Propbank Annotation of Danish Noun Frames

Eckhard Bick
Institute of Language and Communication
University of Southern Denmark
eckhard.bick@mail.dk

Abstract

This paper presents a frame annotation scheme for Danish nouns, with VerbNet-derived frames and semantic roles covering both frame arguments and satellites. The scheme was implemented as a new module for a Danish frame tagger and applied to a 90,000-token Danish treebank with ongoing manual revision. In addition to explicit frames, Constraint Grammar rules are used to map free semantic roles on noun dependents without pre-defined frames, using general syntactic-semantic context clues. We discuss the annotation scheme and present a statistical breakdown and linguistic evaluation of the assigned noun frames and adnominal roles in the corpus.

1 Introduction

There is a long linguistic tradition of frame and role annotation for verbal predications, rooted in verb sense classifications on the one hand (e.g. Levin 1993), and the concept of semantic roles (also called thematic or case roles, Fillmore 1968) on the other. In a frame-based framework, verb categories and semantic roles are seen as interdependent, and predications are annotated for both, usually involving both valency-bound arguments and free (adverbial) satellites of a given verb. Two crucial resources in the area are FrameNet (Baker et al. 1998, Ruppenhofer et al. 2010) and PropBank (Palmer et al. 2005). The former is more lexicographical in its conception and focuses on a one-by-one exhaustive description of individual frames, the latter offers exhaustive proposition annotation of running corpus sentences, with an eye on applications such as Machine Learning (ML).

For Danish, both a FrameNet and a frame tagger (DanGram) have been published (Bick 2011), but unlike some work on larger languages, e.g. the German Salsa corpus (Rehbein et al. 2012), these Danish tools addressed only verbal frames, largely ignoring nominal predications. The work presented here strives to resolve this problem in a three-pronged fashion, with automatic corpus annotation based on (a) systematic derivation of noun frames from verb frames, (b) lexicographical treatment of argument-carrying nouns and (c) free role-mapping rules based on semantic noun classes and syntactic triggers.

2 De-verbal noun frame derivation

Rather than define separate frames for nouns, we think that frames have a sufficient level of abstraction to work across not only syntactic, but also morphological/POS variants. We therefore use the frame inventory of the Danish FrameNet, with around 500 categories[1], as is. For verbs themselves, morphological variation covers participles and gerunds, and allows verbs to fill adjectival or adverbial slots while still retaining their arguments, with parallel constructions in Danish and English, e.g. *the new book, published by Elsevier in 2012*, where the frames arguments are distributed across the head (”book” - TH/theme and the dependents of the participle: ”Elsevier” (AG/agent) and 2012 (LOC-

[1] cf. http://framenet.dk

TMP/temporal location). For noun slots, like English, Danish can use infinitive clauses (*To visit Paris without visiting the Louvre is a weird thing to do*), but inflectional nominalization of verbs (with -n: *råbe - råben* [shout]) is very rare: *Hans evige **råben** efter mere øl* (his constant shouting for more ale). However, Danish has two common and reasonably productive derivational morphemes, *-else* and *-(n)ing* that can be employed for nominalization. The verb's original arguments can optionally be retained and will appear in either genitive or post-nominal PP slots, with the former typically inheriting the subject role, and the latter inheriting object and adverbial roles:

1. Firmaets §AG overraskende **udfasning** /V:udfase/ af bonusordninger §PAT [the company's surprising curb on bonus schemes]

2. **fornyelse** /V:forny/ af offentlige bygninger [repair of public buildings]

3. **opsigelse** /V:opsige/ [cancellation]

As a first step to adapt the DanGram Frametagger[2] for noun frames, we therefore exploited derivational analysis to retrieve verbal frames for -else/-ing nouns where a corresponding verb form could be found by stripping the suffix. In order to increase lexical precision, we excluded nouns that had semantic-class tags[3] incompatible with actions, activities, events and processes, such as *følelse* [emotion, not "to feel"] or *forretning* [shop, not "to do business"]. Unlike English, Danish very productively uses morphological compounding (i.e. without space), and deverbal nouns can morphologically incorporate their arguments, both subjects (*kvindesvømning* [women swimming]), objects (*atomspaltning* [atom cleaving]) and adverbials (*dialysebehandling* [dialysis treatment]). For such compound nouns with a second part ending in *-else/ing*, we applied the act/event condition to both the second part and the noun as a whole (if tagged). In addition, compounds with semantic class differences between whole and second part were deemed unsafe.

