SIMPLE - Semantic Information for Multifunctional Plurilingual Lexica: Some Examples of Danish Concrete Nouns

Bolette Sandford Pedersen
Center for Sprogteknologi
Njalsgade 80
DK-2300 S Denmark
bolette@cst ku dk

Britt Keson
Det Danske Sprog- og Litteraturselskab
Christians Brygge 1, 1
DK-1219 Copenhagen K Denmark
britt_keson@irct org

Abstract

SIMPLE is a large-scale European lexicon project funded by the European Commission with the participation of 12 European countries The aim of the project is to add harmonized semantic information to the LE-PAROLE lexicons[1], which contain morphological and syntactic information In this paper we present some examples of concrete nouns from the Danish SIMPLE lexicon which illustrate two central aspects of the SIMPLE model 1) the expressive power of the Qualia Structure exemplified with a phenomenon relevant to a Scandinavian language like Danish namely the representation of the internal structure of Danish non-deverbal nominal compounds, and 2) the representation of regular polysemy in the Danish SIMPLE lexicon

1 Introduction

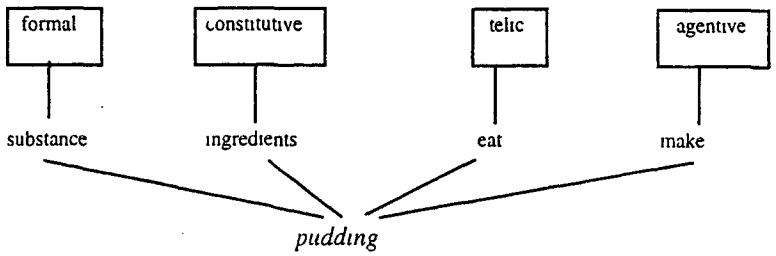
The SIMPLE model is primarily based on three lexical frameworks (Lenci *et al*, 1998) The Generative Lexicon (cf Pustejovsky, 1995), WordNet (cf Miller and Fellbaum, 1991), and EuroWordNet (cf Vossen *et al*, 1998) The basic underlying assumption in the model is that word senses differ in terms of their internal complexity Hence the SIMPLE model consists of three different semantic types (*i*) simple types, which can be characterized in terms of monodimensional relations, (ii) unified types, which involve multidimensional information, and (iii) complex types, which identify regular polysemous classes

One of the basic tasks during the SIMPLE lexicon encoding phase is the assignment of semantic typing to the word senses to be encoded (called the semantic units or SemU's) A set of schematic structures called 'templates' constituting the SIMPLE Ontology (consisting of approx 140 semantic types in all) guides this encoding process A template is a cohort of various different information types which is primarily used by the lexicon encoder to express the semantic type of a word sense, but also to express its domain, definition, predicative representation, argument structure, polysemous classes, etc

The multiple dimensions of meaning are represented in SIMPLE by the use of the Qualia Structure from the Generative Lexicon (Pustejovsky 1995) to represent lexical meaning expressed by means of orthogonal inheritance The Qualia Structure involves four different roles (i) the formal role, which provides information that distinguishes an entity within a larger set, (ii) the agentive role, which concerns the origin of an entity, (iii) the telic role, which concerns the typical function of an entity, and (iv) the constitutive role, which expresses a variety of relations concerning the internal constitution of an entity As an illustration, consider in Figure 1 the meaning components involved in the noun *pudding*

[1] The LE-PAROLE lexicons contain 20,000 entries with corresponding morphological and syntactic information for each of the 12 languages that participated in this project, which was also funded by the European Commission (cf Ruimy *et al*, 1998)

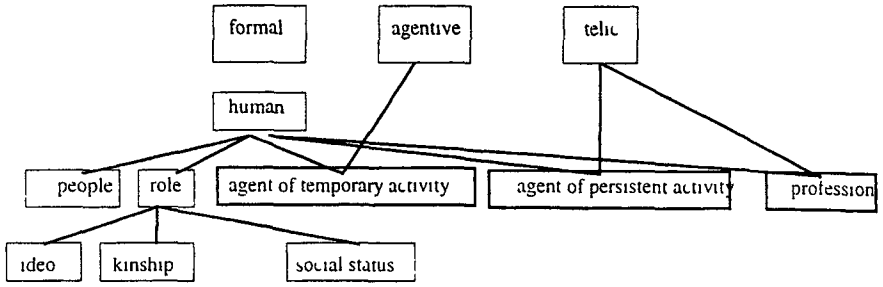
Figure 1: The meaning components of *pudding* (Lenci et al 1998 pp 17)



The central meaning aspects are mirrored in the linguistic contexts surrounding the word, so for *pudding* we could have *John refused the pudding* referring to the eat event, *that's an easy pudding* referring to the make event, *there is pudding on the floor* referring to the substance dimension, and *that was a nice bread pudding* referring to the ingredients of which it is made

