DanNet: From Dictionary to Wordnet

Jörg Asmussen
Society for Danish Language and Literature, Copenhagen, ja@dsl.dk

Bolette Sandford Pedersen
Centre for Language Technology, University of Copenhagen, bolette@cst.dk

Lars Trap-Jensen
Society for Danish Language and Literature, Copenhagen, ltj@dsl.dk

Abstract

This paper deals with some of the methodological implications of constructing a wordnet by reusing a monolingual dictionary originally designed for human users. It illustrates and discusses both advantages and disadvantages of the chosen approach and sketches a number of solutions to cope with some of the downsides of this model.

1 Introduction

DanNet is a lexical-semantic wordnet for Danish. It is being constructed as a joint project between two linguistic research institutions: *Center for Sprogteknologi, Københavns Universitet* (Centre for Language Technology, University of Copenhagen, CST) and *Det Danske Sprog- og Litteraturselskab* (Society for Danish Language and Literature, DSL). The project runs for a period of four years (2005–2008) with a grant of DKK 3 million (approximately €400,000) funded by the Danish Research Councils.

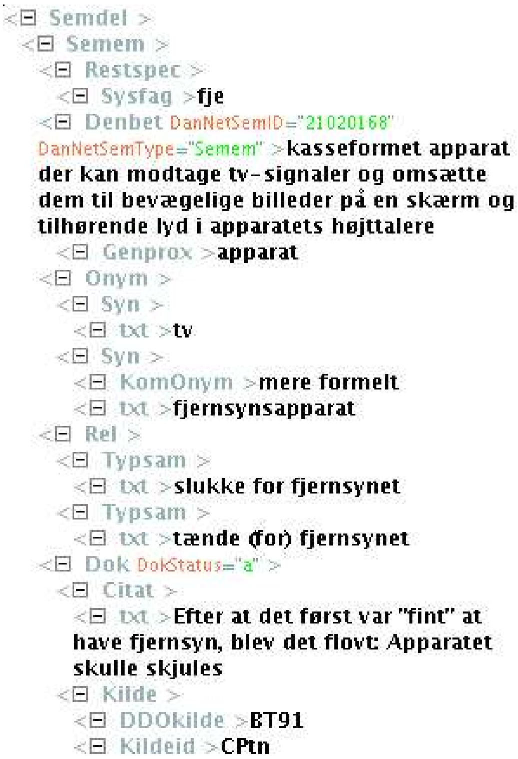
In terms of economic resources, the DanNet project is quite limited which made it an indispensible demand for it to reuse already existing lexical-semantic language resources. One way of reusing material would have been to adapt one of the many existing wordnets to Danish, though the outcome of such an approach is likely to be biased by the lexical-semantic structure of the input wordnet rather than reflecting the one inherent to Danish. I order to avoid such influences, it was decided to reuse existing lexical-semantic resources for Danish, namely *SIMPLE-DK* (Pedersen and Paggio, 2004) giving a detailed formal semantic description of approximately 10,000 concepts and – as the main source – *Den Danske Ordbog* (The Danish Dictionary, DDO, Hjorth et al. (2005)), a new written-from-scratch corpus-based dictionary of contemporary Danish. In this respect, the DanNet project methodologically differs from most other wordnet projects as it does not reuse (and expand) the existing Princeton Wordnet but employs a strictly monolingual starting point.

In the following we will describe some of the structural properties of this dictionary and show how and why it in principle can serve as a useful input to DanNet, and we will show the types of problems that arise and how to cope with them.

2 Characteristics of the DDO

DDO – the first and only truly corpus-based dictionary of Danish – is a printed dictionary in six volumes compiled by DSL and published 2003-2005 (Lorentzen, 2004). It comprises approximately 100,000 words described in 60,000 entries. It gives detailed information on spelling, morphology, pronunciation, meaning, collocations, fixed phrases, syntax, usage, word formation and etymology, and thus addresses a wide variety of potential users. The dictionary is principally based on the Corpus of the Danish Dictionary (DDOC), a reference corpus of contemporary Danish (Norling-Christensen and Asmussen, 1998).

