Some Aspects of the Syntactic Encoding of Nouns in a Computational Lexicon – the STO Project

Sussi Olsen
Center for Sprogteknologi
Njalsgade 80
DK-2300 Copenhagen S
Denmark
sussi@cst.dk

Abstract

In this paper I first give a brief introduction to the syntactic representation of the STO lexicon. Hereafter I focus on some of the problematic phenomena that we encounter during the syntactic encoding of nouns. In principle each noun with one subcategorisation frame is encoded as one syntactic unit even if the noun has two or more semantic readings. Two different syntactic manifestations of the same semantic reading of a noun should however be encoded in the same syntactic unit. When the subcategorisation frame is established, the limits between arguments and modifiers can be difficult to distinguish. Normally modifiers are not considered part of the subcategorisation frame but if corpus search reveals some frequent modifiers, these can be encoded in line with the arguments. Noun compounding is a productive word formation process in Danish. The first element of a compound can fill in the argument slot of the second element, changing the subcategorisation frame, but this is not always the case. Genitive is only encoded when it is part of the subcategorisation frame. For non-deverbal nouns it can be difficult to decide when to encode a genitive. The choice is highly dependent on the number of occurrences in corpora.

The Linguistic Architecture of STO

The Danish STO project, SprogTeknologisk Ordbase, (i.e. Lexical database for language technology) see Braasch et al. [1998] is a national follow-up project of the Danish PAROLE lexicon, see LE-PAROLE [1998] with the aim of creating a large size Danish lexicon for natural language processing.[1]

The background for the architecture of STO is as mentioned above the PAROLE lexicon but parts of it have of course been developed concurrently with the progress of the project. Since the purpose of the STO project is to develop a computational lexicon that can be used for many different NLP applications, we must be aware that we are developing a generic lexical database from which different application specific lexicons can be extracted. This means that the STO database should not be dependent on a specific linguistic theory but rather present detailed linguistic information in an explicit way as independent of theories as possible.

The linguistic information of the lexicon is organised in three independent but coherently linked layers, i.e. the morphological, the syntactic and the semantic layer. Each layer is made up of a system of the respective linguistic properties organised in sets of information, which are called 'units'. Each unit represents the linguistic behaviour of a lemma at that layer, and the complete description of a lemma includes the morphological, the syntactic and the

semantic units related to this lemma. The figure below might facilitate the comprehension of the lexicon architecture.

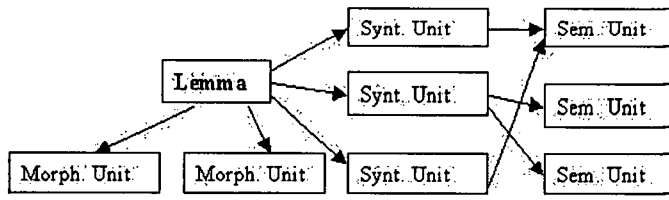

Fig 1: The STO description model

The linguistic information of STO is primarily based on corpus investigations. What is described of a given lemma at the three levels, is what has been found in the corpora we have access to. This means that even some examples of 'bad' language use or errors in spelling and inflection - according to the standardised rules of Danish established by the Danish Language Council – will be included if these phenomena are frequent in the corpus evidence. This kind of information is supplied with an annotation telling that this information is not approved by the Danish Language Council and should not be used for language generation. Of course the size and bias of the corpora might skew the information gathered but we supply the corpus investigations with searches on the web, and in cases where the number of occurrences of a specific phenomenon is too small to rely on, the lexicographer's introspection can overrule the corpus evidence.

See Braasch et al. [2002] for further information of the linguistic specifications of the STO lexicon.

The Syntactic Units for Nouns

The syntactic unit describes one syntactic behaviour of one morphological unit. This behaviour is described in one or more descriptions.

The syntactic units for nouns in the STO lexicon contain information about the subcategorisation frame, an example and possible comments. The subcategorisation frame is described with a mnemonic sequence of letters and digits, facilitating the encoder's work, e.g.

