From Treebank to Propbank:

A Semantic-Role and VerbNet Corpus for Danish

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

Abstract

This paper presents the first version of a Danish Propbank/VerbNet corpus, annotated at both the morphosyntactic, dependency and semantic levels. Both verbal and nominal predications were tagged with frames consisting of a VerbNet class and semantic role-labeled arguments and satellites. As a second semantic annotation layer, the corpus was tagged with both a noun ontology and NER classes. Drawing on mixed news, magazine, blog and forum data from DSL's Korpus2010, the 87,000 token corpus contains over 12,000 frames with 32,000 semantic role instances. We discuss both technical and linguistic aspects of the annotation process, evaluate coverage and provide a statistical break-down of frames and roles for both the corpus as a whole and across different text types.

1 Introduction

The syntactic potential and semantic structure of a language's lexicon can either be encoded explicitly in a dictionary or ontology, or implicitly through annotated data. Rule-based natural-language processing (NLP) will typically rely on the former, machine-learning (ML) systems on the latter. For the semantic annotation of predicate-argument structures, two well-known English ressources each addressing one of these two approaches are FrameNet (Baker et al. 1998, Johnson & Fillmore 2000, Ruppenhofer et al. 2010) and PropBank (Palmer et al. 2005), respectively. While FrameNet categorizes verb senses into frames with semantically restricted "slot-filler" arguments, PropBank departs from syntactically annotated corpus data to assign both roles and argument structure to each verb consecutively. The data-driven approach of PropBank promises better coverage and statistical balance[1], and therefore better automatic ML tagging, but its semantic role inventory and numbered arguments are highly predicate-dependent, and do not support semantic generalization and interpretation as well as FrameNet. A third approach, VerbNet (Kipper et al. 2006), opts for less granularity and a more limited set of roles and predicate classes. In recent years, corpora with such medium-granularity semantic-role annotation have been published for various languages, e.g. German (Mújdricza-Maydt et al. 2009) and Dutch (Monachesi et al. 2007).

For Danish, a VerbNet-based FrameNet (Bick 2011), with similar granularity (35 roles, 200 predicate classes subdivided into 500), achieved reasonable coverage in automatic annotation, but so far no manually validated corpus has been published. The SemDaX corpus (Pedersen et al. 2016) does provide human-validated semantic annotation of a Danish corpus, but only for word senses, and (with the exception of 20 highly ambiguous nouns) only for WordNet super-senses (Fellbaum 1998), not for semantic roles and predicate frames. In this paper, we present a corpus of similar size and composition, but with

[1] Random or running text samples not only guarantees a real-life statistical representation of lexical items, but also forces annotators to confront - and resolve - unforeseen constructions and contexts.

the full structural Verbnet frame annotation proposed by (Bick 2011), and augmented with corresponding frames for nominal predications. For verbs, this implies full sense disambiguation. For nouns, senses were added in the form of semantic-prototype tags, but only fully disambiguated in the case of named entities.

2 The corpus

For our corpus, we use the term *PropBank* in a generic sense, agreeing with Merlo & Van der Plas (2009) that VerbNet-style generalizations are important in the face of new semantic role instances, and that they complement the structural constraints that can be learned from a PropBank. Our data source is Korpus2010 (Asmussen 2015), a sentence-randomized corpus of Danish around the year 2010, compiled and distributed by the Danish Society of Language and Literature (Det Danske Sprog- og Litteraturselskab). The 44.1 million token corpus has a fairly broad coverage and includes both printed, electronically published and non-standard sources:

Danish PropBank section	**token percentage**
Magazines (text files)	44.99%
Magazines (scanned)	12.97%
National newspapers	14.58%
Parliamentary speeches	10.51%
Chat fora, young adults	2.48%
Web sources (various)	5.9%
blogs	8.51%

Table 1: Corpus composition

Our subcorpus of ca. 87,000 tokens (4,924 sentences/utterances) was built using random id-based sentence extraction. Compared to Korpus2010 as a whole, the subcorpus has a higher percentage of blog and chat data, less news and more magazines. A few excerpts were discarded, mostly because the original automatic sentence separation was erroneously triggered by abbreviation dots resulting in incomplete fragments.