In order to achieve a high level of consistency in the semantic description, the dictionary entries were written in groups of semantically related words rather than in alphabetical order. Templates for sense description were developed and applied for the individual groups. Function words were edited in groups of word classes. In order to secure a future data reuse possibility, a fine-grained microstructure was designed, including also elements not meant for presentation in the printed dictionary. Even if the dictionary so far is only publicly available as a printed edition, it was edited in a machine-readable format (SGML/XML), and a first online version



```
<⊟ Semdel >
  <⊟ Semem >
    <⊟ Restspec >
      <⊟ Sysfag >fje
    <⊟ Denbet DanNetSemID="21020168"
    DanNetSemType="Semem" >kasseformet apparat
    der kan modtage tv-signaler og omsætte
    dem til bevægelige billeder på en skærm og
    tilhørende lyd i apparatets højttalere
      <⊟ Genprox >apparat
    <⊟ Onym >
      <⊟ Syn >
        <⊟ txt >tv
      <⊟ Syn >
        <⊟ KomOnym >mere formelt
        <⊟ txt >fjernsynsapparat
    <⊟ Rel >
      <⊟ Typsam >
        <⊟ txt >slukke for fjernsynet
      <⊟ Typsam >
        <⊟ txt >tænde (for) fjernsynet
    <⊟ Dok DokStatus="a" >
      <⊟ Citat >
        <⊟ txt >Efter at det først var "fint" at
        have fjernsyn, blev det flovt: Apparatet
        skulle skjules
      <⊟ Kilde >
        <⊟ DDOkilde >BT91
        <⊟ Kildeid >CPtn
```

Figure 1: Semantic description in the DDO

is planned to be launched by the end of 2007. In a longer perspective, it is obvious for the online dictionary also to make use of DanNet data, e.g. by facilitating onomasiological queries as well.

DDO comprises approximately 100,000 well-structured semantic descriptions including sense definitions. The idea is to reuse a substantial part of them as concepts/synsets in DanNet. In Figure 1 an example is given to show the semantic description of the sense 'tv set' in the entry for *fjernsyn* 'tv'. Let us briefly comment on the different elements used in the entry.

The first completed element is <Sysfag> which is one of the above-mentioned non-visible elements, indicating in this case systematic domain information. This element has been filled in whenever possible, but it is only displayed in the printed dictionary if the sense in question is used technically, e.g. as a professional term. In a DanNet perspective, domain information is used for the relation *concerns*. In Figure 1, the element has been filled in with *fje*, a short code for 'television'. During the coding process the interface editing tool will prompt the coder that 'television' is a candidate to consider for the relation *concerns* of the synset 'tv set'. The relation *concern* is meant to refer to the general domain area of the synset. In some cases it overlaps partially with the information indicated by the ontological type, as in *egnsret* 'regional dish' which has the ontological type Comestible and the <Sysfag> information 'food'. In most cases, however, it adds new, relevant subject information to the synset, as in the case of *fodboldstøvle* 'football boot' with the *concerns* relation 'sports' (and the ontological type: Artefact+Object).

The element <Denbet> contains the sense definition and is followed by yet another non-visible element <Genprox> indicating the hyperonym of the sense as used in the definition. We will come back to this in more detail in Section 3. The element <Denbet> furthermore carries two DanNet-related attributes that ensure a fixed linkage between the DanNet concepts and the definitions in the DDO. Thus DanNet, when finished, can be used as a special semantic or onomasiological acces to an electronic version of the DDO.

The element <Onym> contains information on sense relations. In the example, two synonyms are given, *tv* and *fjernsynsapparat* (both meaning 'tv set'). Synonym information is convenient to reuse in DanNet. In this case the two synonyms can be included directly in the appropriate 'tv' synset together with the headword *fjernsyn*. In the same manner antonym information from dictionary entries (given in a separate element <Ant>) can be reused for establishing the corresponding antonym relation in DanNet. Finally, DDO also contains information on near-synonymy, and this is specified in DanNet as a separate relation.

In Figure 1 the elements <Typsam> contain collocational information, and finally a citation is given in the element <Citat>. These last elements are, however, less relevant for DanNet purposes.

It can be seen how some elements in the microstructure can be exploited more or less directly in a wordnet setting. However, most of the relevant information is not coded in separate elements in the entry, but must be deduced from the sense definitions. Consequently, it is the job of the DanNet editors to extract "manually" the relevant semantic information and convert it into the relations that are crucial to a wordnet, above all the taxonomic structure as reflected by the language. Some considerations on how this process might be facilitated by "automatic" means is given in Section 4.

3 Extracting the hyponymy structure

As has been shown, the <Denbet> element which holds the dictionary definition of a certain lexical sense, has a sub-element, <Genprox>, giving a lexicalised expression for the genus used in the definition. Wherever possible, definitions have been composed according to the classical scheme of giving the closest hyperonym of the definiendum, the *genus proximum*, and the *differentia specifica* that distinguishes the definiendum from its co-hyponyms.

