A Semantic Ontology of Danish Adjectives

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

Abstract

This paper presents a semantic annotation scheme for Danish adjectives, focusing both on prototypical semantic content and semantic collocational restrictions on an adjective's head noun. The core type set comprises about 110 categories ordered in a shallow hierarchy with 14 primary and 25 secondary umbrella categories. In addition, domain information and binary sentiment tags are provided, as well as VerbNet-derived frames and semantic roles for those adjectives governing arguments. The scheme has been almost fully implemented on the lexicon of the Danish VISL parser, DanGram, containing 14,000 adjectives. We discuss the annotation scheme and its applicational perspectives, and present a statistical breakdown and coverage evaluation for three Danish reference corpora.

1 Introduction

This paper describes a multi-dimensional semantic classification system for Danish adjectives. The system has been implemented for a fairly unabridged computational lexicon, with 14,000 adjectival lemmas, and is intended for use with Danish NLP tools in general, and machine translation and semantic correctness grading of generated Danish sentences in particular.

Lexical resources about the semantics of adjectives are much harder to come by than corresponding dictionaries for nouns and verbs, not least in the context of less-resourced languages like Danish. Nouns allow the construction of ontologies based on hyponym-hyperonym relations (e.g. Princeton WordNet, Fellbaum 1998 for English, and DanNet, Pedersen et al. 2009, for Danish), and verbs can be classified using argument relations and restrictions (e.g. FrameNet, Baker et al. 1998 and Ruppenhofer 2010, for English). However, both methods are less ideal for adjectives - only a small set of adjectives takes arguments, hyponym-hyperonym relations are problematic, and traditional WordNet synonym clusters and antonym relations do not constitute a true classification system. One way out is using noun classification as a proxy and linking adjectives to nouns or verbs:

(a) property nouns denoting the property that the adjective describes, e.g. linking "hot", "tepid", "cool", "cold", "ice-cold" etc. to the noun "temperature", a method that works well for antonomy and scale adjectives. Thus, EuroWordNet (Vossen 1998) uses a "near synonymy" relation across word classes, e.g. "obese/obesity", "infamous/infamy".

(b) derivational base: A large percentage of adjectives are morphologically derived from nouns or verbs using suffixes, e.g. "*V-lig"* (V-able), *"N-lig"* (N-like), *"N-fuld"* (being full of s.th.), *"N-løs"* (not having s.th.). In addition, many Danish adjectives are morphologically past or present participles and can thus be linked to a verbal base (*"V-et"* - V-ed, *"V-ende"* - V-ing).

(c) nominal heads: Adjectves can be classified according to their prototypical head houn, using categories like "animal adjective" ("*tame"* - tame, "*vild"* - wild, *"glubsk"* - voracious) or "food adjectives" (*"bagt"* - baked, *"fersk"* - fresh, *"lækker"* - tasty).

However, (c) lumps semantically very different adjectives together (e.g. states, quality, source, purpos etc. for the food category), and neither (a) nor (b) is, on its own, applicable to the entire adjective lexicon, and morphological/derivational links, in particular, are slippery ground, as meaning can change over time, and become less transparent. Thus, "*huslig"* ("housely") does not mean "house-like" (the literal meaning), but rather "house-related" (tasks) or a human psychological trait of "housewife-ness". Also, sometimes the adjective is primary in a derivation relation, as in *"tapper" > "tapper-hed"* (brave > braven-ess), risking a sparceness of information, if the corresponding noun is simply classified as "property" exactly because its core is really adjectival.

GermaNet (Hamp & Feldweg, 1997) addresses the problems with (a) and (b) by establishing a separate semantic class hierarchy[1] for adjectives, with 16 classes at level 1 and 78 classes at level 2, with relations like *"green" > colour > perception* or *"short" > dimension > spatial.* Transparent denominal and deverbal derivations is classified as "pertainyms". For Danish, Nimb & Pedersen (2012) suggest the use of thesaurus data to build a type (c) classification by harvesting "property_of" relations between adjectives and typical collocate classes (e.g. person, thing, feeling, food). However, the authors mention the need for validation, and the current public version of DanNet[2] does not contain a "property_of" feature.