As an example of the semantic types expressed in the SIMPLE Ontology and of how the different dimensions of meaning are involved for each semantic type, consider Figure 2 which shows a subset of the SIMPLE ontology referring to human beings

Figure 2: Subset of the SIMPLE Core Ontology representing human beings



Some examples of word senses encoded as simple types are *russer* (a Russian) under the template type 'people', *jøde* (Jew) under 'ideo', *kusine* (female cousin) under 'kinship, and *ven* (friend) under 'social status' These senses may naturally involve additional dimensions of meaning, however they are not considered type-defining in the SIMPLE model In contrast, word senses encoded under the emphasised boxes above, such as *borddame* (female dinner partner) under 'agent of temporary activity', *alkoholiker* (an alcoholic) under 'agent of persistent activity', and *læge* (doctor) under 'profession', are unified types These are identified by more than one coordinate in the type hierarchy, since they involve more than one type-defining dimension of meaning Thus 'agent of temporary activity' also involves an agentive role For *borddame* this is defined by the act of sitting next to someone at a dinner 'Agent of persistent activity' and 'profession' involve a telic role, which for *alkoholiker* is the act of drinking, and for *læge* is the act of curing

2 Danish nominal compounds denoting containers: expressing their internal structure via the Qualia Structure

The internal semantic structure of Danish deverbal nominal compounds (i e *flisebelægning* (flagging, lit 'flagstone paving'))

can to some extent be identified by the argument structure of the derived verb and the internal ranking of its arguments (cf Ørsnes, 1995) In contrast, non-deverbal nominal compounds generally display a more arbitrary internal structure in Danish (cf Paggio & Ørsnes 1993), and hence they require a higher degree of expressive power in the lexical entries The Qualia Structure as it is expressed in the SIMPLE model provides a good basis for a lexicalised encoding of Danish non-deverbal nominal compounds, as for example nominal compounds denoting containers

The template type 'Container' belongs to the set of templates constituting the SIMPLE Core Ontology It is a unified type that has the unification path 'Concrete entity + Artifact/Agentive + Telic' This indicates that the template denotes a kind of concrete entity, and that it has been augmented with two kinds of additional type-defining information (i) agentive information (namely that these concrete entities are man-made artifacts), and (ii) telic information (namely that these concrete entities are used for a specific purpose to contain things)

All Danish containers encoded under this template type have been encoded with Danish information about the formal role via an is_a-hierarchy As a default, *beholder* (container) is chosen as hypernym for the Danish containers in the is_a-hierarchy Since containers are (man-made) artifacts, the process of their creation is specified via the agentive role For Danish containers this is the process *fremstille* (to create) As is apparent from the unification path for this template, containers are also encoded with the type-defining telic information that their function is to contain things For Danish containers this is specified with the verb *indeholde* (to contain) in the encoding of the telic role

The Qualia Structure used in the SIMPLE model is well-suited for capturing a majority of the internal relationships for Danish non-deverbal nominal compounds denoting containers As an example, the encoding of the meaning of a word like *bæger* (cup) can be further augmented with the telic role for such compounds as *målebæger* (measuring cup), *raflebæger* (lit cup for casting dice = dice cup), or *drikkebæger* (drinking cup) In the SIMPLE model, the encoding of the meaning of *bæger* can also be further specified by including information about the constitutive role for other types of compounds The encodings can include additional information on (i) the material of which the container is made, e g *plasticbæger* (plastic cup), *messingbæger* (brass cup), or *papbæger* (paper cup), or (ii) what the container (prototypically) contains, e g *askebæger* (ashtray), *giftbæger* (poisoned chalice), or *yoghurtbæger* (yoghurt cup)

The following SGML-encoded examples from the Danish SIMPLE lexicon[2] demonstrate the use of the SIMPLE Qualia Structure to distinguish between different types of compound containers The first example shows the encoding of *vinflaske* (wine bottle) This is shown to be a hyponym of *flaske* (bottle) in the formal is_a-relationship The encoding of *vinflaske* differs from *flaske* by the additional constitutive information expressing that it (prototypically) contains *vin* (wine) In the second example, the meaning of *bærepose* (carrying bag) is linked to its hypernym *pose* (bag) in the is_a-hierarchy The encoding of *bærepose* also contains the additional telic information that its prototypical function is to carry something (*bære*) In the last example, the meaning of *blikdåse* (tin can) is shown to be a hyponym of *dåse* (can) in the is_a-hierarchy The encoding of *blikdåse* also contains the additional constitutive information that it is made of *blik* (tin), whereas *dåse* would be underspecified with respect to this role