In the example shown in Figure 1 *apparat*, '(technical) device', is given as genus (or rather as one of the senses that the lexicalised expression *apparat* may have), whereas some of the other information given in the definition must belong to its differentia. All genus specifications have been extracted from the <Genprox> elements together with the definition proper and other semantic information and have been stored in the DanNet database which is accessible through a special interface allowing the editors to manipulate the data in order to build the wordnet. When a new synset is established, the interface automatically suggests probable genuses – a method that speeds up the process of establishing the hyponymy structure of the wordnet. The fixed linkage between the dictionary definitions and the wordnet concepts during editing is also helpful in this process as it always ensures necessary information right at hand.

In the unproblematic cases, the task of the editor is solely to assign the right concept/synset behind a DDO-given genus expression. This includes disambiguation of possible homonymy or polysemy of the genus expression. For example, the word *cell* is used as genus proximum in both the synset 'yeast cell' and 'prison cell', but clearly it has to do with different senses of the word *cell.*

Another more cumbersome task is the general harmonisation of the wordnet to ensure the construction of a reasonably consistent and computationally processible taxonomy. The genus expressions assigned in the DDO were not taken from a predefined set of ontological concepts, but were rather decided upon by the individual dictionary editors on the basis of some general guidelines. To give an example, this means that one editor has chosen the term *lære* 'studies' as a superordinate concept to *informatik* 'informatics' and *bromatologi* 'nutrition science', whereas another has used the term *fag* 'subject' to cover *samfundsfag* 'social studies' and yet another *videnskab* 'science' as hyperonym to *datalogi* 'computer science'. In such cases the DanNet editor will try to change or merge the hyperonyms in order to provide a more balanced network. Often this is done by merging expressions like *fag*, *lære* og *videnskab* into one synset as long as there does not seem to be any consistent principles speaking in favour of maintaining them as separate ones. Likewise, it seems to be arbitrary when *gadespejl* 'outside mirror' is considered to be an *indretning* 'appliance' whereas *støttefod* 'kickstand' is defined as an *anordning* 'arrangement'. Or when *spisekort* 'menu card' is described as an *oversigt* 'list' whereas (the synonym) *menukort* 'menu card' is defined as a *kort* 'card'. If the wordnet is meant for more flexible information retrieval, a harmonisation of hyperonyms is crucial since it will otherwise be very hard to calculate for instance semantic similarity between concepts on reliable grounds.

On the other hand, the DanNet editors are cautious regarding the introduction of other and more fine-grained hierarchies than the ones given in the DDO; generally we strive towards making explicit already given information rather than building deeper hierarchies. To illustrate this, consider the example *stol* 'chair'. From the DDO the following hyperonyms are derived from *stol*: > *siddemøbel* 'sitting furniture' > *møbel* 'furniture' > *genstand* 'object' > *top*. However, in our vocabulary we find also the synonyms *bohave* and *indbo* 'household effects' which are terms typically used in insurance business referring to all moveable items in the home. When coming across such terms in the encoding phase, one could be tempted to introduce new levels in the hierarchy including them, e.g. *stol* > *siddemøbel* > *møbel* > *indbo/bohave* > *genstand* > *top*; however, in such cases we have chosen to

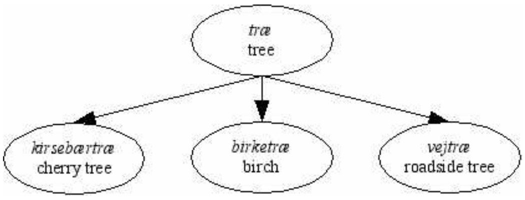


Figure 2: Some hyponyms of *træ* ‘tree’