3 Annotation levels

While our research focus is on the semantic annotation of verb-argument structures, and - widening this scope - the semantic annotation of predications and roles in general, this higher-level annotation is built upon a skeleton of syntactic tags and dependency links, and the corpus can therefore also be used as an "ordinary" treebank. As such, it complements the bigger, but older Danish *Arboretum* treebank, that contains data from Korpus90 and Korpus2000[2]. Both resources are available in the ELRA catalogue (http://catalog.elra.info/).

3.1 Tokenization and morpho-syntax

We use functional units as tokens, rather than strictly space-separated strings, in order to facilitate assignment of higher-level syntactic and semantic tags. Thus, complex prepositions and conjunctions *(i=stedet=for [instead of], på=trods=af [despite], i=og=med [though], for=så=vidt [in so far as])* are used for syntactic perspicuity, and named entities are fused for semantic reasons. Non-adjacent parts of lexemes (verb particles) are marked, but only linked at the syntactic level.

Morphosyntactic annotation adopts an analytical scheme, with separate tags for POS, syntactic function and each morphological feature, rather than complex all-in-one tags. Due to the underlying automatic pre-annotation, native annotation uses Constraint Grammar (CG) abbreviations, but the corpus is available in a variety of treebank output formats, such as the cross-language analytical VISL standard (visl.sdu.dk), MALT xml, TIGER xml and Universal Dependencies in CoNLL format.

3.2 Valency and dependency relations

The corpus strives to make a connection between a verb's valency potential, dependency relations and semantic arguments. Thus, the latter can be viewed as fillers for valency slots projected by dependency links. We therefore mark both the instantiated valency (e.g. <v:vt> for monotransitive), syntactic dependency links and (separate) semantic argument links. This way, we support a triple representation of tree structure:

(a) a shallow, verb-marked lexical valency tag, chosen according to which arguments are actually present in a given sentence

(b) a traditional *syntactic* dependency tree, with prepositions and auxiliaries as heads of PP's and verb chains, respectively, and conjuncts chained onto each other

(c) a semantic (tectogrammatical) tree with only content words as nodes (i.e. main verbs and

[2]Like *Arboretum*, the new corpus will be distributed through the ELRA catalogue of Language Resources.

PP-nominals rather than auxiliaries and prepositions), and with conjuncts linked to the same head in parallel

This multi-layered approach is similar to the semantic role linking scheme used in the Prague Dependency Treebank (Böhmová et al. 2003), and different from the Universal Dependencies scheme (McDonald et al. 2003), which opts for a 1-layer approach by *replacing* syntactic links with semantic ones, while still maintaining "surface-near" non-semantic labels.

3.3 Semantic prototype annotation

Our annotation scheme maps a semantic ontology onto nouns, with around 200 so-called semantic prototype categories[3], organized in a shallow hierarchy. Thus, major categories like <H> (human), <V> (vehicle), <tool> etc. are further subdivided, e.g. <Hprof> (profession), <Vair> (planes), <tool-cut> etc. During treebank generation, these tags are used both for contextual disambiguation and as slot fillers for the identification of verb frames. Even in the face of polysemy, these tags are usually sufficient to pinpoint a given verb sense and frame, because the choice is further constrained by the syntactic function of a verb's arguments, as well as POS and morphological form.

Once frames and roles are established, these can be in turn used, to automatically discard non-appropriate noun senses, ideally leaving only one or at least non-conflicting sense tags[4]. In the current version of the Propbank, manual validation and disambiguation of sense tags has not yet been concluded. Once finished, another task will be to assign, where available, DanNet senses (Pedersen et al. 2008) in a semiautomatic way by mapping one ontology onto another.

