Encoding Attitude and Connotation in wordnets

Anna Braasch and Bolette S. Pedersen
University of Copenhagen, Centre for Language Technology (CST)

The Danish wordnet, DanNet, though part of the global WordNet family, contains some information types that are not generally provided in wordnets such as qualia roles and connotation *of words. Connotation is seen as the set of associations implied by a lexeme in addition to its primary, literary meaning; it is evoked by one (or more) particular feature of the entity referred to and suggests attitudes, emotions and opinions like admiration or disapproval. Thus, lexemes with a connotation have an observable pragmatic effect in texts making them* subjective *or* opinionated.
In the paper, we discuss the relevance of connotation information in lexicons for computational applications in general and present the set of encoded semantic information exemplified by empirical data. We focus on a particular ontological type of entities, namely humans *with the focus on selected hyponyms of* person *that are encoded with a connotation value and discuss the prototypical properties evoking* positive *or* negative *connotations. The qualia structure based approach enables to encode both the prevalent, connotation evoking features and prototypical activities of the person.*
The material encoded with connotation so far consist of 650 nouns and comprises a male, a female and a gender-neutral group, thus it lends itself to comparative examinations concerning the distribution of connotation evoking features and polarity distribution within each individual group and between the groups as well. One of the most striking observations says that (in our material) the negative connotation polarity is predominant; the most important feature of female persons seems to be their positive appearance, and a general disparaging attitude dominates as regards the conduct and manners of male persons.

1. DanNet – a wordnet for Danish

DanNet[1] is a classical wordnet that conforms to the framework given in Fellbaum (1998) and Vossen (1999). In the line of these works, the synsets (synonym set) constitute the central element of the network where a set of previously established semantic relations constitutes the link between the synsets, the has_hyperonym relation being the central. The hyperonymy relation is further supplemented by antonymy relations, meronymy relations, as well as different kinds of functional relations. In contrast to most other wordnets, DanNet has been constructed using the so-called merge approach where the wordnet is built on monolingual grounds and thereafter linked to Princeton WordNet. Since the starting point of DanNet was a corpus-based, newly completed printed dictionary of Danish (*Den Danske Ordbog*; henceforth DDO) accessible in a machine-readable version with hyperonymy information explicitly specified for each sense definition, the motivation for the merge approach was obvious (cf. Pedersen et al. 2009). The approach can be seen in contrast to the more widely seen expand approach where synsets are translated from Princeton WordNet into the target language. The fact that a wordnet for Danish could be semi-automatically built from well-consolidated sense distinctions where the set of senses was actually defined on the basis of corpus data, made it feasible to build a monolingually based wordnet which would be practically useful in NLP tools meant for Danish text material. Apart from DDO that comprises approx 100,000 sense definitions, another resource has inspired the structure of DanNet, namely the SIMPLE lexicons (cf. Lenci et al. 2000); for example, from the SIMPLE model we have taken over the qualia structure (Pustejovsky 1995) comprising four roles. This has enabled us to expand the list of relations to encompass among others also those of the

[1] DanNet is developed in collaboration between Center for Language Technology at the University of Copenhagen and the Danish Society for Language and Literature. The project is granted by The Danish Ministry of Research. Download under an open source license from www.wordnet.dk.

constitutive role under which we organize features on connotation and sex as well as a so-called *concerns relation*, and it gives us a possibility to further specify which characteristics a connotation of a certain sentiment labeling relates to (i.e. behavior, sexual attitude, intelligence etc), see further in Section 3.

Currently, DanNet contains 60,000 synsets and is still under development within the DK-CLARIN project, until end of 2010. DK-CLARIN is the Danish branch of the EU project CLARIN which stands for a common language resources and technology infrastructure. The CLARIN project is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily usable; the Danish branch obviously focuses on the Danish language resources.

DanNet is foreseen to be integrated in NLP systems that include a semantic aspect, such as intelligent information navigation as well as writing aids. Currently, it has been integrated in the Danish version of OpenOffice where it is used as a facility to suggest broader and narrower terms.