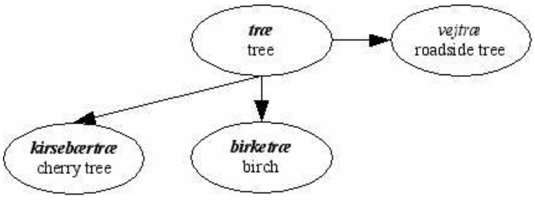


Figure 3: “Orthogonal” hyponymy

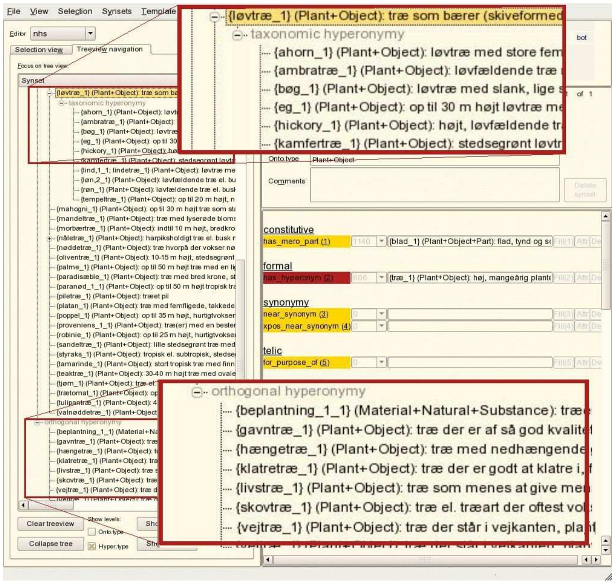
stick to the original DDO organisation since it seems to better represent the intuitive position of the non-specialist use of the language.

As indicated above, the treatment of the ontological structure of the network is generally a central issue in the DanNet project, and in particular the hyponymy relation calls for further investigation.

Interesting patterns emerge when extracting large groups of hyponyms from the DDO; for instance it becomes clear that hyponymy covers several sub-relations, one corresponding to inclusion of formal ontology and others rather denoting different dimensions or roles of the hyperonym.

Where hyponymy can generally be defined as *X is a Y*, taxonomy can be seen as a subtype to hyponymy with the definition *X is a kind/type of Y* (Cruse, 1991; Cruse, 2002). In formal ontologies, this proves to be a crucial distinction since only the latter corresponds to proper inclusion where the inheritance structure is straightforward. Consider the example in Figure 2: Since *vejtræ* ‘roadside tree’ differs from the other hyponyms by not being a type of tree but rather any kind of tree standing in the road side, we have decided to give it another ontological status in the hierarchy. In DanNet we consider it to be “orthogonal” to the rest of the taxonomy (cf. Figure 3), i.e. as a hyponym which describes another meaning component than the rest of the hyponyms. For a full account of this approach based on Cruse’s classification, we refer to Pedersen and Sørensen (2006).

Where taxonomy refers to *X is a kind of/type of Y*, non-taxonomic hyponyms can be defined as *X is (more or less) any kind of Y for which it is the case that...* This means that in fact both cherry trees and birches can function as road side trees. Figure 4 shows how this is reflected in the DanNet interface.



{løvtræ_1} (Plant+Object): træ som bærer (skiveformed
taxonomic hyperonymy
{ahorn_1} (Plant+Object): løvtræ med store fem
{ambratræ_1} (Plant+Object): løvfældende træ
{bøg_1} (Plant+Object): løvtræ med slank, lige s
{eg_1} (Plant+Object): op til 30 m højt løvtræ me
{hickory_1} (Plant+Object): højt, løvfældende tr
{kamfertræ_1} (Plant+Object): stedsegrønt løvtr

Onto.type Plant+Object

Comments

constitutive
has_mero_part (1) 1140 {blad_1} (Plant+Object+Part): flad, tynd og s

formal
has_hyperonym (2) 606 {træ_1} (Plant+Object): høj, mangeårig plante

synonymy
near_synonym (3) 0
xpos_near_synonym (4) 0

telic
for_purpose_of (5) 0

orthogonal hyperonymy
{beplantning_1_1} (Material+Natural+Substance): træ
{gavntræ_1} (Plant+Object): træ der er af så god kvalitet
{hængetræ_1} (Plant+Object): træ med nedhængende
{klatretræ_1} (Plant+Object): træ der er godt at klatre i,
{livstræ_1} (Plant+Object): træ som menes at give me
{skovtræ_1} (Plant+Object): træ el. træart der oftest vo
{vejtræ_1} (Plant+Object): træ der står i vejkanten, pla

Figure 4: Hyponymy in the DanNet editor

4 Extracting other semantic relations

Section 3 showed that the hyponymy relations are fairly straightforward to extract from the dictionary because each definition carries a genus expression in its *Genprox* element, denoting the nearest hyperonym. This section will deal with the complementary part of the definitions where the *differentia* information resides. The question to answer in this section is how much of the definition material can be semi-automatically exploited to establish semantic relations between synsets other than the hyponymy relation that we have already discussed – and moreover, which methods have to be applied to achieve this. Here, we will only sketch out some of the problems involved in such an attempt. A more detailed description can be found in Asmussen (2007).