Dn2GPn-med

This lexical description for nouns starts with a 'D' followed by an 'n' for noun. After that comes a digit which tells whether the noun is avalent (0)[2], monovalent (1), divalent (2) and so forth. When a genitive is subcategorised for, it is expressed with a 'G'. Potential prepositional phrases are described with a P followed by the kind of governor it may take (n for NP, t for that-clause, i for infinitive, w for wh-clause, q for interrogative clause) and by the preposition itself.

This way to encode the arguments is merely an encoding convention, which is used to point to the relevant records in the tables of the database. When the syntactic information is extracted, it can be expanded into other formats.

The background
The principles for the encoding of the subcategorisation frames do not respond to a specific theory either. The analysis of the argument structure for nouns by Jane Grimshaw [1990] has been an inspiration for the Danish PAROLE lexicon and consequently for the STO lexicon. The LINDA manual [Underwood 1996] that presented a set of linguistic specifications for Danish in a Typed Feature Structure based formalism also played a role in establishing the theoretical base. But during the process of encoding thousands of nouns, trying to account for many unforeseen problems, a lot of pragmatic decisions have been taken, implying that the original theoretic decisions have been overruled in some cases. The clearest example of this is the fact that what in this article is called the subcategorisation frames or valency patterns (contained in the syntactic description) do not only comprise the arguments of the nouns but also some frequent modifiers. How this is done, will be explained thoroughly in a later section of this paper, but the reason to do this is that corpus investigations have revealed that some modifiers are so frequent and so central to some nouns that the best solution for a computational lexicon is to include them in a syntactic description so that e.g. a parser can recognise them afterwards.

One or more Syntactic Units

As mentioned above a syntactic unit accounts for the syntactic behaviour of one morphological unit and is expressed in one or more syntactic descriptions. But the encoding of nouns in syntactic units gives rise to some questions concerning when a noun with various syntactic realisations should be encoded in one or several syntactic units. The principles presented below are the ones followed in the STO lexicon but have of course been the subject for many discussions.

A noun which can occur with two different subcategorisation frames should as the main principle be encoded as two independent syntactic units, each of these having its own valency description, e.g.

SynU 1	Lemma	Description	Example
	variation	Dn1Pn-af	variation af temaet (*variation of the theme*)

SynU 2	Lemma	Description	Example
	variation	Dn1Pn-over	variation over temaet (*variation over the theme*)

Table 1: The syntactic units for 'variation'

There is little semantic difference between the two occurrences of 'variation'.

Constructions which we consider alternations, i.e. a noun which can have different syntactic realisations of the same semantic reading, constitute an exception to this principle. Such alternations are encoded in the same syntactic unit, each with its own valency descriptions in order to be able to recognise such constructions as alternations. This means that a noun like 'forpagter' (*tenant)* which can realise the same element either as a subjective genitive or as a prepositional phrase, is encoded in one syntactic unit with two syntactic descriptions:

	Lemma	**Description**	**Example**
SynU 1	*forpagter*	Dn1Pn-af	Forpagteren af gården (*the tenant of the farm)*
		Dn1G	Gårdens forpagter (*lit. the farm's tenant*)

Table 2: The syntactic unit for 'forpagter'

This is also the case with reciprocal constructions like the word 'forlovelse' (*engagement*) which is encoded with three valency descriptions in the same syntactic unit:

	Lemma	**Description**	**Example**
SynU 1	*forlovelse*	Dn1G	Peter og Susannes forlovelse (*Peter and Susanne's engagement*)
		Dn1Pn-mellem	forlovelsen mellem Peter og Susanne (*the engagement between Peter and Susanne)*
		Dn2GPn-med	Peters forlovelse med Susanne (*Peter's engagement with Susanne)*

Table 3: The syntactic unit for 'forlovelse'

Two semantic readings of the same noun, which do not differ in their syntactic realisation, are encoded as one syntactic unit. This means that the Danish word 'krampe' which means either *convulsion* or *staple*, is encoded in one syntactic unit as an avalent noun (Dn0), and that the semantic difference will be registered at the semantic level of the lexicon.