2. Related work

A recent, rapidly growing sub-discipline of computational linguistics, opinion mining - also called sentiment polarity analysis - is concerned with the subjectivity and opinion that a text expresses. Subjectivity analysis determines whether the text content is presented in an objective (factual) or a subjective (opinionated) manner. A subjective text expresses a positive or negative opinion on its topic; this is also called *orientation* or *polarity* of the text as summed up in (Esuli & Sebastiani 2006a). In determining text orientation, a crucial task is to identify the connotation of opinionated words contained in the text. Another aspect, comprising the interaction of subjectivity and meaning and the importance of subjectivity annotation for word sense disambiguation, is discussed e.g. in Wiebe and Mihalcea (2006) with the following conclusion: 'Adding subjectivity labels to WordNet could also support automatic subjectivity analysis.' Further investigation (Akkaya, Wiebe and Mihalcea (2009)) are concerned with the automatic determination of subjective (opinionated) and objective senses of word instances in a corpus, a task called subjectivity word sense disambiguation.

The authors point to the promising perspectives of this method compared to previous work.

Application areas of opinion mining are e.g. ranking of internet pages and automatic opinion sampling previous to elections, as discussed in Esuli & Sebastiani (2007) and Attardi & Simi (2006). Studies within this field have shown that a lexical resource for computational applications containing information about the orientation or polarity of words (aka terms) is a pre-requisite for sentiment analysis and opinion mining. Tools for sentiment analysis, e.g. PolArt also use a so-called subjectivity or polarity lexicon (cf. Klenner & Fahrni 2009). SentiWordNet (cf. Esuli & Sebastiani (2006b)) is such a polarity lexicon which has been semi-automatically derived from Princeton WordNet providing not only the positive/negative polarity of a word, but also fine-grained information on polarity strength.

The DanNet approach of encoding information on connotation polarity differs in several aspects from SentiWordnet. In the following we describe a few characteristic features of our approach.

3. The connotation of words

One of the extensions to the general WordNet framework is to provide word senses implying an associative, subjective element (aka connotation, orientation, opinion) with information on whether the word sense is positive, negative or neutral wrt its connotation. Connotation, in our understanding (cf. Pedersen and Braasch 2009) is an additional dimension to the literal meaning of a word expressing a speaker/writer attitude to the denoted concept or entity, being it admiration, emotion, disapproval, judgment, etc. It is a pragmatic, extra-linguistic feature of words that is intuitively understood by native speakers but difficult to acquire for language learners and difficult to treat in natural language processing as well. We demonstrate the task of describing connotation as a binary category with the polarities *positive* and *negative*, though without determining the strength of the polarity in question, thus being different from SentiWordNet; our present framework is not prepared for polarity strength ranking.

On the other hand, we make use of qualia roles, esp. of the *concerns* relation of the *constitutive role*, which allows us to refer to the particular property (e.g. appearance, behavior, age) of the entity in question. Also, characteristic absence of a certain property can be often expressed here, e.g. *træmand* 'dry stick' (viz. 'a man without having feelings'), by negation the property *følelse* 'feeling' provided in the *concerns* relation. The *role_agent* relation of the telic role gives the prototypical activity (e.g. *drikke/drikke sig fuld* 'drink/ get drunk') of the denoted person. The explicit mention of prototypical properties and activities that evoke personal opinion - and thus form the basis of a given connotation - contributes to the particular qualities of DanNet.

We focus on a specific type of entities, namely *human*s, because nouns denoting persons are very frequent in language and more often than other ontological types, e.g. *food* and *vehicle* nouns, encompass connotations. The majority of person nouns (approx. 4,000 in DanNet) are objective, factual, without expressing any particular attitude to the referred entity, such as persons with a particular occupation, education, nationality, position, kinship, etc. These nouns are regarded *polarity neutral*, which is the default value in our encoding. Other types of person nouns frequently imply a connotation because humans judge each other by various remarkable or striking features in various social and communicative contexts. Here, we focus on selected hyponyms of *person* that are encoded with a connotation value because they exhibit an inherent speaker attitude to the denoted object, based on emotion or opinion like admiration, disallowance or even disdain; such nouns are also called *opinionated terms* (e.g. in Esuli & Sebastiani 2006b).

3.1. Method of identifying subjectivity and connotation polarity

In DanNet, the connotative information is manually encoded – like a number of other information types, and it is first of all based on the definition in DDO that is provided in the encoding tool, both in its full length and also in a shortened form as gloss to the word sense. In a large number of cases, a characteristic corpus example further supplements the definition.