When the DDO editors chose genuses for their definitions, they could stick to some general editorial guidelines to help them select appropriate genuses as they were obliged to fill in the *Genprox* element with an appropriate genus expression. This made the selection of a genus expression a conscious process. But as for the differentiating features there was practically no guidance at all: No specifications existed that stated which features or relations had to be specified for which types of words, and neither was the definition vocabulary nor the definition syntax preassigned in any way. These circumstances are not really problematic as long as the dictionary is a printed book addressing human users, but they seriously complicate computational exploitation of the material, for instance in a wordnet setting.

Therefore, at the current stage, the semantic relations other than hyponymy are encoded more or less on a manual basis. Some relations, however, are inherited from the hyperonym as will be examplified below. In all other cases, the DanNet encoders manually extract the relations that are given in the definition and transform them into wordnet relations – a procedure which is in most cases rather straightforward. As was the case for the hyponym relation, we rarely add new information to DanNet, i.e. information that is not already given in the DDO definition. On the contrary, we reduce information in cases where we find the definition too detailed, or where the content of the definition cannot easily be expressed via relations. As a default guideline, we strive towards expressing at least the most central features of the definition. Consider for illustration the definition below for the lemma *lagkage* 'layer cake':

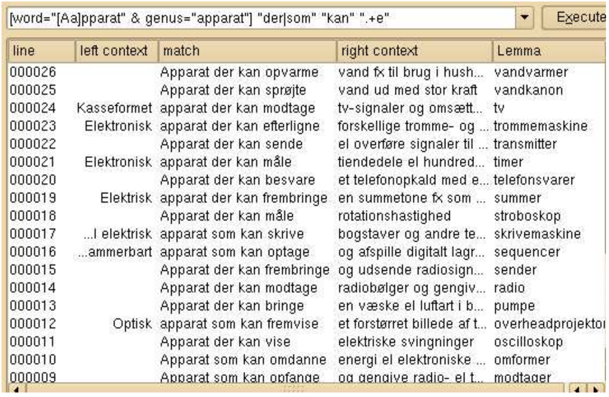
- *stor, cirkelrund kage bestående af 2-3 kagebunde der er lagt i lag med forskellig slags fyld, fx frugt, syltetøj og creme, og pyntet med glasur og flødeskum; skæres ud i trekantede stykker af form som cirkeludsnit* Lit.: 'big round cake consisting of 2-3 baked layers with different kinds of filling in between, such as fruit, jam, and custard and decorated with frosting and whipped cream; is cut out in triangular pieces shaped like circle sectors'

In this case, we manually encode the relations *has_mero_part: bund* 'layer' and *has_mero_part: creme* 'custard', whereas the agentive *made_by* relation is given in more general terms: *tilberede* 'prepare'. The rest of the content is omitted. In addition, the relations *has_mero_madeof: mel* 'flour', and *has_mero_madeof: sukker* 'sugar' are inherited from *kage* 'cake' and the relation *for_purpose_of: spise* 'eat' is inherited from *føde* 'food'.

A preliminary approximation to a potential more automatic exploitation of the dictionary material than the one presented in the *layer cake* example is to consider the whole of definitions as a special type of corpus amenable to common corpus-analytical investigation which may shed light on the structure of the definitions. The definition given in Figure 1 has the genus expression *apparat* 'technical device' whereas the modifier *kasseformet* 'box-shaped' and the VPs *modtage tv-signaler* 'receive tv signals' and *omsætte dem [...]* 'transform them [...]' must be differentia information. Based on these observations it can be hypothesised that for artefact definitions

1. premodifiers (i.e. adjectives) of the genus denote general (physical) properties of the definiendum;
2. VPs after a relative pronoun and headed by *kan* 'can' specify the telic role (i.e. the *for_purpose_of* relation) of the definiendum.

