Towards a Strategy for a Representation of Collocations - Extending the Danish PAROLE-lexicon

Anna Braasch & Sussi Olsen

Center for Sprogteknologi
Njalsgade 80, DK-2300, Denmark
e-mail: anna@cst.ku.dk sussi@cst.ku.dk

Abstract

We describe our attempts to formulate a pragmatic definition and a partial typology of the lexical category of 'collocation' taking both lexicographical and computational aspects into consideration. This provides a suitable basis for encoding collocations in an NLP-lexicon. Further, this paper explains the principles of an operational encoding strategy which is applied to a core section of the typology, namely to subtypes of verbal collocation. This strategy is adapted to a pre-defined lexicon model which has been developed in the PAROLE-project. The work is carried out within the framework of the STO-project the aim of which is to extend the Danish PAROLE-lexicon. The encoding of collocations, in addition to single-word lemmas, greatly increases the lexical and linguistic coverage and thereby also the usability of the lexicon as a whole.

Decisions concerning the selection of the most frequent types of collocation to be encoded are made on empirical data i.e. corpus-based recognition. We present linguistic descriptions with focus on some characteristic syntactic features of collocations that are observed in a newspaper corpus. We then give a few prototypical examples provided with formalised descriptions in order to illustrate the restriction features. Finally, we discuss the perspectives of the work done so far.

1. About the STO project

The aim of the dictionary project STO[1] is to develop a large-scale Danish lexicon for language technology applications using the Danish PAROLE[2]-lexicon consisting of 20,000 general language entries as the point of departure. The establishment of the descriptive model and the linguistic specifications for STO greatly benefits from the experience acquired in the LE-PAROLE work. The lexicon will contain approx. 45,000 general and specialised language entries including semantic information part of which will be based on reuse of data and specifications from the SIMPLE-project[3]. These will result in approx. 100,000 semantic readings (meanings).

[1] SprogTeknologisk Ordbog, literally 'Language Technology Lexicon', i.e. a Danish lexicon for NLP applications. A project initiated by Center for Language Technology in Copenhagen (Braasch et al. 1998).

[2] The LE-PAROLE-project (Preparatory Action for linguistic Resources Organisation for Language Engineering) 1996-1998, developed NLP lexicons for 12 European languages provided with morphological and syntactic information.

[3] The LE-SIMPLE-project (Semantic Information on Multifunctional Plurilingual Lexica) extends the PAROLE-lexica with semantic information.

2. Lexicographic and computational aspects in combination

A considerable number of lexical units in a text are recurring bound word combinations. With the exception of valency patterns, these have until now not been incorporated into the STO lexicon. In order to extend the lexical and linguistic coverage, one of the most important tasks is to encode in the lexicon such word combinations, including collocations. To this end we have to set up a classification and want to develop an encoding strategy that accounts for the specific linguistic properties of collocation types and is compatible with the descriptive model used for single-word lemmas.

In all practical lexicography, one of the most discussed topics is the appropriate selection and description of lexical units that consist of more than a single word. It is well-known that they frequently cause problems not only for language learners but also for native speakers because bound word combinations cannot be understood or produced by using general rules of the language, i.e. they are complex units that cannot be treated fully compositionally (Moon, 1992; Heid, 1998). They can be regarded as coherent and (more or less) lexicalised building blocks of the language and thus they belong to the vocabulary. The lexicalisation of word combinations is a process of step-by-step progression which is influenced by different factors. The process results in a large number of cohesion types that can be classified along various axes (see e.g. in Benson et al. 1986; Alexander 1992).

In this connection, lexicographers are concerned with the following basic questions:

- what kinds of word combinations should be in the dictionary
- where is their proper position in the macro- and microstructure of the dictionary
- with which linguistic information should they be described.