Two types of connotation-related information may be present explicitly or implicitly in the definition and/or example and can thus be extracted from this material: the connotation polarity itself (viz. positive or negative) and the prototypical property (or activity) of the denoted entity evoking the connotation in question. In a very few cases only, a particular semantic label explicates the general, negative use of the word in the sense defined, such as *skældsord* ('infective') e.g. for *klaptorsk* ('silly ass') or *nedsættende* ('derogatory') e.g. for *so* ('slut', viz. an unclean, untidy woman'). This means that the definition/gloss of a noun with

positive or negative connotation contains at least one element, usually an adjective, and very often also some further elements we can make use of. An illustrating example that describes prototypical properties and/or activities of the person denoted: *rappenskralde* ('battleaxe') is defined as *arrig og rapkæftet kvinde eller pige* ('bad-tempered and cheeky woman or girl'); this tells that the referred object is a female person, and the negative adjectives relate to the person's behavior, disapproved by the writer/speaker. If there is neither explicit nor implicit information provided, we regard the word sense being objective without connotation.

Sometimes, however, the definition is lacking or if provided, it is not sufficient to determine the subjectivity and polarity of the word such as *organturist* 'organ (transplant) tourist', viz. being a person who travels to a foreign country and participates in black market organ donations as seller or buyer. In such cases, we first investigate the occurrences of the word in KorpusDK (http://ordnet.dk/korpusdk), a modern general language Danish corpus of 56 million tokens. A rather new word like *organturist* may however not occur at all in this corpus of which the most recent text is approx. 7 years old. Consequently, we have to search the word further, in newer Danish texts on the web where we find 60 reliable occurrences of this word and additionally 544 occurrences of the related activity *organturisme* ('organ tourism'), all exclusively occurring in an emphatic context, e.g. preceded by negative, condemnatory adjectives *umenneskelig, ulovlig, uetisk* ('inhuman', 'illegal', 'unethical'). This reflects not only a socio-cultural position, but also the personal opinion of the speaker/writer. In contrast, *organdonor* ('organ donor'; 12,200 occurrences on the web), *organtransplanteret* ('organ transplant/recipient'; 33 web occurrences) are not preceded by any qualifying adjectives at all; they are obviously neutral, objective terms. In these and similar cases, corpus evidence is the only means to identify subjectivity and connotation polarity. Corpus occurrences are of course also used to detect synonymy relations before assigning a word sense to a particular synset viz. set of synonyms.

3.2. The role of connotation in differentiating synsets

In DanNet, we have chosen to annotate word senses with connotation (as discussed e.g. in Wiebe 2006) for two reasons. First, a lemma might have different subjectivity values depending on the communicative context, e.g. *tøs* denoting a young female person, 'girl or young woman'; one is objective and thus neutral as regards connotation, though having a touch of positive orientation like 'a (very) young, (sweet) little thing', whereas the other one denotes a young woman behaving immorally and contemptibly, thus expressing a subjective opinion with a clearly negative polarity. In such cases, we have two senses and label one of them, here the last mentioned, with a negative connotation. Second, a given concept may be denoted by more than one single word - being synonyms; these are linked together in one synset in wordnets. Therefore, we are aware of the fact when establishing synsets that connotation is an extra-linguistic, paradigmatic component associated with particular word senses, thus this feature is encoded on the word sense level, and ideally, only word senses with identical connotation polarity should be linked together in one synset. This last point can be illustrated in some more detail with the following example. A person who frequently drinks large quantities of alcohol is called *alkoholiker* ('alcoholic') in objective, factual texts without a personal attitude to the topic/target, e.g. in medical reports or courtroom questioning; thus this word is neutral as regards connotation. On the other hand, a number of words denote the same concept, and in this sense they are quasi-synonyms, but they are not interchangeable with each other in all contexts, as their use is limited to different communication situations. Several general language words denote, too, a person addicted to alcohol: words like *drukkenbolt, drukmås, fulderik, fyldebøtte, kvartalsdranker,* etc. ('heavy drunker', 'sot', 'hard/heavy drinker', 'boozer', 'dipsomaniac') are used in everyday texts