[2] The SGML-encodings displayed here are slightly simplified versions of the actual Danish lexicon encodings The example sentences were taken from a composition of several Danish corpora (45 mill running words in all) The definitions marked '(NDO)' were taken from the CD-ROM edition of Politikens Store Nye Nudansk Ordbog (1997)

```
<SemU   id="USEM_N_vinflaske_CON_1"
        naming="vinflaske"  /wine bottle/
        example="en vinflaske kan genbruges syv til otte gange"  /a wine bottle can be reused seven-eight times/
        freedefinition="flaske til vin"    /a bottle for wine/
        weightvalsemfeature1="
                WVSFTemplateContainerPROT
                WVSFUnificationPathConcreteentity-ArtifactAgentive-TelicPROT">
        <RWeightValSemU
                target="USEM_N_flaske_CON_1" /bottle/
                semr="SRIsa">
        <RWeightValSemU
                target="USEM_V_fremstille_1" /to produce/
                semr="SRCreatedby">
        <RWeightValSemU
                target="USEM_V_indeholde_1"  /to contain/
                semr="SRUsedfor">
        <RWeightValSemU
                target="USEM_N_vin_ARD_1"  /wine/
                semr="SRContains"> </SemU>
```

Additional constitutive information ('Contains')

```
<SemU   id="USEM_N_bærepose_CON_1"
        naming="bærepose"  /carrying bag/
        example=' hav mindst en hel bærepose fuld parat til at sætte pa bordet"
                /keep at least a whole carrying bag full ready to put on the table/
        freedefinition="en pose af papir, plastic el stof med hanke som bruges til at bære fx købmandsvarer i (NDO)"
                /a bag with handles made of paper plastic or cloth which is used to carry e g groceries (NDO)/
        weightvalsemfeature1="
                WVSFTemplateContainerPROT
                WVSFUnificationPathConcreteentity-ArtifactAgentive-TelicPROT">
        <RWeightValSemU
                target="USEM_N_pose_CON_1" /bag/
                semr="SRIsa">
        <RWeightValSemU
                target="USEM_V_fremstille_1" /to produce/
                semr="SRCreatedby">
        <RWeightValSemU
                target="USEM_V_indeholde_1" /to contain/
        semr="SRUsedfor">
        <RWeightValSemU
                target="USEM_V_bære_1"  /to carry/
                semr="SRUsedfor '> </SemU>
```

Additional telic information ('Usedfor')

```
<SemU   id='USEM_N_blikdåse_CON_1"
        naming="blikdåse"   /tin can/
        example="en urtepotteunderskal, hvori man omvendt har sat en tom blikdase  fyldes med vand'
                /a flower pot saucer in which one has placed an empty tin can upside down is then filled with water/
        freedefinition="dåse lavet af blik" /can made of tin/
        weightvalsemfeature1="
                WVSFTemplateContainerPROT
                WVSFUnificationPathConcreteentity-ArtifactAgentive-TelicPROT '>
        <RWeightValSemU
                target="USEM_N_dåse_CON_1" /can/
                semr= 'SRIsa">
        <RWeightValSemU
                target= 'USEM_V_fremstille_1' /to produce/
                semr= 'SRCreatedby">
        <RWeightValSemU
                target="USEM_V_indeholde_1" /to contain/
                semr="SRUsedfor">
        <RWeightValSemU
                target="USEM_N_blik_ARS_1" /tin/
                semr="SRMadeof"> </SemU>
```

Additional constitutive information ('Madeof')

3 Regular polysemy in the Danish lexicon

Regular polysemy - when groups of related words display the same ambiguity - is handled in a uniform way in the SIMPLE model via the identification of a set of well-established regular semantic classes, which are adjusted for each of the languages involved While unsystematic ambiguous readings of a word are represented as totally unrelated semantic units, regular polysemous senses can be encoded as interlinked semantic units In the SIMPLE model this is represented by an information slot called *complex*, whose value is the polysemous class to which the semantic unit belongs This strategy relates to Pustejovsky (1995), where regular polysemous classes correspond to complex types, which allows for an underspecified semantic typing of word senses The solution adopted in SIMPLE intends to be a first step towards the future development of underspecified semantic types (Lenci *et al*, 1998)

Empirically-based studies of regular polysemous semantic classes of Danish are at present very scarce (see however Boje & Schøsler (ed) (1992) pp 11-12 and Braasch & Pedersen (forthcoming) for some minor considerations of regular polysemy in Danish nouns as well as Malmgren (1988) for an extensive study of regular polysemy in Swedish, a language which displays polysemous behaviour very similar to Danish) This fact underpins the need for corpus-oriented encoding procedures which have therefore been given a central focus in the the Danish SIMPLE dictionary in the sense that each semantic encoding is supported by corpus examinations[3]

In the Danish lexicon the most productive cases of regular polysemy involving concrete nouns prove to be the following