In natural language processing (henceforth NLP) the property of non-compositionality is a crucial, but until now less elaborate, task to cope with. Generally, NLP systems are based on linguistic rules and regular patterns which describe the predictable and systematic behaviour of language; supplementary non-predictable behaviour and arbitrary choices are treated as exceptions to these rules. Linguistic information represented in a lexicon for NLP applications must be very detailed, unambiguous, explicit, exhaustive and formalised. Therefore, for NLP

systems, e.g. for machine translation, the lexicographer has to consider some additional questions originating from the specific requirements of computational applications.
In the present lexicon project a further essential aspect must be considered: The description of all lexical unit types must fit into the fixed PAROLE-model, and the linguistic specifications for STO (although they still are modifiable) must be followed. In this sense, morphological and syntactic patterns (including valency frames) that are already encoded must be reused in the encoding of new lexical entries.

3. The PAROLE-model of lexical description

As the point of departure we work with the PAROLE-model in a version that has been slightly modified for Danish. The model has originally been developed in the GENELEX-project and was reused in an extended version in PAROLE. It has a modular architecture comprising three independent, but linked layers of description according to a traditional division of linguistic information into morphological, syntactic and semantic types. The model is generic without a declared commitment to a particular linguistic theory. However, it is heavily inspired by the unification-based theory of Head-Driven Phrase Structure Grammar (Pollard & Sag 1994) which makes use of only very few grammar rules. All important syntactic and semantic processes are driven by information contained in the lexical entries.

One of the implications of the modularity is that linguistic behaviours of words are described independently and based purely on features observable at the particular levels in terms of morphological, syntactic and semantic units. A morphological unit contains the exhaustive description of inflection, information on part-of-speech, spelling variants and a few more properties. A syntactic unit contains information about the syntactic structures compatible with the lemma including valency, raising/control. Other syntactic properties of its prototypical syntactic environment can also be described here. The semantic level is not instantiated yet in our lexicon. Morphological and syntactic units are linked to each other according to their connection with the particular lemma.

Thus, this model does not operate with a pre-defined lexical unit similar to that in paper dictionaries. However, a 'dictionary entry' containing the lemma with all represented morphological and syntactic (and semantic) information can be compiled from the relevant units of the three layers of description. This description method has the advantage of not being static with regard to a presentation of the lexical item together with all related information in a single dictionary entry. In paper dictionaries, information is only linearly accessible beginning from the top of the entry.

Decisions regarding the representation of fixed expressions and collocations as lemmas or sublemmas in the structure of the lexicon are therefore in our context not of primary theoretical relevance, confer the discussion in Moon (1992, esp. pp.501-502) and Heer Henriksen (1995).

On the one hand, by using appropriate facilities of the database wherein the lexical data are stored (ORACLE), it is possible to link, to fetch and to present information from the three layers of the lexicon in several ways. On the other hand, from the practical point of view it is necessary to decide on systematic solutions. In the case of totally invariable word combinations it is appropriate to treat them in the same way as simple lexemes, i.e. as units of the morphological layer. In the same way a systematic treatment of bound word combinations, i.e. complex lexemes, must be decided on.

4. Criteria for discerning free and bound word combinations

Concordances produced by using the corpus tool XKWIC (see section 5 below) provide us with information about lexical co-occurrences in our corpus. The starting point is to study the findings in the concordances from two points of view. In computational corpus research, the statistical view on the frequency of word co-occurrences (see e.g. Sinclair 1991, p.109 ff) is the most prevalent one. The significance of co-occurrences of two or more words within a given 'collocational span' shows the degree of mutual affinities between these words. This quantitative criterion is very important, but used alone it would result in a too broad definition of the term 'collocation' which is inappropriate for practical lexicographical work. When used in combination with linguistic criteria that are more or less commonly agreed on, it provides a firm basis for pragmatic decisions (discussed e.g. in Cowie 1983; Cruse 1986; Benson 1986).

A preliminary definition of a bound word combination is formulated as follows: a frequently co-occurring word combination of two or more components showing a certain degree of structural and meaning cohesion. Frequent co-occurrences of words range from free word combinations over bound word combinations with increasing internal affinity and cohesion to fully frozen units.

Figure 1 (below) shows a classification of co-occurrences, deliberately oversimplified for illustration purposes; it is worth noting that there are many overlaps and probably also gaps between the categories mentioned below.