Modifiers and Arguments

When the subcategorisation frame of a noun has to be established, it may some times be difficult to distinguish the arguments of the noun from a modifier. Normally, modifiers are not integrated into the syntactic description.

The noun 'fortykning' (thickening) occurs in the corpus with arguments as well as modifiers

en fortykning af tarmen → argument
(*a thickening of the intestine*)

en fortykning på tarmen → modifier
(*a thickening on the intestine*)

but only the argument is encoded Dn1Pn-af.

However, we sometimes encode modifiers as part of the subcategorisation frame if they occur frequently - compared to the occurrences of other arguments or modifiers - and simultaneously are part of the 'core meaning' of the noun. It will be valuable for a parser for instance, to be able to recognise such modifiers. An example is 'afbræk' (*interruption*) that occurs 15 times in the corpus. In ten of these occurrences it appears with a prepositional phrase initiated by 'i' (*in*) e.g.

'et mindre afbræk i produktionen'
(lit: *a minor interruption in the production*)

and 'afbræk' is consequently encoded as a monovalent noun with the prepositional phrase initiated by 'i' as argument (Dn1Pn-i).

Large part of the frequent modifiers occurring with the nouns, however, is adverbial phrases indicating time or space, and these should certainly not be encoded.

Compounds and their Syntactic Behaviour

Nominal compounding is a very productive word formation process in Danish combining two nouns into one. Lots of the nouns encoded in the STO lexicon are nominal compounds. It is well known (see among others Ørsnes [1995]) that the non-head argument of a nominal compound can fill in an argument slot, thereby reducing the number of arguments of the compound, e.g.

Søsterens arrangement af brylluppet (Dn2GPn-af) →	Søsterens bryllupsarrangement (Dn1G)
(*The sister's arrangement of the wedding*)	(*The sister's wedding arrangement*)

In other cases the non-head argument does not fill in any argument slot. The noun 'program' (*programme*) is encoded with three linguistic descriptions in the lexicon:

Dn1Pn-for	→	et program for de unge (*a programme for the young*)
Dn1Pni-om	→	et program om tegning/at tegne (lit. *a programme about drawing/to draw*)
Dn1Pni-til	→	et program til bekæmpelse af fattigdom/til at bekæmpe fattigdom (lit. *a programme for fight against poverty/to fight against poverty*)

The nominal compound 'standardprogram' (standard programme) is encoded with all three valency patterns since 'standard' does not fill in the argument slot for the object. The nominal compound 'tegneprogram' (drawing programme), on the contrary, is encoded as an avalent noun (Dn0) since the non-head 'tegne' (to draw) fills in the argument slot.

Lexicalisation of compounds

As mentioned above, the STO lexicon contains many nominal compounds and here we are facing the problem of when to lexicalise a compound and when not. Since nominal compounding is a productive word formation process in Danish, we could have chosen not to lexicalise nominal compounds at all. Having the information of what linking element ('fuge element') the actual noun takes as the first part of a compound, any compound can be produced automatically. But it is not possible to account for the syntactic or semantic behaviour of such a compound automatically. Since it is not possible to examine the whole set of potential nominal compounds, we have decided to lexicalise the most frequent compounds found in corpora. The STO database includes information of the binding elements of each noun that is not a compound, so it will be possible to create compounds that have not been lexicalised.

Genitive

Only genitives subcategorised for are encoded in the syntactic descriptions. These are always to be encoded if they occur regularly in the corpus. This means that de-verbal nouns almost always are encoded with a genitive e.g. 'erkendelse' (*acknowledgement)*

direktørens erkendelse af at der var sket en fejl → Dn2GPntwi-af
(*the director's acknowledgement of the fact that a mistake had been made*)

It is important to notice the fact that relational adjectives often indicate the possibility for a genitive, see Ørsnes and Paggio [1994], here exemplified with a relational adjective denoting nationality:

Den irakiske invasion af Kuwait → Iraks invasion af Kuwait
(The Iraqi invasion of Kuwait → Iraq's invasion of Kuwait)

Possessive genitive is not encoded, and this implies that most concrete nouns are not encoded with a genitive since these normally only appear with possessive genitive. There are, however, exceptions to this rule.