2 Existing resources

DanNet (and its dictionary precursor STO[3]) is one of two large sets of lexical resources used in Danish language technology. However, it only contains about 3,000 adjectives, with a flat 12-category ontology, and while there is information about hyperonym relations to either other adjectives or nouns, 23% are linked directly to the top node "property" or "property:physical" without any real classificational information. The other resource is the lexicon of the Danish VISL parser, DanGram (Bick 2001), containing 103,000 non-name lemma entries, of which about 14,000 are adjectives. The lexicon specifies syntactic word-order information for 11,400 of these, comprising obligatory predicative or attributive use, and so-called "modificational zones" (ordering in case of multiple prenominal adjectives).

<pred> predicative use only: *alene (alone), beliggende (situated), slut (finished)*

<att> attributive use only: *al (all), aldersmæssig (age-related), aldrende (becoming older)*

<mod1> (specificational): *bestemte* alvorlige organiske sygdomme (*certain* serious organic diseases)

<mod2> (descriptional): bestemte *alvorlige* organiske sygdomme (certain *serious* organic diseases)

<mod3> (classificational): bestemte alvorlige *organiske* sygdomme (certain serious *organic* diseases)

<jj>, ad-adjectival, adjectives that modify other adjectives

On top of these syntactic tags, the adjective lexicon also contains some semantic tags. However, while DanGram's noun ontology[4] and Danish FrameNet (Bick 2011) have been used in numerous NLP projects (treebanks, CALL, MT etc.), so far no corresponding semantic system for adjectives has been published. Our current work strives to review, systematize and document existing semantic tags, and to introduce and implement a completely new ontology, more akin to the GermaNet system, where each category in addition to its semantic feature values also should allow the prediction of the semantic class of its typical head noun.

---

1 http://www.sfs.uni-tuebingen.de/GermaNet/adjectives.shtml (accessed 14 January 2019)

2 version 2.2 (https://cst.ku.dk/projekter/dannet/)

3 a Danish "word database" with 68,000 entries and morphological, syntactic and semantic information: https://cst.ku.dk/sto_ordbase/

4 http://visl.sdu.dk/semantic_prototypes_overview.pdf

3 Category scheme

In our proposed system, the primary semantic tags used for adjectives have the form <j....> and are combinatorially restricted feature prototypes, meaning that they specify a feature type of a certain semantic head (noun) class. For instance, <jshape> modifies concrete objects, and <jpsych> (psychological feature) combines with human heads (<H...>), but also actions (<act>) and semiotic products (<sem>).

There are 110-120 tags in all[5], most of which can be lumped in 14 or - with subclasses - 25 umbrella classes, most of them linked to prototypical head types. For instance, all tags within the *people* groups imply [+hum] (human), <jappro> (appropriateness) and <jbehave> combine with actions [+act], and <jsem> is about features of works of art, plans, laws or speeches [+sem]. For some category definitions and examples, see table 3.