- animal / food
- geographical location / human group
- fruit / plant
- human group / institution
- semiotic artifact / information

Other well-known polysemous pairs are not productive in Danish, as for example 'people / language' and 'flower / colour', where only a few examples of each can be found This difference relates to the distinction made by Apresjan (apud Malmgren, 1988) between productive and regular polysemy Here productive polysemy refers to cases where more or less the whole group of nouns within a semantic class display the same polysemy relations, whereas regular polysemy refers to cases where at least two words - but not the whole class - follow the same polysemy pattern

Below is shown an example of the semantic encoding of a proper noun denoting a Danish city, which belongs to the productive 'geographical location / human group' polysemy This example shows the encoding of the 'human group' sense of the word This can be seen in the corpus example ('example'), the definition ('freedefinition'), and the kinds of qualia roles encoded

We believe that the corpus-oriented approach used during the encoding of the Danish SIMPLE lexicon facilitates the identification of new polysemous classes, since the differences in distributional patterns of the encoded words senses are a good indication of whether a regular polysemy relation could be involved

[3] We apply the corpus tool Xkwic on a composite corpus of 44 mill running words (consisting of 'Berlingske Korpus', 'Bergenholtz Korpus' and 'Parole Korpus')

```
<SemU   id="USEM_N_Dragør_HUG_1"
        naming="Dragør" /Dragør - Danish city/
        example="Dragør må i år af med godt 31 mill kr til den kommunale udligning"
                /This year Dragør must pay approx 31 mill crowns to the community equalization/
        freedefinition="de mennesker der bor i Dragør" /The people living in Dragør/
        weightvalsemfeature1="
                WVSFTemplateHumanGroup
                WVSFTemplateSuperTypeGroupPROT
        <RWeightValSemU
                target="USEM_N_befolkning_HUG_1" /population/
                semr="SRIsa">
        <RWeightValSemU
                target="USEM_N_indbygger_HUM_1" /citizen/
                semr="SRHasasmember">
                <RWeightValSemU
                target="USEM_N_Dragør_GEO_1"
                semr="SRPolysemyHumanGroup-GeopoliticalLocation"></SemU>
```

Regular polysemy relation

4 Conclusion

Danish is a typical Scandinavian language with respect to nominal compounding and patterns of regular polysemy In this paper we have demonstrated how a large-scale, plurilingual, and multifunctional lexicon project like SIMPLE facilitates a flexible semantic encoding that can capture universal semantic principles as well as these language-specific characteristics Considering the current status of language technology for a 'small' European language such as Danish, the scope of the SIMPLE project makes it a truly pioneering project The development of this harmonized large-scale semantic lexicon for 12 European languages will enable system developers to implement sophisticated language technology that will also encompass small European languages in the future

References

Politikens Store Nudansk Ordbog på cd-rom, Version 2 1 1997 Politikens Forlag, København

Boje, F & L Schøsler (ed) (1992) 'DISEM - A Semantic MT-Component' in *CST Working Papers no 1, Center for Sprogteknologi*, Copenhagen

Braasch, A & B Pedersen (forthcoming) 'En stor sprogteknologisk ordbog for dansk - med særligt fokus på håndtering af flertydighed i en niveaudelt ordbog', in *7 Møde om Udforskning af Dansk Sprog*, Århus University

Lenci, A , F Busa, N Ruimy, E Gola, M Monachini, N Calzolari, A Zampolli, El Guimier, G Recourcé L Humphreys, U Von Rekovsky, A Ogonowski, C McCauley, W Peters, I Peters, M Villegas (1998) 'Specifications', SIMPLE Work, *Linguistic Deliverable D2 1*, Pisa

Malmgren, S (1988) 'On Regular Polysemy in Swedish', in *Studies in Computer-Aided Lexicography*, Almquist & Wiksell, Stockholm

Ruimy, N O Corazzari, E Gola, A Spanu, N Calzolari, A Zampolli (1998) 'The European LE-PAROLE Project The Italian Syntactic Lexicon', in *First International Conference on Language Resources & Evaluation*, Granada, Spain

Paggio, P & B Ørsnes (1993) Automatic translation of nominal compounds A case study of Danish and Italian' in *Revista di Linguistica 5 1* pp 129-156, Rosenberg & Sellier, Torino

Pustejovsky, J (1995) *The Generative Lexicon*, Cambridge, MA, The MIT Press

Vossen, P , L Bloksma, H Rodrigues, S Climent, A Roventini, F Bretagna, A ALonge, W Peters (1998) 'The EuroWordNet Base Concepts and Top Ontology', *Deliverable D017, D034,D036, WP5, LE2-4003*

Ørsnes, B (1995) *The Derivation and Compounding of Complex Event Nominals in Modern Danish* PhD Dissertation, University of Copenhagen