reflecting both a strong socio-cultural and personal attitude towards such persons. In such cases, the connotation (polarity) is the feature differentiating between two synsets – one without connotation {*alkoholiker, alkoholist, alkoholmisbruger, dranker*} ('alcoholic', 'alcoholist', 'alcohol addict', 'drunkard') and the other synset with negative polarity {*drukkenbolt, drukmås, fulderik, fyldebøtte, kvartalsdranker*}. Figure 1 is a screen shot of this synset encoded in the DanNet tool[2] showing its own gloss, ontological type, connotation and the qualia roles with relations and values inherited from its immediate hyperonym {*alkoholiker…*} and further hyperonyms in turn, {*misbruger*} ('addict') and {*hoved,…*} ('person',…) as well. These are the constitutive role/concerns relation with the values *adfærd* 'behavior, conduct' and *afhægighed* 'dependence' and typical activities provided in the telic role/role_agent relation with the values *drikke* ('drink (too much)') and general human activities such as *tale, tænke* and *leve* ('speak', 'think', 'live') as well.

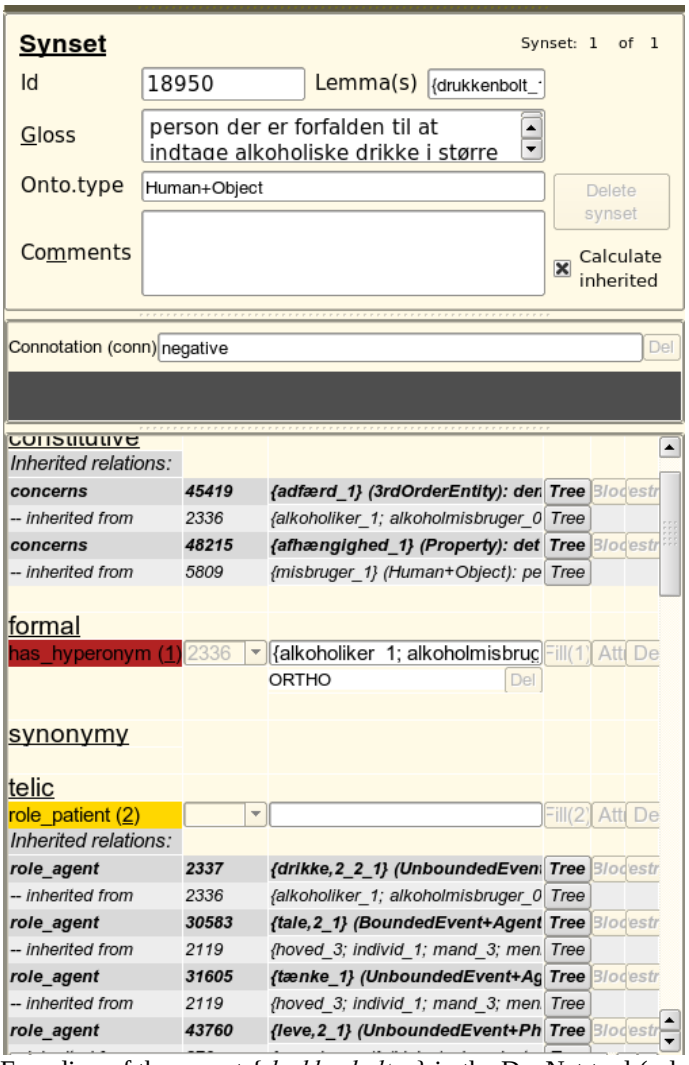

Figure 1. Encoding of the synset {*drukkenbolt…*} in the DanNet tool (selected part)

Looking up these words in KorpusDK, we can confirm the reasonability of establishing two synsets: only 37 (viz. 8.2 %) of the 448 corpus occurrences of *alkoholiker* contains a preceding adjective, and most of these do not express any personal opinion but state e.g. the gender or a kind of condition of the person, such as *passiv, tørlagt, forhenværende, latent, hjemløs, gammel, brutal, hærdet* ('passive', 'dry', 'former', 'latent', 'homeless', 'old',

[2] The DanNet tool is developed by Nicolai Hartvig Sørensen, Society for Danish Language and Literature.

'brutal', 'hard-bitten'), etc., only five of the 25 different co-occurring adjectives reflect the opinion of the speaker/writer: *sølle, desperat, forsumpet, selvforkælende, værdiløs* ('poor', 'desperate', 'apathetic', 'down-at-heel', 'self-pampering', 'worthless'), all occurring only once, thus the percentage of the opinionated, viz. negative occurrences of *alkoholiker* is ~1.1 % (of the 448 total). In contrast, the co-occurrence of adjectives with members of the second synset with inherent negative connotation shows a quite different pattern.