- **people**: <jpsych> (feelings), <janat> (body features), <jage>, <jstate-h>, <jsick>, <jclo-h> (clothedness), <jappear> (appearance)
- **effecting**: <jaff> (affection), <jeff> (effecting), <jaff-h> (affected), <jimp> (important),
- **quality**: <jqual> (quality), <jpower>, <jskill>, <jappro> (appropriate), <jlike> (liked), <jreg> (regulated)
- **properties**:
  - *inherent*: <jprop>, <jtype>, <jbuild> (building), <jornam> (ornamental)
  - *+measure*: <jsize>, <jweight>, <jtemp> (temperature), <jspeed>,
  - *-measure*: <jshape>, <jsurf> (surface), <jsub> (composition), <jmat> (material), <jchem> (chemical), <jcol> (color), <jlight>
  - *state*: <jstate>, <jdam> (damage), <jnormal>, <jres> (result),
  - *sensed properties*: <jpercep> (perception)
- **quantity**: <jquant> (quantity), <jdegree>, <jcont> (content), <jsetop> (set operation), <jmanner-q>
- **identity**: <jident> (identity), <jauth> (authentic), <jcomp> (comparison), <jname>
- **cognitive**:
  - *thought*: <jcog> (cognitive), <jideo> (ideological), <jlike-h> (liking), <jmeta>
  - *speech*: <jcom> (communication), <jling> (language)
  - *epistemiological*: <jfact> (fact, true, likely), <jfame>
  - *semiotic* [+sem]: <jsem>, <jgenre>, <jdomain>, <jstruct> (structure)
- **event**: <jevent>, <jprocess>, <jchance>, <jchange>, <jcause>, <jsit> (situation)
- **doing**: <jact> (action), <juse>, <jhand> (handled), <jmove>, <jmanner>, <jbehave>, <jmethod>, <jres> (resulting), <jcrea> (created), <jlink>, <jtarget>
- **culture**:
  - *food*: <jfood>
  - *society*: <jsoc> (social), <jpol> (politics), <jinst> (institution), <jrel> (religious), <jprof> (professional), <jright> (entitled)
  - *domain jargon*: <jtech> (technical), <jjur> (law), <jmed> (medicine)

[5] This number of categories was deemed a reasonable level of granularity for empirical reasons. For practical purposes (parsing and corpus annotation), having too many increases the error rate in automatic tagging and risk introducing nuances that border on vagueness and often cannot be reliably distinguished by human annotators either. Too few categories, on the other hand, will mean a generalisation and abstraction level that misses out on many interesting semantic distinctions and is too course for contextual disambiguation tasks.

- *cultural products*: <jV> (vehicles), <jVwater> (ships), <jclo> (clothing features)
- *money*: <jmon> (money), <jmon-h>, <jposs> (owned), <jposs-h> (owning), <jval>(value)

- **nature**: <jbio>, <jA> (animals), <jB> (plants), <jL> (place feature), <jwea> (weather)
- **auxiliary**: <jbe>, <jcan> (possible), <jmust>, <jmay> (allowed), <jwill> (ready to)
- **space**: <jnat> (nationality), <jgeo> (geography), <jloc> (location), <jdir> (direction), <jori> (origin), <jpos> (position)
- **time**: <jtime>, <jord> (order), <jper> (period)

Sentiment and polarity markers

A number of feature types exhibit a plus/minus polarity, for instance <jtemp> (temperature: hot/cold), <jlike> (*liked* or *disliked*), <jappro> (*appropriate* or *inappropriate*). This polarity is resolved by means of <Q+> and <Q-> tags that are primarily meant as sentiment analysis tags, but will also double in almost all cases as polarity distinctors. "-h" marks a separate subclass for human heads, e.g. <poss> ("owned") and <poss-h> ("owning"). Where necessary, other, more specific, non-standard semantic head types can be added by means of a <H:...> tag, e.g. <H:furn> for "polstret" (padded).

4 Frames for adjectives

A small, but important, proportion[6] of Danish adjectives can take valency-governed arguments, almost all in the form of prepositional phrases (pp's). In these cases it is possible to say that the adjective is the core constituent of a predication, much like verbs or de-verbal nouns. We classify these constructions using an equivalent verbnet frame, and both frame and argument structure are provided in the adjective lexicon.