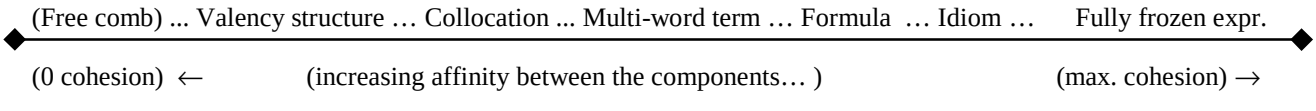

Figure 1: Internal cohesion of co-occurring words seen as a cline

With reference to the terms used in this classification, we deal with the class of collocations. Another terminology is used e.g. in Benson (1986), where collocation is considered a wider term for grammatical collocations (in our classification: valency structures) and lexical collocations (in our classification: collocations).

The word combinations extracted from our corpus are very heterogenous wrt their internal structure, syntactic function, degree of fixedness, semantic transparency, etc. In our case, the most important properties to be taken into consideration are restrictions on syntactic and lexical variability which basically differentiate bound word combinations from free combinations. In this respect, it is also important to discern bound word combinations consisting of a verb and a prepositional phrase from valency instances of a verb, having particularly strong subcategorisation and selectional restrictions.

The examples below illustrate that collocations (1) and (2) look very similar to instances of valency (3) and (4) on the surface:

(1) *tage til genmæle*
'reply' (lit.: take to reply)

(2) *tage [ngt.] i øjesyn*
'inspect [smth.]' (lit.: take [smth.] into eye's view)

(3) *tage til Berlin / i sommerhuset / på indkøb*
'go to Berlin /to the summer house /shopping'
(lit.: take to Berlin/ in the sommer house / on shopping)

(4) *tage [ngt.] i skuffen / fra skabet*
'take / get [smth.] from the drawer/ from the closet'

A valency structure contains a content word (verb, noun or adjective) and a grammatical structure (i.e. prepositional phrase, infinitive, finite or infinite clause) that the content word subcategorises for. Lexical entries in the STO lexicon contain a description of their individual subcategorisation requirements expressed in formalised valency patterns. Navarretta (1997) describes the development of valency descriptions of Danish verbs within the PAROLE-model. This method provides core syntactic information about structural compatibility in a simple and economical way.

Collocations consist of (groups of) content words (nouns, verbs, adjectives and adverbs) where one of the constituents typically carries the meaning and is the syntactically and semantically fixed part (base), while the other one has a weak meaning (collocate) and can be interchanged e.g. with a synonym or an antonym.

A rather comprehensive task is to deal with the lexical compatibility of words that occur in collocations because the choice of the semantically weak constituent is arbitrary and not predictable. In combination with the restrictions on lexical compatibility, collocations often show restricted internal variation of inflection and structure compared to parallel free word combinations.

In the following, we discuss the linguistic features of collocations that we regard as useful criteria for a subclassification and for the selection of frequent collocation types to be dealt with.

Basically, collocations are semantically transparent because of the recognised meaning of the base i.e. semantic core of the expression. However, going through our list of collocation candidates we experienced that the degree of transparency can vary quite a lot, therefore it is by no means straightforward to use semantic cohesion of co-occurring words as a primary classification criterion. Therefore, we concentrate our investigations on the following properties:

- syntactic label of the whole collocation (i.e. phrase type: VP, NP, ADJP or ADVP)
- part-of-speech (or syntactic label, if appropriate) of both constituents (base and collocate)

Additionally, it is necessary to check whether the collocation contains a unique component (not existing as independent lemma outside the collocation e.g. *øjesyn*) in order to ensure that all constituent words are encoded in the lexicon as single-word lemmas for reasons of searchability.

5. An outline of the practical work

Our investigation into recurring bound word combinations is based on two Danish corpora. The first and largest one comprises 20 mill. tokens from newspaper texts, the second one is a corpus of 4 mill. tokens from newspapers, magazines and books. None of the corpora are part-of-speech-tagged nor lemmatised, therefore the processing of corpus evidences involves several manually controlled steps, e.g. the manual partitioning of concordances into subsets based the part-of-speech information. Extension of the available corpora as well as tagging of the corpora is in progress. We use the XKWIC corpus tool (Christ 1993) for the corpus investigations.

