi-automatically converted from the EUROTRA lexicon, but its construction is closer to the PAROLE model than the EUROTRA lexicon.

[9] The DDO contains 40 million tokens.

[10] The Danish morphology has been specified by Ole Norling Christensen and Anna Braasch and has been provided by the former who extracted and converted entries from the Danish Dictionary word base.

Det angår ikke mig at han er rejst
(It does not concern me that he has left)

The function REALSUBJ is assigned to deep subjects introduced by an expletive. The function PREPDEPENDENT is not a verbal one, but it is assigned to the complements following a preposition.

4.2 Optionality of Positions and Alternation of Syntactic Constructions

We have chosen to encode both optionality of **Positions** and alternation of Syntactic Constructions.

We encode optionality of complements, with the restrictions posed by the PAROLE model, i.e. when no constraints apply to the reciprocal order and to the undissociability of some **Positions**. E.g. the two following cases are covered by the same syntactic construction:

Sidste år skrev jeg mange breve til dig
(Last year I wrote many letters to you)

and

Sidste år skrev jeg mange breve
(Last year I wrote many letters)

We encode in the same **SynU** alternation of syntactic constructions having the same syntactic function. The most common cases are objects or prepositional complements being realized as NPs or as subordinate clauses (i.a. infinitive, that-clauses, interrogative clauses) as can be seen in the following two examples:

Han elsker film/ at se film/ at du går med ham i biografen
(He loves films/to watch films/that you go with him to the cinema)

Jeg tvivler på dig/ på at han vil komme
(I doubt you/that he will come (lit.on that he will come)

4.3 The Self

The following **Selfs** have been recognized:

- main verbs taking *have*(have)/*være*(be) as auxiliary (one for each auxiliary)
- main verbs in passive form (i.e. verbs which in (Dansk Sprognævn 86) are encoded as active verbs, but which have different active readings in passive form[11]
- modal verbs
- auxiliary verbs taking *have*(have)/*være* (be) as auxiliary (one for each auxiliary)
- phrasal verbs taking *have* (have) as auxiliary (an entry for each particle)
- phrasal verbs taking *være* (be) as auxiliary (an entry for each particle)
- reflexive verbs (e.g. *beslutte sig til noget* (decide on something)
- reflexive phrasal verbs (an entry for each particle)

The auxiliary taken by a verb and the verbal type (MAIN, MODAL, AUXILIARY) are encoded by adding appropriate syntactic features to the **Self** while the passive form is encoded by the morphological feature PASSIVE. Reflexive and phrasal verbs are composed of two terminals, a main verb followed by a reflexive pronoun and a particle respectively.

We distinguish between verbs with reflexive constructions, such as

jeg vasker mig
(I wash myself)

and reflexive verbs, i.e. verbs where the reflexive form has another meaning and/or construction than the corresponding non-reflexive one e.g.

han brækkede sit venstre ben
(he broke his left leg)
han brækkede sig
(he vomited)

Only the latter type is treated as a reflexive **Self**, the former is just a main verb subcategorizing for a reflexive pronoun object.

Following (Quirk *et al.* 1985; Allan *et al.* 1995) we also distinguish between phrasal and prepositional verbs. Only the former are treated at the **Self** level.

4.4 Control

Control information is obligatory in syntax, and whenever it is possible to determine control syntactically, this is encoded. The PAROLE model provides for the encoding of both binding constraints and types of control (equi and raising constructions), as can be seen in the following examples:

Peter hader at svømme
(Peter hates to swim)

The subject of the matrix verb "hader" is also the subject of the subordinate clause "at svømme" (subject control construction).

Peter overtalte hende til at hente kaffen
(Peter convinced her to get the coffee)

The object of the matrix verb "overtale" is the subject of the prepositional dependent subordinate clause "at hente kaffe" (object control construction).

Peter synes at være træt
(Peter seems to be tired)

[11] E.g. *Såret heler/heles* (The wound is healing), where *heles* is in passive form.

The formal subject of the matrix verb "synes" is the subject of the subordinate clause "at være træt" (raising construction). Also possible restrictions on the subject of subordinate that-clauses are marked, e.g.

Jeg$_i$ *bifalder at du*$_{not_i}$ *læser*
(I approve of you studying)

The subject of the subordinate clause must be different from that of the matrix verb.