1. forelsket i *(in love with)* - FN:**like**/head§COG/i§TH [cognizer - theme]
2. bange for *(afraid of)* - FN:**emote_obj**/for§CAU/head§EXP [cause - experiencer]
3. benovet over *(embarrassed about)* - FN:**affect_exp**/head§EXP/over§CAU
4. beslægtet med *(related to)* - FN:**relate**/med§COM/head§TH [theme - co-argument]
5. blind for *(ignorant of)* - FN:**neglect**/for§TH/head§AG [agent - theme]
6. dygtig til *(good at)* - FN:**can**/head§AG/til§ACT'icl [agent - action]
7. sur på *(angry at)* FN:**emote**/head§EXP/på§CAU'H [experiencer - cause]
   sur over at *(angry because)* FN:**emote**/head§EXP/over§ACT'fcl [experience - action]
8. ond mod *(mean against)* - <FN:**affect_exp**/head§AG/mod§EXP'H> [agent - experiencer]
9. afhængig af - FN:**depend**/head§EXP'H/head§SOA'act/head§BEN/af§CAU
   person hooked on s.th. - FN:depend/**head§EXP'H**/head§SOA'act/head§BEN/af§CAU
   action depending on s.th. - FN:depend/head§EXP'H/**head§SOA'act**/head§BEN/af§CAU
   city relying on tourism - FN:depend/head§EXP'H/head§SOA'act/**head§BEN**/af§CAU

Each noun frame entry (FN) lists first the corresponding verb frame and then a slash-separated list of possible semantic role arguments[7] (marked §) with their slot filler conditions (1-9). We distinguish

[6] Currently, about 300 adjectives have been assigned frame-carrying valencies. As for verbs and nouns, structural complexity correlates with token frequency, so frame-capable adjectives are overrepresented in running text, with a token ratio higher than their type ratio.

between primary conditions and secondary, optional subconditions (present in 6-9). Primary conditions are placed before the role concerned, secondary condition after it. The former are syntactic slot conditions (either 'head' or a bound preposition lexeme), the latter are categorial conditions concerning semantic class (e.g. 'H'=human, 'act'=action), or form conditions such as ’icl’ (non-finite clause, 6) or ’fcl’ (finite clause, 7).

In the Danish data, adjectives only rarely have two completely different frames. More common are cases where there is some variation within the same frame, with different prepositions (7) or different semantic slot fillers (9) corresponding to different semantic roles. In these cases it is optional, whether frames are duplicated (7) or fused by appending argument variants (9).

5 Coverage statistics

In order to evaluate coverage, we tagged a Danish reference corpus consisting of DSL's period corpora, Korpus90, Korpus2000 and Korpus2010 (Asmussen 2015), covering modern post-war Danish up to the 90s and the years around 2000 and 2010, respectively. The first corpus has a broad genre and period scope, including some spoken data. The second is dominated by news and magazine texts and the third includes online material of various types. Together, the three corpora can be said to provide a fair cross-section of modern Danish.

Based on DanGram's morphological disambiguation, and a POS error rate under 1%, the corpus set contained 5.6 million adjective tokens distributed across 27,280 adjective types. In this count, hapaxes were ignored - inspection showed them to be mostly spelling errors and ad hoc foreign loan words. In about 1% of adjective tokens (37% of types), the parser had to use live compounding analysis[8]. Table 1 shows adjectival coverage percentages for both semantic class tagging (j-tags) and domain tagging (D-tags), first for all words, then separately for live compound analysis.

Tag type	**% tokens**	**% types**
semantic class tags (j-tags)	99.24	85.10
domain tags (D-tags)	95.64	73.89
j-tags / compounds	93.96	93.40
D-tags / compounds	75.99	76.82

Table 1: Corpus overage (all words)

As can be seen from the percentages, general running text coverage is very good (99.4%), but due to obvious Zipf-curve effects type coverage is considerably lower (85.1%). Live compounds have a worse token coverage, but better type coverage. Though surprising at first glance, this can be explained by the fact that the class-controlling second parts of compounds are dominated by relatively few, well-know suffixes and participles, leading to a good type-coverage. At the same time, because the individual compounds are all rare compared to ordinary adjectives, there is no pronounced positive effect of counting tokens rather than types.