To find more definitions with these structural characteristics, the hypotheses can be reformulated as corpus queries: A rough approximation of the first hypothesis is to find all definitions in the corpus with genus expression *apparat* with exactly one word – an assumed premodifying adjective – immediately to the left of it. A quick search through the definition corpus reveils that groups of premodifiers are quite common as well. The corpus query should therefore cover these cases also. The total inventory of premodifying adjectives used in conjunction with the genus expression *apparat* is (frequencies in brackets): *elektrisk* (23) 'electric', *elektronisk* (16) 'electronic', *optisk* (5) 'optical', *mekanisk* (4) 'mechanic', *lille* (4) 'small', *kasseformet* (3) 'box-shaped', *transportabelt* (2) 'portable', *ballonbåret* (1) 'balloon-carried', *computerbaseret* (1) 'computerbased',



```
[word="[Aa]pparat" & genus="apparat"] "der|som" "kan" ".+e"
```

line	left context	match	right context	Lemma
000026		Apparat der kan opvarme	vand fx til brug i hush...	vandvarmer
000025		Apparat der kan sprøjte	vand ud med stor kraft	vandkanon
000024	Kasseformet	apparat der kan modtage	tv-signaler og omsætt...	tv
000023	Elektronisk	apparat der kan efterligne	forskellige tromme- og ...	trommemaskine
000022		Apparat der kan sende	el overføre signaler til ...	transmitter
000021	Elektronisk	apparat der kan måle	tiendedele el hundred...	timer
000020		Apparat der kan besvare	et telefonopkald med e...	telefonsvarer
000019	Elektrisk	apparat der kan frembringe	en summetone fx som ...	summer
000018		Apparat der kan måle	rotationshastighed	stroboskop
000017	...l elektrisk	apparat som kan skrive	bogstaver og andre te...	skrivemaskine
000016	...ammerbart	apparat som kan optage	og afspille digitalt lagr...	sequencer
000015		Apparat der kan frembringe	og udsende radiosign...	sender
000014		Apparat der kan modtage	radiobølger og gengiv...	radio
000013		Apparat der kan bringe	en væske el luftart i b...	pumpe
000012	Optisk	apparat som kan fremvise	et forstørret billede af t...	overheadprojektor
000011		Apparat der kan vise	elektriske svingninger	oscilloskop
000010		Apparat som kan omdanne	energi el elektroniske ...	omformer
000009		Apparat som kan opfange	og gengive radio- el t...	modtager

Figure 5: Telic role in DDO definitions

programmerbart (1) 'programmable', *fladt* (1) 'flat', *mindre* (1) 'smaller', *teknisk* (1) 'technical', *tryktluftdrevet* (1) 'powered by air compression', *stort* (1) 'large', *rørformet* (1) 'tube-shaped' – a total of 16 different premodifiers occurring (partly grouped together) in 57 *apparat* definitions. The total of *apparat* definitions in the dictionary is 203, and it seems unlikely that none of the 16 listed modifiers should not be relevant to any of the remaining 152 definitions. A closer look at the modifiers furthermore reveils some peculiar cases: A *computer monitor* can either be *box-shaped* or *flat*, wheras a *tv set* only can be *box-shaped*; an *oven* is *technical* but nothing is mentioned about its shape; a *hearing aid* is *small* (but not *electronic*) whereas a *pacemaker* is *electronic* (but not *small*). The examples show that specification of physical attributes in the *apparat* definitions are quite scattered and inconsistent. For a printed dictionary for humans this is hardly a big problem since the users will know how to interpret the information anyway, but it makes algorithmic exploitation of the material almost impossible.

Figure 5 shows hypothesis 2 converted into a corpus query[1] and part of the resulting concordance together with the corresponding lemmas.

A total of 19 different verbs[2] are covered by this query, of which the most frequent ones are *frembringe* (4) 'generate', *modtage* (3) 'receive', and *måle* (3) 'measure', whereas each of the remaining verbs occurs only once. The total number of concordance lines is 26 which clearly shows that hypothesis 2 only covers a tiny part of the 203 *apparat* definitions. This may have two reasons: The first one is that the definition is not captured by the query because its pattern is extended. This is e.g. the case in the definition for the synonym pair *fjernskriver/telex* 'teleprinter' – the interposed part is surrounded by square brackets:

- *elektrisk apparat som [ligner en skrivemaskine og som tilsluttet et særligt netværk] kan sende og modtage skriftlige meddelelser*
  Lit.: 'electric device [resembling a typewriter which connected to a special network] can send and receive written messages'

The second reason is that the *for_purpose_of* relation may be incorporated in other structural patterns than the one given in hypothesis 2. A quick browse through some *apparat* defintions shows that there is a variety of possibilities:

1. **Pattern:** genus expression *der/som bruges til at* VP-inf *med*
   Lit.: genus expression *that is used for to* VP-inf *with*
   Example: *apparat som bruges til at spinde garn med*
   Lit.: 'device that is used for to yarn thread with'
   Occurrences: 3