3.4 NER annotation

Named entity recognition and classification (NER) of proper nouns and numerical expressions is needed to supplement semantic noun classification, and important for verb frame identification. In principle, the same ontology could be used, but the underlying parser already implements a separate scheme with around 20 NER categories, which can be seen as an extension of - and in some cases synonyms of - the semantic noun tags. Following the MUC conference standard, there are six main categories: person <hum>, organization <org>, place <top>, event <occ>, work of art <tit> and brand <brand>. Because some names have a cross-category potential, <civ> (civitas) was added for places that can act (e.g. build or go to war), <inst> for site-bound organizations or activities and <media> for names that can function as both titles and organizations (e.g. newspapers and certain websites). In these cases, the co-tagged semantic role label will functionally complete the NER categorization of a given name, for instance §AG (agent) vs. §LOC (location) for towns or countries. Tokenization is an important issue in NER, because many names (almost half in our corpus) are multi-word units (MWU) and need to be recognized before they can be classified. To do so, the input parser relies on both pattern matching/reprocessing, a gazetteer lexicon and contextual rules applied after the POS-tagging stage. In the published corpus, both NER tokenization and classification was manually revised. 3.6% of all non-punctuation words in the corpus are names, with a MWU proportion of 42%, and an average MWU length of 2.4 parts.

3.5 Syntactic function and dependency

Dependency links are the necessary backbone of a predicate-argument frame, and syntactic function tags (subject, different object types, subject and object complements, valency-bound adverbials etc.) are useful as argument slot-filler conditions in the automatic assignment of frames. Annotation errors at the syntactic level will therefore often lead to frame and verb sense-errors. Because of this interdependency, inspection/revision of either annotation level helps identifying errors at the other one, too, effectively creating a traditional treebank and a propbank at the same time.

Structurally, however, syntactic trees and Propbank trees are not identical, because the latter propagate ordinary dependency links to meaning-carrying words. Thus, each argument in our corpus carries at least two head-id links, one for the immediate syntactic head (e.g. preposition, first conjunct, auxiliary), and one for the semantic relation (to a verbal or nominal predicator). Furthermore, while traditional dependency links only allow one head, semantic

[3] The category set adopted here is described at: http://visl.sdu.dk/semantic_prototypes_overview.pdf

[4] A non-conflict is a combination of a subset tag with its superset tag, or underspecified <sem-r> (readable) and <sem-l> (listenable) for "værk" (work of art).

relations may ask for multiple heads due to "transparent" arguments (e.g. relative pronouns) , unexpressed arguments (subjects of infinitive verbs) or coordination ellipsis. Thus, in Fig. 1, *Majbrit*, dependency-wise the subject of *sige* ("speak"), is not only role-linked to the latter as §SP (speaker), but also - as §EXP (experiencer) - to the predicate in a depending relative clause, *kan lide* ("likes"), and finally - as §AG (agent) - to an infinitive clause, *bo* ... ("live ..."), three levels further down in the dependency tree. In this case, ordinary treebank dependencies suffer from an argument overlap, impeding tasks like information extraction. Propbank annotation, on the other hand, clearly marks three different predications:

Majbrit §SP	*siger* (says) <fn:say>	... §SOA / §MES
Majbrit §EXP	*kan lide* (likes) <fn:like>	*at bo* ... (to live) §TH
Majbrit §AG	*at bo* (to live) <fn:lodge>	*sammen med fire andre* (with four others) §AG-COM

Table 2: Overlapping propositions

Word	*Lex*	*Secondary, **Frame***	*POS, morphology*	*Synt. funct.*	*Sem. role*	*Dep.*
...						
siger	sige	<v:vq><**fn:say**>	V PR AKT	@FS-STA		#14->0
Majbrit	Majbrit	<fem>	PROP NOM	@<SUBJ	**§SP:14** **§EXP:20** **§AG:24**	#15-14
,	,		PU	@PU		#16->0
som	som	<clb> <rel> <H>	INDP nG nN	@SUBJ>		#17->19
godt	godt		ADV	@ADVL>		#18->20
kan	kunne	<auxmod> <aux>	V PR AKT	@FS-N<		#19->15
lide	lide	<v:vt><**fn:like**>	V INF AKT	@ICL-AUX<	**§ATR:15**	#20->19
hyggen	hygge		N UTR S IDF NOM	@<ACC	**§TH:20**	#21->20
ved	ved		PRP	@N<		#22->21
at	at		INFM	@INFM		#23->24
bo	bo	<v:vp><**fn:lodge**>	V INF AKT	@ICL-P<	**§CIRC:21**	#24->22
sammen =med	sammen =med		PRP	@<PIV		#25->24
fire	fire	<card> <value>	NUM P	@>N		#26->27
andre	anden	<diff>	DET nG P NOM	@P<	**§AG-COM**	#27->25