Lemma	Corpus occ.	Occ. with adj.	Distribution of adjective polarities on corpus occurrences
alkoholiker (alcoholic)	448	37 (8 %)	***Neg: 5 (1 %)*** Neu: 32 (7 %); Pos: 0
alkoholmisbruger (alcohol addict)	25	6 (25 %)	***Neg: 1 (4 %)*** Neu: 5 (21 %); Pos: 0
alkoholist (alcoholist)	13	2 (15 %)	***Neg: 0*** Neu: 2 (15%); Pos: 0
dranker (drunkard)	83	12 (15 %)	***Neg: 3 (4 %);*** Neu: 9 (11 %); Pos: 0
drukkenbolt (heavy drunker)	59	13 (22 %)	***Neg: 9 (15 %);*** Neu: 3 (5 %); Pos:1 (2 %)
drukmås (soak)	24	9 (38 %)	***Neg:7 (30 %);*** Neu: 1 (11 %); Pos:1 (11%)
sut (boozer)	12	8 (65 %)	***Neg:7 (57 %);*** Neu: 1 (8 %); Pos: 0
fyldebøtte (sot)	7	1 (15 %)	***Neg:1 (15 %);*** Neu: 0; Pos: 0
kvartalsdranker (dispomaniac)	7	2 (28 %)	***Neg: 1 (14 %);*** Neu: 1 (14 %); Pos: 0
spritter ((meths) drinker)	9	3 (33 %)	***Neg: 1(11 %);*** Neu: 2 (22 %); Pos: 0
fulderik (drunk)	33	3 (9 %)	***Neg: 1 (3 %);*** Neu: 2 (6 %); Pos: 0

Table 1. Persons addicted to alcohol and the preceding negative/neutral /positive opinion adjectives

Klenner & Fahrni (2009) point to the fact that word polarities are combined to NP, VP and sentence polarities saying that sentiment orientation is compositional. They describe the compositional regularities for NP's consisting of an adjective and a noun and point to the polarity defining role of the adjective. In contrast, person nouns in our material exhibit a strong inherent connotation, and they usually 'select' preceding adjectives with a connotation polarity identical to their own. The distribution figures of adjective polarities (Table 1) may also indicate a considerable variation of polarity strength, e.g. *sut* where 57 % corpus occurrences contain a negative adjective like *forhutlet, forsumpet* ('shabby', 'seedy'), whereas the comparable figure for *fulderik* is just 3 % viz. the (slightly) negative adjective *højrøstet* ('noisy'). This point, however, has not been examined here in detail.

An interesting case is represented by the word *dranker* ('drunkard'), which is encoded as a member of the neutral, not-opinionated synset denoting a person addicted to alcohol, though the percentage of negative preceding adjectives is considerably higher (4 %) than the corresponding figure for *alkoholiker* ('alcoholic') (~1 %). The big difference in the number of corpus occurrences (83 against 448) makes the comparison of the two negative adjective percentages probably unreasonable. Therefore, we searched for some additional, decisive context features.

Complementary information concerning the semantic characteristics of *dranker* and *alkoholiker* is found in the corpus by looking at their corpus evidences where they occur with other nouns, connected by coordinating *og* ('and') or disjunctive *eller* ('or') conjunctions. A large common set of such frequently co-occurring objective nouns such as *narkoman, kriminel, husvild, prostitueret* ('drug addict', 'offender', 'homeless', 'prostitute') justifies the decision to include also *dranker* into the objective synset. At the same time, *dranker* seems to be used more informally, as three negative opinionated words viz. *luder, plattenslager* and *junker* ('whore', 'trickster' and 'junkie') appear too in the near context of *dranker*, though each once only.

A word sense with a connotation (positive or negative polarity) is encoded as a hyponym to an objective, not opinionated word sense denoting the same entity, without expressing a speaker attitude or judgement. In the case discussed here, the synset {*alkoholiker…*} is hyperonym of two negative synsets: {*drukkenbolt, drukmås, fulderik, fydebøtte, kvartalsdranker*} and {*bums, sut, spritter*}. The reason for encoding two synsets with negative connotation as near-synonyms is that the members of the last mentioned synset have an additional specifying characteristic being *subsistensløs* (‘destitute’) or *hjemløs* (‘homeless’). The screen shot from DanNet Hyperonymy Visualizer (Figure 2) illustrates this structure and shows also the hyperonym of the synset {*alkoholiker…*}, which is *misbruger* ‘addict’.