In order to provide guidelines for encoding of collocations, we divided the practical work into the following sub-tasks:

- automatically producing concordances of common nouns and verbs that are already encoded in our lexicon and where the lexicographer noted in a comment that they occur in a great number of recurrent word combinations
- compiling lists of collocation candidates on the basis of these concordances with various sorting aspects to detect frequent collocation types
- manually selecting and extracting a few types for detailed analysis
- comparing the findings with the descriptive model and deciding on an appropriate description strategy
- setting up initial guidelines for linguistic description of the types selected
- starting testing and refining/extension-cycle

The core task was to select the most frequent types from the list of collocation candidates that are classified in terms of the following properties

- syntactic label of the collocation type: verbal phrase
- the part-of-speech of the base is

noun/nominal phrase - Vcoll = V+N/NP
prepositional phrase (PP) - Vcoll = V + PP

It is important to note in this connection, that in Danish the canonical word order of constituents in these types is: collocate – base, which is similar to English but different from German:

(5) *tage del i[ngt]*
'take part in [sth]',
'an [etw.+D] Anteil nehmen'

6. Representation of collocations in the PAROLE-model

In the PAROLE-project the encoding of single-word lexical items was in focus, and to our knowledge no attempts have been made yet by other language groups to encode complex lexical items, although the model is prepared also for this task. The model has its advantage in being very detailed and explicit and is provided with a comprehensive descriptive language.

6.1 Description with focus on the syntactic level

The following linguistic features are regarded as having primary relevance for the description of collocations

- complex structure containing at least one auto-semantic (content) word
- restricted morphological variability of the components compared with their free occurrences;
- restricted (morpho-)syntactic variability;
- a certain degree of meaning cohesion (restricted transparency).

In addition, a syntactico-semantic feature can be made explicit: collocations can function in texts similarly to single-word units. Monolingually a collocation, e.g. *stille krav* 'make a demand' (lit.: 'set demand') can often be substituted by a single word synonym *kræve* 'demand'; they are also often translated into a simple target language lexeme. In this paper, we do not discuss purely semantic features like the base-collocate relationships, and their impact on the semantic part of the description is only briefly mentioned.

The features above can combine in several different ways and they are almost inseparably bound to each other which makes a strictly modular desription a cumbersome task. Therefore it is useful to develop a method based on extensive use of patterns in order to describe (morpho) syntactic features of collocations piece by piece.

A pattern is in this sense a generalised description of a particular linguistic behaviour consisting of a unique combination of relevant information pieces which are expressed in terms of feature-value pairs. This is consistent with the method used for description of inflectional behaviours (we have implemented approx. 550 patterns) and for syntactic behaviours (approx. 700 patterns). A pattern in our model may describe one single, several or a large number of instances. (A pattern having just one single instance describes an exceptional behaviour.)

6.2 Towards a formalisation of syntactic restriction information

In the following section, we give a number of simplified examples in order to illustrate a pattern construction procedure. The linguistic properties described in these examples are recognised for each of the selected search words in a large number of corpus occurrences. One of the frequent Danish verbs, *tage* 'to take' has in its various inflected forms roughly 29,000 instances, of which the most frequent eight collocations make a total of approximately 8,000 occurrences, including the collocation *tage ansvar* 'take/shoulder the responsibility' with 3,128 occurrences. However, we are aware of the fact that such findings have rather limited value because of the size and the composition of the corpus (mainly newspaper texts).

Below, we focus on a few restriction types that affect subtypes of verbal collocations (Vcoll) in different ways.

General remarks

(a) The verb *tage* 'to take' is usually transitive outside these collocations, thus it is relevant to record restriction on passivisation.
(b) If the component (i.e. base noun or collocate verb) to be described behaves identically in free and in collocation-internal uses wrt a particular linguistic feature, then the collocation will not be marked for this feature.
(c) For the sake of clarity, we first mark each restriction separately, and finally the markings are combined into unique patterns that cover all restrictions for the particular collocation sub-type as it is shown in the last example.