Nouns that express 'family relations' and 'professions' are encoded with genitive. So 'søster' (*sister*) will be encoded with Dn1Pn-af:

Caroline Mathilde, der var søster til den engelske konge, George III
(*Caroline Mathilde, who was a sister of the English king, George III*)

and Dn1G:

prinsesse Anne of Denmark, senere Queen Anne, der var Christian IVs ældre søster.

(princess Anne of Denmark, later Queen Anne, who was Christian IV's older sister)

Abstract nouns are not encoded with genitive even if corpus occurrences reveal cases with frequent genitive that could not clearly be classified as possessive, e.g.

landets sprog
(The language of the country)

The important exception to this rule is nouns that express properties. It is however not easy to decide what a property is. We use the SIMPLE ontology [Lenci et al. 2000] as a point of departure for this decision. According to this ontology, properties can be divided into 4 categories:

Quality – *goodness, honesty*
Social Property – *authority*
Physical Property – *colour, strength*
Psychical property – *intelligence*

It often concerns de-adjectival nouns 'ærlighed' (*honesty*), but also non-derived nouns may be classified as properties, e.g. farve (*colour*).

'Broad' descriptions

The syntactic descriptions of STO comprise different syntactic constructions belonging to the same semantic reading, if possible. In the corpora we find the following realisations of the noun 'flytning' (*removal, displacement*)

deres flytning til Buenos Aires — genitive and P-til
(their displacement to Buenos Aires)

før Ingolf Gabolds flytning af TV-avisen falder på plads — genitive and P-af
(before Ingolf Gabold's displacement of the TV-news is arranged/settled)

flytning af Hærens Operative Kommando fra Århus til Karup
(displacement of the Army's Operational Command from Århus to Karup)
P-af + P-fra + P-til

In no instances does the genitive appear together with the three possible prepositional phrases. Nevertheless 'flytning' (*displacement*) is encoded with the description Dn4GPnPnPn-af-fra-til since an NP with 'flytning' and all the slots filled in is perfectly grammatical.

But this principle means that we do not distinguish between subjective and objective genitive in the syntactic encoding.

We could have chosen to encode the noun 'placering' (*placing*) as two syntactic units, the first one covering the description with the subjective genitive

SynU 1	Lemma	Description	Example
	placering	Dn2GPn-af	DSB's placering af den kommende S-togsordre i Tyskland (*The Danish Railway Company's placing of the future order for S-trains in Germany)*

Table 4: The first syntactic unit for 'placering'

and the second one covering the construction with the objective genitive which can alternate with a PP governed by 'af'

SynU 2	Lemma	Description	Example
	placering	Dn1G	hotellernes placering (lit: *the hotels' placing*)
		Dn1Pn-af	placeringen af hotellerne (*the placing of the hotels*)

Table 5: The second syntactic unit for 'placering'

But this encoding will cause over-generation due to the fact that the PP with 'af' of both syntactic units covers the same element. To avoid the over-generation, we have chosen to encode the noun in one syntactic unit (the first syntactic unit of the two units described above), well aware that we are loosing the information of whether the genitive is objective, blocking for the agentive element, or subjective.

Conclusion

The paper has presented some of the main problems we have encountered during the syntactic encoding process of nouns in the STO project. The linguistic architecture of STO is clear and stringent but when it has to be applied to data in greater scales, a lot of difficulties show up.

It often seems that the linguistic knowledge and the introspection of the lexicographers do not match corpus evidence. This is problematic since the STO lexicon is supposed to be corpus based and not all the conflicting occurrences can be said to be due to corpus gaps or the like. E.g. the principles for encoding genitive are quite clear and should be easy to follow, but here it seems that corpus occurrences often substantiate fewer genitives than the introspection of the lexicographer. It can be difficult to decide whether this is due to the corpora size or the fact that these are not well balanced or whether the genitive simply does not occur with this noun. In these cases it is recommended that the lexicographers search the web and only in cases where it is proved that the genitive does occur regularly or where there is very few occurrences even on the web to account on, the lexicographer should let his/her introspection overrule corpus evidence.