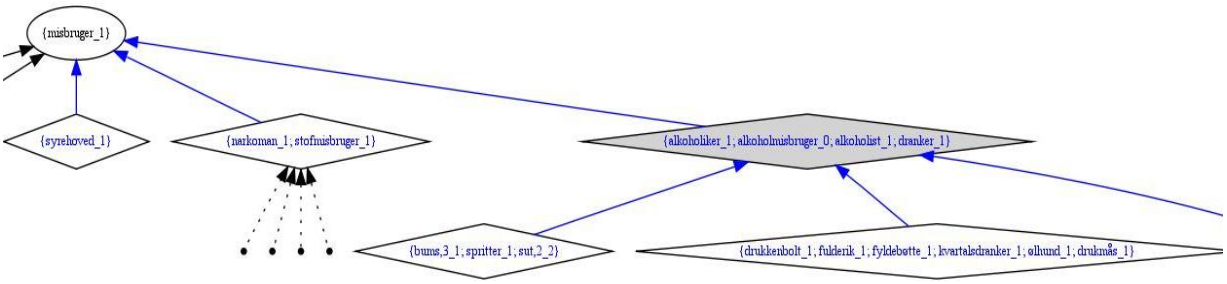

Figure 2. Hyperonymy structure of *alkoholiker* as shown by the visualizer of the DanNet tool

A DanNet browser[3] (www.andreodk.dk) can be accessed on-line; it renders intelligible the gloss, semantic structures and relations of the lemmas, presenting the last release (update by 30 March 2010) of the open source version of the DanNet data to the user.

Obviously, many of the opinionated terms are used in informal, colloquial communication, slang or even in abusive language. However, in DanNet we do not establish different synsets on the basis of usage differences only. This means that terms both used in general, everyday, informal and colloquial language and in slang as well may appear in the same synset if they denote the same entity and share connotation value (or they are not opinionated.) The following two synsets illustrate this method. All members of the first one {*bil, vogn, øse, karet, slæde…*} mean ‘car’, and all are objective, not-opinionated terms, though the first two are general language words, whereas the remaining three belong to slang, whereas the word senses in the second synset {*spand, skramlekasse, smadderkasse*} (‘crate’, ‘old crock ‘, ‘banger’) have all negative connotation meaning a bad, old, battered car.

4. Taking stock of the material encoded in DanNet

Currently, 650 person nouns (~16 % of the total 4,000) are encoded with connotation information; hereof are 97 specified by their nearest hyperonym as female and 76 as male persons, the rest (475) can denote both female and male persons. In DanNet (as mentioned in Section 1), the *constitutive role* encompasses semantic relations and properties that form the internal structure of the concept. The *concerns relation* of this role is used to express the feature that evokes the connotation associated with the noun such as appearance, intellect, stature, behaviour, sexuality, temper, manners, morals, physical power, and *connotation* is an *attribute* to the constitutive role that holds a positive or negative value.

[3] The browser is developed by Anders Johannsen, University of Copenhagen in 2009.

4.1. Polarity distribution

A current comparison of the general distribution of connotation polarities for person nouns provided with this attribute is the following: 506 nouns (~78 %) have negative connotation and 144 (~22 %) have positive connotation, i.e. the number of nouns with negative connotation is 3.5 times as high as with positive.

In more detail, also the distribution of connotation polarities on the female, male and gender-neutral groups shows some obvious tendencies, though the encoding of connotation is currently in progress, thus the figures are still preliminary. The connotations are predominantly negative in each group being between 67 % and 81 %; accordingly, positive connotations range from 18 % to 30 %.

4.2. Distribution of connotation evoking features

The connotation evoking features (viz. provided values in *concerns*) in the female and male groups can be roughly systematized in eight and nine major, generalized features respectively, whereas the gender-neutral group is more scattered with a distribution on 18 (or even more) generalized features. Defining a generalized feature means that closely related evoking features are grouped together in one major feature for the sake of clarity, although they may focus on different details or aspects, e.g. *appearance/shape/stature* comprises also body weight, height, clothes, etc. The respective distribution figures are summarized in two gender-specific tables that make comparisons easy.