Control for prepositional dependent clauses is also encoded:

Jeg tænker på at rejse til Nepal
(I am thinking of going to Nepal)

where the subject of the matrix verb "tænke" is the same as that of the prepositional dependent infinitive subclause "at rejse til Nepal" (subject control).

4.5 Syntactic Alternations

We have currently decided to link the following syntactic alternations:

- dative shift alternation
  Jeg giver hende en bog/Jeg giver en bog til hende
  (I give her a book/I give a book to her)
- alternation of constructs which do or do not contain expletives
  Der går en mand med en hvid hat ned ad gaden
  (A man with a white hat walks down the street)
  (lit. There walks a man down the street)

  Manden med den hvide hat går ned ad gaden
  (The man with the white hat is walking down the street)
- simple reciprocal alternation
  Jeg mødtes med Poul/Poul og jeg mødtes
  (I met with Poul/Poul and I met)
- alternation of causatives/unaccusatives
  Jeg lukker døren/Døren lukker
  (I close the door/The door closes)

4.6 Descriptions and Constructions

In the model, as it is at present, it is not possible to encode the preposition which introduces a position at the **SynU** level. Thus two similar syntactic frames, only differing wrt. the preposition introducing one position, must be assigned two different **Constructions** and thus **Descriptions**. It results in a large number of **Constructions/Descriptions** especially when encoding verbs. This may be acceptable for case marking prepositions. However, it is a problem when encoding weakly bound prepositions where a large number of prepositions can be subcategorized for by the same verb (but not always the same set of prepositions for all verbs).

Currently we have only encoded generic entries for weakly bound prepositional constructions and for movement verbs combining with different directional particles. We list, in the comment fields for each **SynU**, the prepositions/particles the current lexical entry admits. This must be changed time permitting. We have also chosen not to encode idiomatic expressions.

As an example a simplifyed **Description** for a divalent verb taking a subject and an oblique prepositional phrase (the preposition *i* (in) followed by an NP or by an infinitive with subject control) is given in Figure 1.

```
<DESC IDD="Dv2PPnic-i"
  NAME="divalent: NP, PP(NP/infinitive subj. control) prep=i"
  EXP="Han invilligede i (det kompromis/at tale med hende)
      (He accepted that compromise/speaking with her)"
  MU="indvilligede">
  <ST LBL=Clause>
     <POS FUNCTION=SUBJECT>
       <ST LBL=NP>
          <FT NAME=COREF VALUE="COI">
     </POS>
     <POS FUNCTION=OBLIQUE>
       <ST LBL=PP>
          <FT NAME=PREP VALUE="i"MU="i" SAT=Y>
             <POS FUNCTION=PREPDEPENDENT>
                <ST LBL=NP>
                   <FT NAME=COREF VALUE="COJ">
                   <FT NAME=CASE VALUE="UNMARKED">
                <ST LBL=Clause>
                   <FT NAME=INTROD VALUE="at"MU="at" SAT=Y>
                   <FT NAME=MOOD VALUE="INFINITIVE">
                   <FT NAME=SYNSUBCAT VALUE="SSINFINITIVE">
                   <FT NAME=COREF VALUE="COI">
                   <FT NAME=CONTROLT VALUE="SUBJECTCONTROL">
             </POS>
     </POS>
</DESC>
```

Figure 1: Description for a Divalent Verb

5 Problems

The difficulties which we have met encoding syntax for Danish verbs are of different kinds: general design problems, lexicographic problems, problems related to the adaptation of language specific phenomena to a model which did not account for them and problems related to the nature of the project.

The first group of problems include general design decisions such as whether or not to encode optionality of complements and generalizations by means of links between **Descriptions**. We chose the solutions which gave more information and which we found more useful to both NLP applications and human users, but these solutions are not necessarily the easiest for the lexicographer encoding the lexicon.

Lexicographic problems include defining "clear" criteria for deciding when a complement can be considered optional and for distinguishing complements, which must be encoded, from adjuncts which are not treated in the lexicon. Here we have followed a practical solution, being guided by the co-occurrences of a lexical entry in the corpus. If a complement is omitted in non-elliptical constructions it is a good candidate for being optional, and if the same adjunct/complement co-occurs frequently in the corpus with the actual lexical entry it will be coded as a complement.