It seems to be a rather simple principle that each reading of a noun should have its own syntactic unit unless we are dealing with alternations, but it seems that the definition of an alternation is not quite unproblematic. Simultaneously the principle of letting a broader subcategorisation frame cover different realisations of a noun as long as it would be possible to fill in all the argument slots without changing the reading of the noun, sometimes makes it very complicated to decide the subcategorisation frame.

It can also be difficult to decide whether an element is an argument or a modifier but when frequent modifiers, central to the noun in question, might be part of the subcategorisation frame, the problem is not totally solved. Then the problem just lies in deciding which modifiers are frequent and central enough to be included in the subcategorisation frame.

In a computational lexicon data have to be more formalised than in an ordinary dictionary. This can lead to some interesting generalisations but it is also makes it very problematic to account for all the linguistic subtleties encountered in the language. The cases outlined in this paper are brilliant examples of this fact.

Acknowledgements

Since the contents of this paper build entirely upon the results of the syntactic encoding of the STO project, I want to thank all the project participants but particularly Nicolai H. Sørensen, Bolette Pedersen and Stig W. Jørgensen for their contributions to the principles of the syntactic encoding of nouns in the project, and thus to this paper.

Endnotes

1 The STO project gets funding from the Danish Ministry for Science, Technology and Development for three years. Center for Language Technology is the co-ordinator of the project and the work is carried out in collaboration with Institute for Computational Linguistics, Copenhagen Business School, Institute for General and Applied Linguistics, University of Copenhagen and The Institute of Business Information Technology.

2 We call all nouns that do not occur with arguments for 'avalent', so we do not make the traditional distinction between argument-taking noun types that can be avalent in some readings and the kind of nouns that never takes arguments.

References

[Braasch et al. 1998] Braasch, A., A. B. Christensen, S. Olsen & B.S. Pedersen, 1998. A Large-Scale Lexicon for Danish in the Information Society, in: *Proceedings from First International Conference on Language Resources & Evaluation*, Granada.

[Braasch et al. 2002] Braasch, A., A. Buhr, C. Navarretta, S. Nimb, S. Olsen, B.S. Pedersen, N. Sørensen 2002. *SprogTeknologisk Ordbog - Lingvistiske Specifikationer*, Technical Report, version 5, Center for Sprogteknologi, Denmark.

[Grimshaw 1990] Grimshaw, Jane B, 1990. *Argument Structure*, MIT Press, Cambridge, Mass., US.

[Lenci et al. 2000] Lenci, A., N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonoski, I. Peters. W. Peters, N. Ruimy, M. Villegas, A. Zampolli, 2000. 'SIMPLE – A General Framework for the Development of Multilingual Lexicons', in: T. Fontenelle (ed.) *International Journal of Lexicography Vol 13*. pp. 249-263. Oxford University Press.

[LE-PAROLE 1998] LE-PAROLE, 1998. *Danish Lexicon Documentation.* Internal report, Center for Sprogteknologi, Copenhagen.

[Underwood et al. 2000] Underwood, N.L. (ed.), 2000. *The LINDA Manual,* Working Papers, Center for Sprogteknologi, Copenhagen.

[Ørsnes & Paggio 1994] Ørsnes, B. and Paggio P., 1994. Maskinoversættelse af Substantivkomposita. In Baron, I. (ed.) NORDLEX-Projektet: *Sammensatte substantiver i dansk,* vol. 20 of LAMBDA, pp 135-57. København.

[Ørsnes 1995] Ørsnes, Bjarne, 1995. The Derivation and Compounding of Complex Event Nominals in Modern Danish - an HPSG Approach with an Implementation in Prolog, University of Copenhagen.