Fig.1: Multiple semantic heads
*(... **says** Majbrit, who **likes** the coziness in **living** together with four others)*

3.6 Verb frame annotation

Our annotation of proposition-argument structures is based on the category set of the Danish FrameNet (Bick 2011), which uses ca. 500 classes based on the original VerbNet senses, albeit with a modified naming system[5] and additional subclassification. Thus, though syntactic alternations such as diathesis or word order are not considered frame-distinctors, the Danish FrameNet differs from both WordNet and VerbNet by introducing polarity antonyms like *increase - decrease*, *like - dislike*, and a self/other distinction (*move_self*, *move_other)*. The scheme also avoids large underspecified classes, subdividing e.g. *change_of_state* into new classes like *heat - cool*, *activate - deactivate* and *open - close*.

A first-pass frame annotation was performed by running a frame-mapper program exploiting existing morphosyntactic and semantic class tags as well as argument-verb dependencies assigned by the DanGram parser. Action and event nouns with argument slots and a de-verbal morphology (in particular Danish *-else/-ing* verbs) were annotated with the corresponding verb frames. All verbs and deverbal nouns were then manually inspected together with their arguments, which led to corrections in 15-20% of all frames. About a quarter of these were due to syntactic annotation errors or, sometimes, faulty POS-tagging[6]. Given the high lexeme coverage of the Danish FrameNet, very few verbs were left completely frameless, so most of the errors were mistaggings due to frame patterns not foreseen in the Danish FrameNet lexicon, often involving phrasal constructions with incorporated adverbs

[5] We wanted the class names to on the one hand be real verbs, on the other to reflect hypernym meanings wherever possible. Therefore, we avoided both example-based names (common in VerbNet) and - mostly - abstrac concept names (common in FrameNet) that are not verbs themselves.

[6] The parser's error rate for POS is about 1%, and 5% for syntactic function.

and prepositions (e.g. *slå ned på* [hit down at] - "stop"-frame), or idiomatic expressions with non-literal nominal arguments (e.g. *bære frugt* [carry fruit] - "succeed"-frame). In these cases, the frame tagger defaults to the first frame listed for a given valency, or to a basic transitive or intransitive frame, if there is no valency match in the lexicon either.

In order to speed up manual revision work, missing frames were added to the FrameNet lexicon, and missing valencies to the parser lexicon, and automatic annotation was then repeated for the remaining, not-yet-revised part of the Propbank in steps of about 20%.

3.7 Semantic role annotation

Our Propbank assigns semantic roles to both predicate arguments and free (adverbial) satellites not valency-bound by their head. The former are automatically mapped onto syntactic predicate-argument "skeletons", together with the chosen verb sense, once a given frame is chosen. For a correct syntactic tree, errors in such roles will always manifest as frame/sense errors, too. Satellite roles, on the other hand, depend less on the verb, and have to be tagged from local clues alone, e.g. the preposition and semantic noun class of an adverbial PP. The annotation scheme distinguishes between 38 argument-capable semantic roles and an additional 14 roles that can only occur as satellites.