A problem concerning the use of the PAROLE model has been that of currently "finding" solutions to phenomena specific to a language (or to a group of languages) that the model did not account for. One example is the encoding of expletive constructions such as

> *Der kommer en mand med hat*
> (A man with hat is coming)
> (lit. There comes a man with hat)
> and
> *Det luner med brænde i kaminen*[12]
> (The burning wood in the fireplace gives warmth)
> (lit. It gives warmth with the burning wood in the fireplace)

which are very common in Danish.

Another difficulty for Danish is to encode all the **SynUs** for the 20,000 **MuSs** required by the project, because Danish has a relatively "poor" basic vocabulary (e.g. in comparison with English). On the other hand the combinatorial behaviour of words enlarges the vocabulary and this is especially true for verbs combining with other word classes such as prepositions and particles. Thus many **SynUs** may be related to one **MuS**.

5.1 Partial Results and Concluding Remarks

For the time being we have encoded 3000 syntactic entries for approx. 2900 verbal lemmas (3766 **Descriptions** are linked to these entries), 455 **Description** types and 289 **Constructions**[13].

Encoding many **Descriptions/Constructions** to characterize similar syntactic frames is certainly time consuming, but once they are encoded into the lexicon it is quite easy to connect new morphological entries to them.

Although LE-PAROLE is a big project with many participants from different countries it has been enriching to discuss the linguistic problems together and trying to find common guidelines for resolving them within the short time we have.

The project's aim is to build lexical resources, thus LE-PAROLE is a practical project. But, although many choices we have made are of pragmatical nature they are, in most cases, supported by the material in the corpus.

The usability of lexica of the PAROLE type for NLP applications is self-evident. However, navigating in our still "uncompleted" lexicon we can already observe some phenomena which are quite interesting for a human user, such as common "meaning" similarities of **SynUs** sharing the same **Description** and regularities/irregularities of phenomena such as the use of expletive constructions.

Further work includes encoding the remaining entries/**Descriptions** etc., externally verifying the entries and the usability of the lexicon, and exploring the linguistic phenomena of related syntactical units in a more accurate way.

Obvious future projects are encoding of idiomatic expressions, adding semantic information to the existing syntactic entries, encoding domain-specific words, comparing the different PAROLE lexica, exploring the possibility of semiautomatically acquiring syntactic entries from the corpus.

[12] The prepositional phrase *med brænde i kaminen* is the deep subject of the clause.

[13] The figures regard only the verbs.

References

(Monachini 96) M. Monachini & N. Calzolari. Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages. EAGLES Document EAG-LSG/IR-T4.6/CSG-T3.2, EAGLES, May 1996.

(Consortium 93) GENELEX Consortium. *Report on Syntax Layer.* Technical Report, Genelex Consortium, 4.0, december 1993.

(Schøsler & Van Durme 96) L. Schøsler & K. Van Durme. The Odense Valency Dictionary: An Introduction. In *Odense Working Papers in Language and Communication*, volume No.13. Odense, 1996.

(Calzolari (ed.) 96) N. Calzolari (ed.). Parole Linguistic Resources: Technical Specifications Overview. WP4: Deliverable N.4, MLAP PAROLE, 1996.

(Jørgensen & Kirchmeier-Andersen (eds.) 91) L. Jørgensen & S. Kirchmeier-Andersen (eds.). Rule Coding Manual for Danish. Technical Report, V.1.1, Copenhagen, 1991.

(Sanfilippo *et al.* 96) A. Sanfilippo & et al. *Subcategorization Standards. Report of the EAGLES Lexicon/Syntax Group.* Technical Report, SHARP, Oxford Science Park, Oxford, UK, March 1996.

(Allan *et al.* 95) R. Allan & P. Holmes & T. Lundskjær-Nielsen. *Danish: A Comprehensive Grammar.* Routledge, 1995.

(Music & Navarretta 96) B. Music & C. Navarretta. Documentation of the Danish Lingware. Technical Report, LRE 61029, Deliverable E-D8-DK, Copenhagen, 1996.

(Sprognævn 86) Dansk Sprognævn. *Retskrivningsordbogen.* Gyldendal, 1986.

(Quirk *et al.* 85) R. Quirk & S. Greenbaum & G. Leech & J. Svartvik. *A Comprehensive Grammar of the English Language.* Longman, 1985.

(Calzolari & Zampolli 94) N. Calzolari & A. Zampolli. Standards to Make Natural Language Resources Shareable Resources. In *Proceedings of the International Workshop on Shareable Natural Language Resources*, Nara, 1994.