If (a) purely heuristic (i.e. non-compound) analyses, (b) lexicon-registered erroneous forms and (c) foreign words are excluded (about 1,800 types or 11,500 tokens), coverage increases, as could be expected.

---

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

[8] These are cases, where a word was unknown in the sense, that it could not be reduced to a lemma or a compound found in the lexicon, but where the parser was able to come up with a likely compound analysis of its own at run time.

Tag type	% tokens	% types
semantic class tags (j-tags)	99.39	90.53
domain tags (D-tags)	95.82	78.78
j-tags / compounds	95.12	94.47
D-tags / compounds	77.01	77.75

Table 2: Corpus overage (recognized words and compounds only)

Table 3 contains a breakdown of the 22 statistically most important tag types by frequency (covering 80% of tokens and 52% of types), providing definitions and examples. For sense discrimination and other NLP tasks, it is an advantage that the category distribution curve is relatively even, with small differences between neighbouring frequencies, and even the top category below the 10% mark in token terms. By comparison, DanNet contains not only fewer items, but also exhibits a much steeper frequency curve, indicating less discriminatory power. Thus, when looking at type frequencies, our system "peaks" at 6%, with a spread over several, very different categories, while DanNet links 34% of adjective types to just "Property", and equally 34% to the hyperonym "beskaffenhed" (type). Even when classes and hyperonyms are combined, 23% are linked to combinations of Property/Property:physical and "beskaffenhed".

Tag	definition	% tokens	% types	examples
<jsize>	size	9.51	1.36	kæmpestor, lav, bred
<jqual>	quality	7.8	2.96	god, dårlig, ringe, pæn, smuk
<jnat>	nation, region, town	7.49	**5.81**	afghansk, chilensk, aarhusiansk
<jtime>	time	5.62	1.52	tyveårs, fortsat, sen, sjælden
<jstate>	state, non-human	4.69	2.04	frisk, åben, lukket, vakkelvorn
<jcog>	cognition	4.25	**3.19**	gennemtænkt, klar, enkel
<jquant>	quantity	4.2	0.74	halv, hel, rigelig, samlet
<jpsych>	psychological, feeling	3.79	**5.45**	vred, varmhjertet, arbejdsom
<jimp>	importance, impact	3.48	1.63	(u)vigtig, nødvendig, afgørende
<jage>	age	3.17	1.94	alderældst, attenårig, ung
<jord>	order (successive)	2.57	0.24	efterfølgende, gradvis, sidste
<jident>	identity	2.53	1.67	konkret, samme, selveste
<jsoc>	social	2.5	1.71	offentlig, privat, fri, uafhængig
<jappro>	appropriate	2.45	1.22	(u)egnet, rigtig, forkert, farlig
<jpol>	politics	2.11	1.56	sprogpolitisk, blokfri, autonom
<jnormal>	normal	2.09	0.56	almindelig, særlig, elementær
<jcol>	colour	2.04	3.8	grøn, lyseblå, ternet, tigerstribet
<jmanner>	manner	1.97	3.02	klodset, uorganiseret, mesterlig
<jfact>	fact, truth, probability	1.95	0.96	sand, korrekt, sikker, (u)mulig
<jdegree>	degree	1.92	1.29	gennemført, ekstrem, drastisk
<jbehave>	behaviour	1.86	3.27	anmassende, barbarisk, barnlig
<jtype>	type (underspecified)	1.67	**6.19**	-mæssig, kvindelig, -betonet
		79.66	52.13	

Table 3: Semantic class distribution

Some of the categories in table 3 have a much higher type/token ration than others, indicating a larger lexical spread, and more work for the lexicographer per annotated token. This is true not only for the

underspecified "type" category, but also for people's geographical provenance (<jnat>), cognition adjectives (<jcog>) and states-of-mind (<jpsych>).