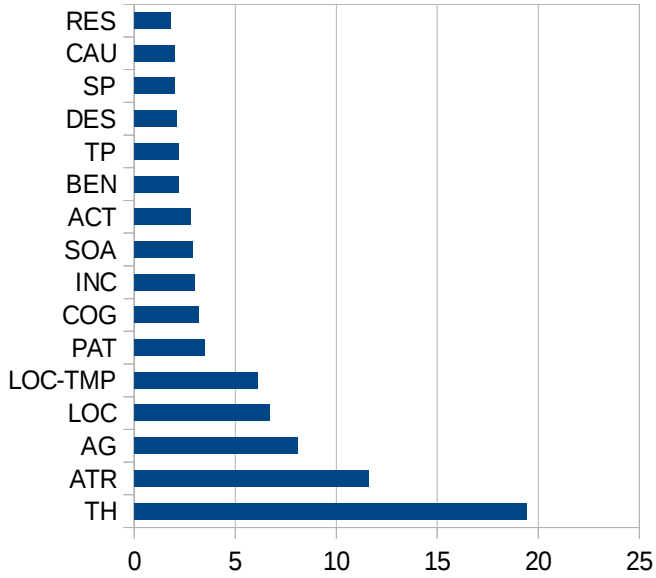

Fig. 2: Semantic-role token percentages

As can be seen from Fig. 2, the 5 main roles account for over 60% of all cases. Compared to the unrevised frame annotation of newspaper data reported in (Bick 2011), place and time locations (§LOC, §LOC-TMP) figure more prominently and there are more incorporates (§INC). The most likely explanation for this is not so much the difference in source genres, but rather the more complete coverage of satellite (free) adverbials and the exhaustive treatment of all verb particles and incorporated nouns in the corpus.

4 Annotation Procedure

The current annotation is a 1-annotator linguistic revision of an automatic annotation, with parallel improvements in the underlying DanGram parser[7] and Danish FrameNet lexicon[8] followed by intermittent re-annotation of not-yet-revised portions, in 20%-steps. The lack of multi-annotator cross-checks, while not standard procedure, has the advantage of reduced cost and more data per time unit. As a side effect, there is a certain consistency advantage compared to at least an *incomplete* multi-annotator setup where not *all* annotators revised *all* data, or where annotators could not agree.

The revision was performed twice - first with a focus on main verbs and valency-bound arguments, then with a focus on non-verbal predications and satellite roles. The first pass also resolved V/N part-of-speech errors, and consequently major tree structure errors, together with argument function errors. Unlike arguments and verbs, which warrant 100% semantic tagging, there is less linguistic consensus as to which tokens should be marked semantically, with satellite roles. Apart from lexically pre-marked material (in particular space, direction and time adverbs), nouns and names are the most likely semantic role carriers. For the latter, a complete, separate inspection pass was carried out, for the former, a mini Constraint Grammar was run on the already-annotated corpus to mark missing roles. The simplified marker rule below looks for nouns in the nominative without a §-marked role. Excluded are top-level nodes (ADVL [unattached adverbial] and NPHR [function-less NP]), plus transparent nouns (*slags/art* [kind of], quantifiers etc.). In order to ensure that only the main noun in an NP is addressed (e.g. *amerikaneren professor Pentland*), there are NEGATE checks for

[7] visl.sdu.dk/visl/da/parsing/automatic/parse.php

[8] http://framenet.dk

immediate parent or child nodes without interfering frame carriers (such as verbs).

```
MAP (§NOROLE) TARGET N + NOM - (/§.*/r)
  (NEGATE *p (/§.*/r) - (<fn:.*>r)
        BARRIER (<fn:.*>r))
  (NEGATE *c (/§.*/r @N<)
        BARRIER (<fn:.*>r))
  (NOT 0 @ADVL/NPHR OR <transparent>) ;
```

5 Evaluation and statistics

We evaluate our propbank statistically, in order to assess corpus parameters such as lexical spread, representativeness, frame and role type frequencies. In addition, the relative distribution of these semantic categories across text types, as well as their interdependence with other, lower-level linguistic categories is of interest, given that this is the first time a comprehensively annotated and revised Danish corpus is available for this level of annotation.

At the time of writing, the corpus contained 10,708 instances of main verbs, covering 1275 different lexemes, 100% of which were annotated with frames. By comparison, only 9.6% of the ca. 15,000 nouns were frame carriers, albeit with a much higher type/token ration (741 lexemes/1722 tokens) than for verbs. Frames for other word classes were only assigned to about 190 adjectives (65 lexeme types), a few direction adverbs and a single determiner:

- bange for [afraid of] + §CAU
- følsom over for [sensitive to] + §STI
- syd for [south of] + §LOC
- anden end [other than] + §COMP

In addition, 81 attributively used, prenominal participles (52 types), received "naked" verb frame senses, without arguments[9], mostly corresponding to the default sense of the underlying verb.