2. **Pattern:** genus expression *til at* VP-inf *med/på/i*
   Lit.: genus expression *for to* VP-inf *with/on/in*
   Example: *apparat til at afspille cd'er med*
   Lit.: 'device for to play-back CDs with'
   Occurrences: 11

3. **Pattern:** genus expression *der/som* VP-fin
   Lit.: genus expression *that* VP-fin
   Example: *apparat der måler og viser et køretøjs hastighed*
   Lit.: 'device that measures and displays the speed of a vehicle'
   Occurrences: 42

[1] The corpus query system used is the IMS Corpus Workbench (Christ, 1994) embedded in a Python-based server-client application and accessed through a Qt-based graphical user interface, both developed at DSL.

[2] The verb proper very often is semantically too underspecified to yield a meaningful *for_pupose_of* relation. For this reason, DanNet has introduced an *fpo_object* relation as well.

4. **Pattern:** genus expression *til* NP
   Lit.: genus expression *for* NP
   Example: *apparat til optagelse og afspilning af lyd*
   Lit.: 'device for the recording and play-back of sound'
   Occurrences: 29

5. **Pattern:** genus expression *der/som er specielt beregnet til at* VP-inf
   Lit.: genus expression *that is specially designed for to* VP-inf
   Example: *apparat som er specielt beregnet til at optage og afspille tale*
   Lit.: 'device that is specially designed for to record and play-back *speech*'
   Occurrences: 1

Patterns 1–5 cover 86 definitions. Together with the pattern from hypothesis 2, 70% of the *apparat* definitions are covered by six patterns. Once these patterns have been established, it gets more feasible to automatically extract the semantic information necessary to determine the *for_purpose_of* relation. But still, 30% of the definitions can probably not be processed automatically at all, as the variety of different syntactic ways to indicate semantic relations in definitions cannot be covered by a few algorithmic rules. And the process of formulating these rules is in itself rather "manual" and time-consuming. Furthermore, extraction with high precision would require a syntactically annotated definition corpus.

If dictionary definitions really are to be exploited automatically they are required to be constructed in a more predictable way with an explicitly defined syntax where certain syntactic patterns correspond to certain semantic relations.

Another considerably more coarse way to isolate differentia expressions which may concern the telic role is to use a statistical approach where a frequency list of tokens in definitions with the genus expression *apparat* is compared to a frequency list of tokens in the definitions corpus as a whole. Salient tokens from the *apparat* corpus can be determined by some statistical test such as log likelihood (Dunning, 1994) or mutual information (Church and Hanks, 1989). By applying a modified version of mutual information we get the following salient tokens in the *apparat* corpus that possibly may give some hints of the telic role in question:

- ***afspille* 'to play-back':** *grammofon*, *cd-afspiller* 'CD player', *afspiller* 'player', *sequencer*, *diktafon*
- ***afspilning* 'play-back':** *kassettespiller* 'cassette recorder', *hjemmevideo* 'video cassette recorder', *kassettebåndoptager* 'cassette recorder', *båndoptager* 'tape recorder'
- ***måle* 'measure':** *stroboskop*, *måler* 'measuring tool', *timer*, *løgnedetektor* 'lie detector', *ekkolod* 'sonar'
- ***måler* 'measuring tool':** *gasmåler* 'gas meter', *speedometer* 'speed indicator', *omdrejningstæller* 'evolutions meter', *benzinmåler* 'fuel gauge', *fotofælde* 'speed camera'
- ***måling* 'gauging':** *elmåler* 'electric meter', *trykmåler* 'pressure gauge', *luxmeter*, *spirometer* 'aeroplethysmograph', *gyrometer*, *alkometer*, *newtonmeter*, *magnetometer*, *instrument*, *kalorimeter*
- ***målinger* 'measurements':** *måleinstrument* 'measurement device', *radiosonde*, *satellit*, *fartskriver* 'tachograph'

By taking this type of automatically generated lists showing salient tokens in definitions with the genus expression *apparat* together with the according lemmas the wordnet editor may get hints on which synsets should be supplied with the same telic role info.

The examples show that it is possible to use some approaches from corpus linguistics to get a first impression of the structure of dictionary definitions, but the interpretation of the correlation between elements in the differentia part of the definition and their apropiate semantic function can only be performed by the DanNet editor. A fully automated transformation of dictionary definitions to a wordnet seems hardly possible although certain corpus-analytical methods may show a useful tool in some cases. Thus, to determine the *for_purpose_of* relation, the established patterns could be used to extract verbs from the definitions that express this relation and these verbs could be presented for the DanNet editor as possible descriptors of the *for_purpose_of* relation among which the editor then could choose the appropriate ones.