6 Applications

The DanGram parser is used in a number of ongoing research projects, where improving adjective annotation might have an impact.

Greenlandic machine translation

Since Greenlandic linguistic tradition, based on morphological clues, does not recognize the existence of adjectives in the language, it is a non-trivial task to match Danish adjectives to Greenlandic lexical items. Often, the Greenlandic "adjective candidate" can translate into either a noun or an adjective in Danish. With a semantic-combinatorial classification of Danish adjectives, it might be easier to decide whether a word matches the semantics of a potential head noun, and hence should be treated as an adjective, or not.

Sentence grading

One interesting area within Intelligent Computer-Aided Language Learning (ICALL) is the automatic generation of exercises, and the grading of possible solutions. For instance, an ICALL system can generate sentences or question-answer pairs based on known vocabulary. If this is done solely based on syntactic slots, however, a large proportion of the suggested sentences will be meaningless. Thus, when using an adjective, it has to match the semantic type of its syntactic slot, normally defined by a noun. "Red ideas" and "angry houses" should be weeded out, while slight or metaphorical mismatches ("angry machines" or "red elephants", if recognized as such, might even contribute to making an exercise interesting and fun.

Sentiment analysis for hate speech

Hate speech research has lately drawn considerable public and political interest, as well as funding. Both in terms of technology (extracting and recognizing hate speech from online data) and linguistics, it is useful to be able to perform semantic annotation, and looking at what kind of adjectives are used in connection with hate speech target objects (immigrants, Muslims, Jews) is one way of decoding the linguistics of hate speech. Both sentiment analysis and adjective semantics are interesting in this regard, and to the best of our knowledge, no complete sentiment mark-up has ever been published for Danish adjectives.

7 Conclusions and outlook

We have presented a full-fledged lexico-semantic annotation scheme for adjectives and shown that the implemented Danish version can achieve 99% token coverage and 90% type coverage, while exhibiting a shallow frequency distribution curve with a high discriminatory potential.

It will be interesting to see if ongoing NLP work in the area of machine translation, semantic sentence grading and hate speech recognition can be made to profit from an improved lexical base for adjective annotation.

8 Acknowledgments

I would like to thank my colleague Anders Hougaard, now an associate professor at my university, for his valuable work on the semantic and syntactic classification of Danish adjectives during the early stage of the VISL project[9] in the late 90s. The domain tags mentioned in this paper are to a large degree motivated by and based on his contributions.

[9] http://visl.sdu.dk, a cross language grammar initiative at SDU with both an NLP and a teaching perspective

References

Asmussen, Jørg. *Corpus Resources & Documentation*. Det Danske Sprog- og Litteraturselskab, (http://korpus.dsl.dk, last updated 2018)

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

Bick, Eckhard. "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, 2011.

Bick, Eckhard. "En Constraint Grammar Parser for Dansk." In *8. Møde om Udforskningen af Dansk Sprog, 12.-13. oktober 2000*, ed. Peter Widell & Mette Kunøe, pp. 40-50, Århus University, 2001.

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

Hamp, Birgit, and Helmut Feldweg. "GermaNet - a Lexical-Semantic Net for German." In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid, 1997.

Nimb, Sanni, and Bolette S. Pedersen. "Towards a richer wordnet representation of properties – exploiting semantic and thematic information from thesauri." In *LREC 2012 Proceedings*. Istanbul, Turkey, 2012

Pedersen, Bolette S., Sanni Nimb, Jørg Asmussen, Nicolai H. Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. "DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary." *Lang Resources & Evaluation* 43, 269–299, 2009.

Vossen, Piek, ed. *EuroWordNet: A Multilingual Database with Lexical Semantics Networks.* Dordrecht: Kluwer Academic Publishers, 1998

Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. *FrameNet II: Extended Theory and Practice.* 2010 *(*http://framenet.icsi.berkeley.edu/ index.php?option=com_wrapper&Itemid=126)