The corpus contains examples of 454 different frames, covering 91.9% of all frame types in the Danish Framenet, and 598 (or 44.7%) of the possible frame type combinations (e.g. "udgive sig for" - <fn: imitate && role_as>). Since frames are used to disambiguate the valency potential of a given verb and to define its senses, it is also possible to quantify verb polysemy in the corpus. All in all, we found different 2153 verb senses[10], amounting to an average of 1.69 senses per verb lexeme, albeit with huge differences between lemmas (table 3).

verb	**sen-ses**	**3 most frequent frame senses (number of instances)**
gå	54	run 30, leave 14, reach&&participate 13
være	47	be_copula 1522, exist 260, be_place 217
komme	41	reach 69, appear 24, occur 17
tage	36	take 38, do 19, run 10
holde	27	run_obj 8, persist 8, defend_cog, endure, hold, keep, like, sustain 6
have	26	have 346, own 18, cope 13
stå	21	be_place 27, be_attribute 11, spatial_conf 8
sætte	21	put 12, start 6, decrease, change_body_position 4
lægge	20	put_spatial 14, suggest 5, create_semantic 3
slå	18	beat 6, confirm, deactivate, integrate, succeed 2
se	18	see 85, notice 22, be_attribute 16
gøre	18	do 78, turn_into 43, cause 11
få	18	get 165, cause 37, obtain 12

Table 3: Most polysemous verbs

In almost all cases, sense differences come with differences in argument structure or phrasal particles etc., but the inverse is not true - there may well be more than one syntactic realization of a given verb sense. Thus, there are 24.6% more valency-sense combinations for the verbs

[9] This is a gray zone - a number of Danish deverbal adjectives could arguably also be read as -ende/-et participles [-ing/ed], but for now we simply followed the choices made in the parser lexicon, assigning frames to attributive participles only where they were productively derived from verbs. Another decision was not to tag the heads of such attributive participles with argument roles referring back to their own modifiers (e.g. *voksende* [growing] + §PAT. Postnominal participles, on the other hand, are all argument/satellite-carriers in Danish, and hence assigned roles.

[10] By comparison, the Danish FrameNet contains 11174 verb senses for 7033 lexemes, i.e. 1.59 senses per verb. Hence, our corpus data is slightly more ambiguous, even on a type basis, than the lexicon, probably because the corpus only covers the 18.1% most frequent verbs, and 19.7% of all senses in the FrameNet lexicon, lacking many rare, but unambiguous verbs.

in the corpus than just verb senses[11]. Interestingly, senses have a much more even frequency distribution for some verbs (e.g. "holde" [hold] and "slå" [hit]) than for others ("være" [be], "have" [have]).

Semantic role statistics are complicated due to the fact that one token may participate in several different frames across the sentence, and therefore carry multiple role tags. All in all, there were 20,437 semantic role tags for verb arguments, and 5252 role tags for verb satellites, corresponding to a 79.6% / 20.4% distribution. Arguments were more likely to share a token than satellites (with an average of 1.1 roles per token for the former, and 1.026 for the latter). For the rarer non-verbal frame-heads (mostly nouns), the argument-satellite balance was almost the opposite (29.7% / 70.3%), with 1303 argument roles and 3077 satellite roles, and a few multi-tag tokens (1.007 tag/token for arguments and 1.008 for satellites).

For classifying semantic roles as arguments or satellites, and to mark them for verbal or non-verbal head type, we used the same method described in ch. 4 for consistency checking, namely CG mark-up rules, exploiting syntactic function tags and frame relation links as context.

6 Linguistic text profiling

We also examined the distribution of both frames and semantic roles across different text types, hoping to identify text type- (or even genre-) specific traits in a semantically generalized fashion, different from - and arguably more linguistic and generalized than - standard techniques such as bag of words.