In loan words, Danish also allows the Latin equivalent of its native -else/-ing derivation, where *-ere* verbs correspond to *-ion/-ation* nouns: *adoptere - adoption, approksimere - approksimation*. Not least in the scientific domain, these words constitute a sizeable section of the lexicon, and many *-ere* verbs have been moved into the common domain, taking *-else/-ing* suffixes and allowing productive prefixation, e.g. *afnazificere/-ing* ["denazify"], *detailregulere/-ing* ["regulate in detail"]. While adding the -ere/-ation derivation to the frametagger did improve coverage, there is a substantical risk of gaps due to irregular stemming, e.g. the phonetically motivated c/k shift in *kvalificere - kvalifikation* [qualify - qualification], making the method less automatic and more dependent on derivational lexicon entries.

3 Lexicon scheme for nominal frames

Apart from proofreading automatic derivational analysis and entering verb stems for irregular *-ation* nouns, we also introduced the option of entering complete nominal frames into the lexicon from scratch. This solution is obviously much more labour-intensive, but allows the treatment of argument-taking nouns without any de-verbal morphological clue.

Each noun frame entry (FN) lists first the corresponding verb frame and then a slash-separated list of possible semantic role arguments[4] (marked §) with their slot filler conditions (1-5). We distinguish between primary conditions and secondary, optional subconditions (present in 1-3). Primary conditions are placed before the role concerned, secondary condition after it. The former are syntactic slot conditions (left/genitive position, self and bound preposition lexeme), the latter are categorial

[2] https://visl.sdu.dk/visl/da/parsing/automatic/parse.php

[3] DanGram uses - and tags - a shallow ontology of around 200 so-called semantic prototypes, among them <act> [+CONTROL,+PERFECTIVE, <activity> [+CONTR,-PERF], <event> [-CONTR,+PERF] or <process> [-CONTR,-PERF]

[4] The Danish FrameNet foresees about 35 argument-capable roles and an additional 15 satellite roles

conditions concerning semantic class. Form conditions such as ’icl’ (non-finite clause) or ’fcl’ (finite clause) can be used both instead of a preposition condition or as a subcondition on the argument of the preposition (4).

1. hjælp - FN:**help**/til§BEN’all/til§FIN’act/fra§AG [help for/with]
2. krav - FN:**demand**/p§TH/om§ACT/til§REC’H/til§TP’all [demand for/to]
3. betaling - FN:**pay**/af§REC’H/af§CAU’act/for§CAU/til§REC/med§INS [payment to/for]
4. hensyn - FN:**adjust**/til§BEN [consideration]
5. ret - FN:**allow**/til§ASS/icl§ACT [a right to]

We created a new module for the DanGram Framtagger identifying nominal frames by trying to match conditions on argument slots and then assigning the corresponding semantic roles and tagging instantiated verb frames on the noun in question. If an NP has a human genitive dependent, agent role (§AG) will be used as a fall-back if no genitive condition with another role is found the lexical frame entries for the noun itself or, if relevant, its derivational verbal base. A special case of deverbal nouns are cases, where the noun denotes not the predicating core of the frame (as in -*else/ing*), but rather one of the arguments, usually the subject. Even without a realized predicator, such nouns still evoke their frame and will take genitive or preposition arguments representing other roles in the frame, as in the *hosting*-frame for *vært*, where the word itself is the agent (self§AG), while events and beneficiaries can be added as genitive [gen] or with the preposition for:

1. vært - FN:socializeO/self§AG/for,gen§BEN’H/for,gen§EV’occ [host for/to sb/an event]

4 Free role mapping

As for verbs, some PP dependents of nouns are not valency-bound by their head (arguments), but simple free satellites (adjuncts), with a low selection preference for a specific noun or frame. Thus, the majority of nouns can take a location complement, and most nouns with a deverbal component allow time complements. In these cases, a semantic role can be assigned to the adjunct complement, but no specific frame will be triggered by doing so, leaving the head noun untagged. Consider the following examples from our corpus, all of which contain an §EXT (extension) role complement mediated by the preposition *på*:

1. *nedskæringer på 750 millioner* [cuts amounting to 750 million] - frame: decrease
2. *håndteringsbeløb* på 50 kroner [a handling fee of 50 crowns] - compound, frame: cost
3. *fedtindhold* på 0,5% [a fat content of 0.5%]
4. *ikke så interessant efter 11 bind på 2 timer* [not so interesting after 11 volumes in 2 hours] - implied frame: read

These cases exhibit a cline from strongest-bound complement (1) to weakest-bound complement (4). Thus, in (1) the *decrease*-frame is recoverable from a verbal template *skære ned på*, which inherently implies a degree/extension of ”cutting down”. In (2), no verb is recoverable, but compound analysis reveals a second part noun *beløb* [amount], that has a frame listed in the lexicon, and a semantic class implying measurability. The second part of the compound (3), on the other hand (*indhold* [content]), evokes the *containing* frame, that only loosely implies degree, and only in connection with the first part (*fedt* [fat]). In (4), finally, the real predication (frame *read*) is elliptic, and only implied by a potential reading object (*bind* [volumes]). For (1) and (2), our corpus annotation task can rely on lexicon information, once the frame-carrying lexeme is identified. For (3) and (4), however, role mapping has to be performed without identifying a frame first. For this, we use Constraint Grammar mapping rules relying on the semantic class of the role carrier and its head preposition. Even where no semantic class is available, the degree/extension role can often be inferred

from modifiers (numbers) or even hinted at by inflexion (plural), as long as the trigger-preposition (*på*) is present and linked to a noun.

The rule below maps the EXT role (extension) on nouns of class <unit>[5], if they have the right preposition as parent (p) and a noun LINKed as grandparent, and if the immediate left context (-1) is either a numeral (NUM) or a fraction word (NUM-FRACT) preceded (-1) by a number or the article ”*en*”.

```
MAP (§EXT) TARGET N-UNIT (p (”på” PRP) LINK p N)
      ((-1 NUM) OR (-1 NUM-FRACT LINK -1 NUM OR (”en”))) ;
```

5 Results

Our noun frame scheme and annotator module were developed as part of a larger Propbank project for Danish, and applied to a 87,000-token treebank automatically pre-annotated with morphosyntactic tags, (ambiguous) semantic class potential for nouns and dependency links. The corpus is based on the larger, sentence-randomized Korpus2010 (Asmussen 2015) and covers a variety of both printed, electronically published and described sources, among them national newspapers (15%) and magazines (58%), blogs (8.5%), chat fora (2.5%), parliamentary speeches (10.5%) and various Internet sources (6%), such as a recipe website. While the Propbank is subject to revision at all levels of annotation, and will contain full mark-up for all verbal frames, we are here only concerned with noun frames and their distribution.

Our method tagged 9.6% (1342) of the about 15,000 nouns in the corpus as frame carriers, and identified 4477 ad-nominal roles. Of these, about 30% were linked to (and identified through) a noun frame, while the remaining 70% were assigned by free mapping rules. About half (2300) of the adnominal role carriers were themselves nouns, 26% were clauses (especially relative clauses), and 13% names. Predictably, we found a clear tendency for some roles to be frame-projected arguments (ACT, RES, CAU, TH, PAT) while others were mostly identified by free mapping rules (ATR, ID, LOC, ORI, EXT).

Tag	**role**	**% all**	**% frame arg**
ATR	attribute	27.11	4.9
LOC	location	14.2	7.2
TH	theme	9.0	67.9
TP	topic	7.0	38.8
ID	identity	6.4	6.3
PAT	patient	4.0	65.5
AG	agent	3.3	29.5
BEN	beneficiary	3.0	48.1
ORI	origin	2.9	23.1
FIN	purpose	2.7	63.4
HOL	whole	2.1	77.4
ACT	action	2.1	90.3
CAU	cause	2.0	69.7
EXT	extension	1.9	26.5
RES	result	1.5	80.9

Table 1: Semantic role distribution

[5] Note that the rule uses not the /unit/ tag itself, but a set, N-UNIT, which has been defined elsewhere in the grammar and allows the inclusion of other classes that unit itself, such as currency or containers, or even the addition of individual words, like the English-inspired fod [foot], which is not an ordinary Danish unit