Table 2 provides an overview of the connotation evoking features and the distribution of polarities for female persons. A connotation may be evoked by more than one single prevailing feature; in such cases the noun is encoded with more than one *concerns* value. The total of *concerns* is therefore higher (108) than the number of females with connotation (97). The same goes for male persons, cf. Table 3.

Evoking feature based on *concerns*	**Evoked + connotation**	**Evoked - connotation**	**Number of nouns with the feature**	**% of nouns with the feature**
Appearance/shape/stature	18	13	31	28.7 %
Sexual behaviour/sex appeal	7	16	23	21.3 %
Temper/mind/character	5	14	19	17.6 %
Conduct/manners	1	16	17	15.7 %
Status/function/efforts	3	5	8	7.4 %
Intellect/ability	0	5	5	4.6 %
Age/experience/maturity	1	2	3	2.8 %
General	1	1	2	1.8 %
TOTAL	**36**	**72**	**108**	**100%**

Table 2. Connotation evoking features of females

Evoking feature based on *concerns*	**Evoked + connotation**	**Evoked - connotation**	**Number of nouns with the feature**	**% of nouns with the feature**
Conduct/manners	3	23	26	30 %
Appearance/shape/stature	6	9	15	17 %
Sexual behaviour/sex appeal	2	12	14	16 %
Age/experience/maturity	0	10	10	11.5 %
Temper/mind/character	2	8	10	11.5 %
Status/function/efforts	0	5	5	5.7 %
Physical power	3	0	3	3.5 %
Intellect/ability	1	1	2	2.3 %
General	2	0	2	2.5 %
TOTAL	**19**	**68**	**87**	**100 %**

Table 3. Connotation evoking features of males

The material encoded lends itself to be examined more in detail. Comparing the figures in Table 2 and Table 3, following specialties can be observed.

- The ratio of positive/negative connotations in the female group is exactly 1:2, whereas the corresponding ratio for males is 1:~3.5; in other words, there are at the overall level significantly more male nouns than female ones with negative polarity.

- The connotation evoking features describing male and female persons show slightly different weight and order of priority in our material, but the order of their magnitude is rather similar, viz. the most frequently encoded feature represents around 30 % of the connotation evoking features and the slightest ones are around 2 %. A feature encoded particularly for male persons is, not surprisingly, physical power, though with a modest occurrence percentage only.

- Female persons are predominantly judged by their appearance (28 %) and sexual behaviour (21 %), whereas for male persons the most prevalent features evoking the connotation are their conduct/manners (30 %) and appearance (17 %). Interestingly, the percentage sums of the two predominant features in the respective groups are very similar: for females 49 % and for males 47 %.

- As regards the distribution of positive/negative polarities of the appearance feature, the two noun groups show a striking difference. Females are described mainly with positive terms (18), even though the number of terms with negative polarity is only ~35 % lower (13). For males, the distribution has quite different figures: the negative polarity is by far the most dominant (23), only three nouns of males have a positive polarity. The above differences may lead us to a conclusion saying that for females, positive appearance is a predominant feature, whereas for males the appearance is less striking, and most of the opinionated terms alluding to men's appearance have negative polarity.

- Conduct/manner seems to be the most specific connotation evoking feature of males, with approximately twice as high percentage of male nouns (30%) as females (15,7 %) In other words, people's conduct and manners seem to evoke in general a negative connotation, as the ratio of positive/negative polarities is 1:16 for female nouns and 1:7.7 nouns.

As regards the largest, gender-neutral group of nouns, the list of connotation evoking features is much more comprehensive and varied, also because of the diversity of the persons denoted. In this group, the most frequently encoded features are conduct/manners, appearance/shape/stature, intellect/ability, general morality, temper/mind/character, age/experience/maturity. Further candidates for the top-ten-list of features are criminality, social and economic status, performance/success/failure, state of health/mind, personal attitude/opinion. The feature sexual behaviour/morals/sex appeal seems to be less prevalent for these nouns, obviously because they do not imply the gender of the person denoted.