Role rank	News	Maga-zines	Blog	Forum	Parlia-ment	Reci-pes
	4909	14898	1026	740	1412	317
1	**ORI**	**MES**	**DES**	**STI**	**FIN**	**CIRC**
2	**BEN**	**CAU**	**LOC-TMP**	**COG**	**TH-NIL**	**PAT**
3	**SP**	TH-NIL	COG	**EXP**	**ACT**	**EXT-TMP**
4	**ID**	**TP**	**INC**	**ATR**	**RES**	BEN
5	**EV**	ID	AG	TH	TP	DES
6	**COMP**	**SOA**	EXP	**COMP**	CAU	**LOC**
7	EXP	LOC	ATR	ID	SOA	**TH**
8	AG	RES		**REC**	ORI	ATR
9	SOA	SP			INC	ACT
10	LOC	REC			BEN	

Table 4: Relative role ranks across text types

The semantic role ranks in table 4 are computed by normalizing in-text relative frequencies according to all-corpus relative frequencies. Only roles more frequent than the corpus average are listed, and each role is bold-faced where it ranks highest in the table overall. By using normalized frequencies rather than absolute ranking, text differences are emphasized, and patterns become more salient.

Thus, the reporting style of news text rhymes with top ranks for §SP (speaker), §ID (typical of explaining name appositions) and §EV (events). The high rank for §ORI (origin) is symptomatic of quote sourcing ("according to .." etc.). Furthermore, the news texts evaluate facts by comparing them (§COMP) and by discussing who was affected, profited or suffered (§BEN, benefactive). Magazines, though a similar text type in other linguistic aspects, address more specific audiences and topics (§TP), and are - by comparison - more interested in bringing a message across (§MES) and making claims (§SOA, state-of-affairs).

Blogs and discussion fora are the most personal text types in our corpus, and are characterized by opinions and cognitiveness (§COG), relaying experiences (§EXP, §STI) and describing or judging things (§ATR, attribute). In addition, blogs, often written as a personal timeline or travel report, rank high for time markers (§LOC-TMP) and destinations (§DES). Interestingly, blog writers have high scores for non-literal language, with a lot of verb incorporations (§INC). While in theory also 1-person text types, parliamentary speeches are very different from blogs and fora, and more argumentative than the rest of the corpus, scoring high on intention/planning (§FIN), results (§RES) and discussed actions (§ACT). Also, these speeches rank higher than even news texts for impersonal constructions linked to formal subjects (§TH-NIL, "*det er X der*", *"der + s-passive"*).

Finally, a small but "spicy" section of the corpus is dedicated to recipes, which are known to stand

[11] This count includes difference between transitive and ditransitive use and between NP and clausal objects, but not the difference between active and passive

out even in morphological ways (imperatives, uninflected nouns, unit numbers). In terms of semantic roles, recipe sentences are all about *changing* (e.g. frying) things (food, §PAT patient) in certain *circumstances* (§CIRC) for a certain *amount of time* (§EXT-TMP).

For frames, a text-wise break-down is informative, too, and provides a useful means of abstraction compared to simple lemma frequencies. Thus, frames help lump together both morphological and POS variation (*unhire* = fyring, massefyring, fyre) and lexical variation (*steal* = snuppe, nappe, stjæle). For the weighted ordering in table 5, relative in-text frequencies were used, with a weighting exponent of 1.5 for the numerator. Compound frames were split into atomic frames.

Text type	Weighted top-ranking frames (absolute numbers in parentheses)
News	be_copula (221); **unhire** (7); behave (8); **dispute** (5); **trade** (5); have (54); reach (30); succeed (13); **run_obj** (19); **tell** (22); lower (4)
Maga-zines	be_copula (712); **say** (137); have (163); **exist** (130); **assume** (69); **become** (97); affect (43); **get** (88); cause (68); relate (56); be_part (41)
Blog	**long** (5); serve_to (2); steal (3); be_copula (138); belong_to (3); **send** (10); **know** (25)
Forum	**like** (7); appear (9); be_copula (62); **hear** (3); add (3); **know** (10); **inquire** (3)
Parlia-ment	**compensate** (7), **exist** (61); **ensure** (17); **exempt** (2); adjust (16); **improve** (15); **unestablish** (4); be_copula (160); agree (15); suggest (16); exaggerate (3)
Recipes	**prepare_food** (23); supply (10); **combine** (7); **cover_ize** (4); add (5); pour (2); **put_spatial** (4); put_deposit (3)