The identified noun frames covered 741 different lexemes, amounting to a type/token ratio of about 1:2, indicating a higher lexeme spread than for verbs (type/token ratio of 1:8), probably due to the fact that both frequent and infrequent verbs are frame carriers, while many frequent nouns are not. All in all, 255 different frames were found, covering about half the frame inventory in the Danish FrameNet. Table 2 lists the most frequent noun frames, and for each of them, the three most frequent lexeme realisation. As can be seen, some frames are dominated by a single lexeme, while others have a more even lexeme spread. *be_part* and *run_obj* are examples of frames, where the carrier token often is itself a frame participant (i.e. deserving a role tag in its own frame), but most carrier words (and not only the 50% or so of verb-derived *-else/ing* words) function as predicators for their frame and can only carry role tags for a higher-level, containing frame.

Frame	**n**	**lexemes**
be_part	57	del 44, halvdel 5, led 2
investigate	40	undersøgelse 30, forskning 3, analyse 2
run_obj	39	formand 12, leder 5, forvaltning 4
future_having	35	mulighed 32, udbud 2, anvisning 1
decide	31	bestemmelse 12,regel 8,afgørelse 6
discuss	30	debat 9, samtale 4, forhandling 4
relate	29	forhold 9, spørgsmål 6, relation 5
cause	28	årsag 7, grund 6, konsekvens 5
adjust	26	regulering 9,omstilling 7,hensyn 5
explain	25	forklaring 8, eksempel 7, redegørelse 6
create	24	udvikling 9, udmøntning 7, fremstilling 3
allow	24	ret 7, adgang 5, godkendelse 3
tell	23	oplysning 11, historie 4, meddelelse 2
assess	22	vurdering 16, beregning 3
pay	20	råd 4, udgift 3, ressource 3
help	19	grundlag 9, støtte 3, hjælp 3

Table 2: Frame distribution

Since manual revision of the noun frames is work in progress, it is difficult to say how many frames we missed, but inspection indicates that 20% of the ad-nominal roles assigned by free mapping were in fact arguments rather than satellites,warranting a frame tag on their head noun. Given that certain roles are more likely to be arguments than others, these could be flagged for prioritized inspection, if linked to a frameless head. Another potential indicator for false-negatives are non-transparent nouns (i.e. excluding Danish equivalents of ”kind [of]”, ”lot [of]”, ”handful” etc.) that did not receive any role-tag. A low figure of 5% for this category indicates a good coverage for our method, not least because every second of such nouns is a simple genitive modifying a non-deverbal noun, with a very low chance of being a role carrier.

6 Conclusions and outlook

We have shown how a combined method of verbo-nominal derivation, lexical argument-slotfiller information and ontology-based role mapping rules can be used to extend annotation of a Danish Propbank from verbal to nominal frames with a reasonable coverage. Our automatic annotation allows lexeme- and category-based statistics amenable to linguistic information and conducive to an informed prioritization of future manual revision work. Thus, a first layer of noun frame annotation will be available in the upcoming 2017 release of the Danish Propbank.

References

Asmussen, Jørg. 2015. *Corpus Resources & Documentation*. Det Danske Sprog- og Litteraturselskab, http://korpus.dsl.dk

Bick, Eckhard. 2011. A FrameNet for Danish. In: *Proceedings of NODALIDA 2011*, May 11-13, Riga, Latvia. NEALT Proceedings Series, Vol. 11, pp. 34-41. Tartu: Tartu University Library.

Baker, Collin F. Baker; J. Fillmore; J. Charles; John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the COLING-ACL*. Montreal, Canada

Fillmore, Charles J. 1968. The case for case. In Bach & Harms (Ed.): *Universals in Linguistic Theory.* New York: Holt, Rinehart, and Winston. 1-88.

Palmer, Martha; Dan Gildea; Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics, 31:1.*, pp. 71-105, March, 2005.

Rehbein, Ines; Josef Ruppenhofer; Caroline Sporleder; Manfred Pinkal. 2012. Adding Nominal Spice to SALSA - Frame-Semantic Annotation of German Nouns and Verbs. *Proceedings of KONVENS 2012*, Vienna. pp. 89-97.

Ruppenhofer, Josef; Michael Ellsworth; Miriam R. L. Petruck; Christopher R. Johnson; Jan Scheffczyk. 2010. *FrameNet II: Extended Theory and Practice.* http://framenet.icsi.berkeley.edu/