5. Summing up

In DanNet, we decided to provide the connotative information that goes beyond the pure denotation of words. The utility of this supplementing information is investigated in relation to some other wordnets and tools, and we have stated that information on subjectivity and connotation polarity provided in a lexicon helps to identify the attitude or bias of a text. We

have shown that the DanNet framework is well-suited to represent both connotation polarity and features evoking the connotation. We have outlined the methodology adopted, which employs information provided in dictionary definitions and corpus evidences. Further, we have discussed the important role of connotation in differentiating synsets that denote the same entity.

We have developed our points further and illustrated the encoding of connotation in DanNet by examples of person nouns. The figures of connotation evoking features of gender-specific and gender-neutral nouns have been discussed and compared on the basis of overview tables. This has led us to some interesting observations about the gender-specific distribution of these features and the corresponding polarity distribution as well. The figures suggest that more than three fourths of the opinionated nouns denoting humans have a negative polarity, and the most prevalent features evoking connotations are different for males and females. A further relevant observation is that the prevalent polarities for the same connotation evoking feature are highly different for the two genders: female appearance seems to be remarkable, if it is positive, whereas remarkable male appearance is negative.

The examination of the material encoded so far has shown some clear and interesting tendencies as regards distributional matters, and some preliminary conclusions on the distribution of connotation on evoking features and polarities could been presented. The work is still in progress, and we expect that a growing number of words, including adjectives provided with connotation information will form a fully reliable basis for comprehensive and valid final conclusions.

References

Akkaya, C.; Mihalcea, R.; Wiebe, J. (2009). 'Subjectivity Word Sense Disambiguation'. http://www.cs.pitt.edu/~wiebe/pubs/pub1.html [access date: 15 February 2010].

Attardi, G.; Simi, M. (2006). 'Blog Mining through Opinionated Words'. http://trec.nist.gov/pubs/trec15/papers/upisa.blog.final.pdf [access date: 15 February 2010]

Esuli, A.; Sebastiani, F. (2006a). 'Determining Term Subjectivity and Term Orientation for Opinion Mining. http://acl.ldc.upenn.edu/eacl2006/main/papers/13_1_esulisebastiani_192.pdf [access date: 15 February 2010].

Esuli, A.; Sebastiani, F. (2006b). 'SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining'. http://www.isti.cnr.it/People/F.Sebastiani/Publications/LREC06.pdf [access date: 15 February 2010].

Esuli, A.; Sebastiani, F. (2007). 'PageRankingWordNet Synsets: An Application to Opinion Mining'. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague. 424–431.

Fellbaum, C. (ed.; 1998). *WordNet – An Electronic Lexical Database*. Cambridge, Massachusetts / London, England: The MIT Press.

Klenner, M.; Fahrni, A. (2009). 'PolArt: A Robust tool for Sentiment Analysis'. In *NODALIDA 2009*. Odense.

Lenci, A.; Bel, N.; Busa, F.; Calzolari, N.; Gola, E.; Monachini, M.; Ogonowski, A.; Peters, I.; Peters, W.; Ruimy, N.; Villegas, M. & Zampolli, A. (2000). 'SIMPLE – A General Framework for the Development of Multilingual Lexicons'. In: *International Journal of Lexicography* 13. 249–263.

Pedersen, B.S.; Braasch, A. (2009). 'What do we need to know about humans? A view into the DanNet Database'. In *NODALIDA 2009*. Odense.

Pedersen, B.S.; Nimb, S.; Asmussen, J; Sørensen, N.; Trap-Jensen, L.; Lorentzen, H. (2009). 'DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary'. In: *Language Resources and Evaluation*. Springer.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, Massachusetts: The MIT Press.

Vossen, P. (ed.; 1999). *EuroWordNet, A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.

Wiebe, J.; Mihalcea, R. (2006). *Word Sense and Subjectivity. Joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics. (COLING-ACL 2006)*. http://www.cs.pitt.edu/~wiebe/pubs/pub1.html [access date: 15 February 2010].

Electronic resources

DanNet = The Danish WordNet. Download under an open source license from http://www.wordnet.dk

DanNet browser. http://www.andreord.dk [access date: 18 February 2010].

DDO = *The Danish Dictionary*. http://ordnet.dk/ddo/

KorpusDK= Danish Corpus. http://ordnet.dk/korpusdk/ [access date: 15 February 2010].

Princeton WordNet. http://wordnet.princeton.edu/ [access date: 1 February 2010].