Table 5: Frame ranking across text types

Frame analysis more or less supports the picture suggested by semantic role distribution, but is somewhat more concrete, and provides more insight into topics. Thus, news text is about firing people (unhire), disputes, trade and how to run things (run_obj). Parliamentary debates are about reform (ensure, improve, unestablish) and discussion (agree, suggest). The high rank for the *exist*-frame is a form-trait, and due to impersonal constructions *(der er)*. People in blogs and fora are a bit more emotional (*long*, *like*), and information is essential (*know*, hear, *inquire*). The most concrete text type is recipes (*prepare_food*), where frames are about physically manipulating things (*combine*, *add*, *put*, *pour*, *cover_ize*).

7 Conclusions and outlook

We have presented a first proposition bank for Danish, with extensive annotation of both argument and satellite roles, for both verbal and nominal VerbNet frames. Offering both syntactic and semantic tree structures, and three levels of node annotation (syntactic function, semantic ontology and semantic role), the corpus aims to serve multiple ML and linguistic purposes. By way of example we have discussed frame- and role-based text profiling.

In terms of additional annotation, a useful next step would be to improve the semantic annotation of pronouns by adding anaphorical relations. The current, sentence-randomized corpus, however, will allow this only for in-sentence relations. The same is true for another type of relational annotation, discourse analysis, and a future version of the corpus should therefore include a running text section from a source, where this is not a copyright problem.

Also, using randomized sentences from a multi-source corpus, while providing a good statistical sample of a language, is not the best way to beat Zipf's law. Therefore, in order to extend per-type coverage for verb senses in the Danish FrameNet, future work should include a second propbank section, where sentences are extracted from an automatically pre-tagged Korpus2010 not randomly, but based on which verb senses they contain.

References

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

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

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.

Böhmová, Alena ; Jan Hajič; Eva Hajji; Barbora Hladká. 2003. The Prague Dependency Treebank: A Three-Level Annotation Scenario. In: Anne Abeillé (ed.): Text, Speech and Language Technology Series. Vol. 20. pp 103-127. Springer

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In Proceedings of ACL 2013

Fellbaum, Christiane (ed.). 1998. WordNet: An Electronic Lexical Database. Language, Speech and Communications. MIT Press: Cambridge, Massachusetts.

Johnson, Christopher R. & Charles J. Fillmore. 2000. The FrameNet tagset for frame-semantic and syntactic coding of predicate-argument structure. In: Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL 2000), April 29-May 4, 2000, Seattle WA, pp. 56-62.

Kipper, Karin & Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extensive Classifications of English verbs. *Proceedings of the 12th EURALEX International Congress.* Turin, Italy. September, 2006.

Merlo, P. & Van Der Plas, L. (2009). Abstraction and generalisation in semantic role labels: Propbank, Verbnet or both? In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1, pp 288–296. ACL

Monachesi, P.; G. Stevens; J. Trapman. 2007. Adding semantic role annotation to a corpus of written Dutch. In: Proceedings of the Linguistic Annotation Workshop. pp 77–84. ACL

Mújdricza-Maydt; Éva & Silvana Hartmann; Iryna Gurevych; Anette Frank. 2016. Combining Semantic Annotation of Word Sense & Semantic Roles: A Novel Annotation Scheme for VerbNet Roles on German Language Data. In: Calzolari et al. (eds). Proceedings of LREC 2016.

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.

Pedersen, B.S.; S. Nimb; L. Trap-Jensen. 2008. DanNet: udvikling og anvendelse af det danske wordnet. In: Nordiske Studier i leksikografi Vol. 9, Skrifter published by Nordisk Forening for Leksikografi, pp. 353-370

Pedersen, Bolette Sandford; Braasch, Anna; Johannsen, Anders Trærup; Martinez Alonso, Hector; Nimb, Sanni; Olsen, Sussi; Søgaard, Anders; Sørensen, Nicolai. 2016. The SemDaX Corpus - sense annotations with scalable sense inventories. In: Proceedings of the 10th LREC (Slovenia, 2016).

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/ index.php?option=com_wrapper&Itemid=126