5 Conclusions

As has been demonstrated, the "manual" exploitation of the dictionary is quite straightforward as the DanNet editors can reuse quite a lot of the semantic information given, especially synonym/antonym information, cf. Section 2, and hyponymy information, cf. Section 3. The DanNet editing tool supports this type of information quite well and thus substantially facilitates the coding process. However, due to the free "human" style of formulating definitions in the dictionaries, it proves to be much more difficult to exploit them automatically and transform them into semantic relations, cf. Section 4.

A widely discussed topic in the wordnet environment is the choice between the "expansion approach" and the "merging approach" when building wordnets. It is generally accepted that the former approach is easier, cheaper and secures better consistency between wordnets, but involves a genuine risk of linguistic bias, whereas the latter presents a more loyal picture of linguistic conceptualisation in a specific language, but may for the same reason be less compatible with other wordnet structures, and in addition, this strategy is more labour intensive and thus correspondingly resource-demanding unless it can be based on already existing lexical-semantic resources as a comprehensive monolingual dictionary. The DanNet project has chosen to follow the second option by creating a wordnet on the base of a dictionary and only ensuring compatibility via the common core of wordnet base concepts. From a project management point of view it goes without saying that the advantage of such an approach can hardly be overrated. In addition, however, we are convinced that the result reflects the sense relations of the Danish language better than it would have if we had built DanNet independently.

As a consequence of the chosen approach, DanNet may be even further away from an idealised tidy and homogenous taxonomy than other wordnets are. In order to converge to the requirements of formal ontology and terminologists, the "orthogonality" feature on hyponymy has been introduced in DanNet denoting whether or not the hyponymic relation is taxonomic. This means that one can chose to select only the taxonomical part of the vocabulary for specific ontology-related purposes and thereby omit more heterogeneous and conflicting parts of the vocabulary.

Another consequence of using DDO as the main source is that DanNet implicitly takes the position of the layman. This means that the hyponymic structures exposed are sometimes deep, sometimes disturbingly flat, and that several logical gaps can be found in the system. This is probably the most important lesson one can learn from working with "real language" data deduced from corpora and compiled in a dictionary. However, we are inclined to believe that this is an indispensable fact about human language that deserves to be accounted for in the wordnet.

References

Jörg Asmussen. 2007. Korpuslinguistische Verfahren zur Optimierung lexikalisch-semantischer Beschreibungen. In Werner Kallmeyer, editor, *Jahrbuch des Instituts für Deutsche Sprache 2006*. de Gruyter, Berlin / New York.

Oliver Christ. 1994. A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX'94*, Budapest.

Kenneth Church and Patrick Hanks. 1989. Word association norms, mutual information and lexicography. In *ACL Proceedings, 27th Annual Meeting*, Vancouver.

D. Alan Cruse. 1991. *Lexical Semantics*. Cambridge University Press, Cambridge.

D. Alan Cruse. 2002. Hyponymy and Its Varieties. In R. Green, C.A. Bean, and S.H. Myaeng, editors, *The Semantics of Relationships: An Interdisciplinary Perspective, Information Science and Knowledge Management*. Springer-Verlag.

Ted Dunning. 1994. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, (19):61–74.

Ebba Hjorth, Kjeld Kristensen, Henrik Lorentzen, Lars Trap-Jensen, Jørg Asmussen, et al., editors. 2005. *Den Danske Ordbog 1–6*. DSL & Gyldendal, København/Copenhagen.

Henrik Lorentzen. 2004. The Danish Dictionary at large: presentation, problems and perspectives. In *Proceedings of the 11th EURALEX International Congress*, volume 1, pages 285–294, Lorient. Euralex.

Ole Norling-Christensen and Jørg Asmussen. 1998. The Corpus of The Danish Dictionary. *Lexikos. Afrilex Series*, 8:223–242.

Bolette Sandford Pedersen and Patrizia Paggio. 2004. The Danish SIMPLE Lexicon and its Application in Content-based Querying. *Nordic Journal of Linguistics*, 27(1):97–127.

Bolette Sandford Pedersen and Nicolai Hartvig Sørensen. 2006. Towards sounder taxonomies in wordnets. In *Proceedings from the OntoLex Workshop in association with LREC 2006*, Genova.