Formalised markings

[ ] obligatory slot to be filled in e.g. with an object NP
< > optional slot that can be filled in e.g. with a prepositional object PP
( ) syntactic function of a constituent
{ } restriction to encode

6.2.1 Inflection: number and definiteness

Definiteness of nouns is in Danish expressed in two ways by a suffix or front article. The number of nouns is expressed by means of a suffix. In cases where the number of the noun cannot be recognised by a suffix or cannot be inferred from the noun-adjective agreement properties, we consider the noun singular indefinite (cf. Allen et al., 1995).

Vcoll = V + N(obj)
(6) *tage <u>kørekort</u>* N(obj){sing.indef}
'to take driving lessons/to pass one's driving test'
(lit.: to take driving licence)

Vcoll = V + N(obj) +PP
(7) *tage <u>æren</u> for [ngt]* N(obj){sing.def.}
'take (the) credit for [sth]'

Vcoll = V + N(obj) <+PP>
(8) *tage <u>ansvar/ansvaret</u> <for [ngn/ngt]>*
N(obj) {sing.indef /def.}
'take/shoulder the responsibility for sth'

(9) *tage hensyn til [ngn/ngt]* N(obj) {plu.indef.}
'show consideration for someone'/'take sth into consideration'

Although the noun 'hensyn' does not have a plural indefinite suffix, the number is inferred from the suffix of an attributively used adjective because of the agreement in number.

(10) *tage afsked med [ngn]* N(obj){sing.indef.}
'take one's leave/ take leave of someone'

Vcoll = V + NP(obj)
(11) *tage sin afsked* N(obj){determined by poss.pron, agreement with the subj.}
'resign' (lit.: take one's resignation)

6.2.2 Passive transformation of the collocation as a whole
Danish has two possible ways of expressing passive: the *'–s'* passive and the '*blive*' passive' marked as *'s'* and *'b'*, respectively (for further description see Allan et al., 1995). If neither of the passive forms is applicable, the marking is no_pass. The marking of passivisation restrictions below is not showed in detail (several combinations of passivisation restrictions are possible).

Vcoll = V+ N(obj)
(12) *tage kørekort* VP{no_b_pass}
'to take driving lessons/to pass one's driving test' (lit.: to take driving licence)

Vcoll = V + N(obj) + PP
(13) *tage æren for [ngt]* VP{no_pass}
'take (the) credit for [sth]'

The collocations below include an obligatory slot for a direct object which allows for passive transformation. Since both passive forms are applicable, the collocation pattern will not include information about the passivisation feature (cf. above: General remarks, (b)).

Vcoll = V + [NP(obj)] + PP
(14) *tage [ngt] i brug*'
'put [sth] into service' (lit.: take [sth] into use)

(15) *tage [ngt] i øjesyn*
'inspect [sth]' (lit.: take [sth] into eye's view)

6.2.3 Insertion of a modifying element
Adverbial modification of the verbal collocate - and thereby of the collocation as a whole - is nearly always possible without loss of the lexico-syntactic cohesion. However, this does not hold for idiomatic expressions, like *tage sit gode tøj og gå* 'walk out' (lit.: take one's good clothes and leave) but this is outside the scope of our presentation.

Modification of the base noun by attributively used adjective is either not allowed, restricted or freely allowed, depending highly on the particular Vcoll-subtype.

Adjective insertion is **not possible** at all
Vcoll = V + N + PP
(16) *tage bestik af [ngt]* N{n_a}
'take stock of [sth]'

Vcoll = V + PP
(17) *tage til genmæle* N{n_a}
'reply' (lit.: take to reply)

Adjective insertion **is possible**, but **semantically highly restricted** to a finite set of intensifying lexical items, e.g. *særlig* 'special', *stor* 'big', *afgørende* 'decisive'.

Vcoll = V + N + PP
(18) *tage hensyn til [ngt]* N{r_a}
'take [sth] into consideration'

Adjective insertion is **possible** and **semantically only weakly restricted** therefore we do not mark the noun for restrictions, cf. General remarks (b)

Vcoll = V + N <+PP> <+PP>
(19) *stille krav <til [ngn/ngt]> <om[ngt]* N
'make demand on [someone] for [sth]'

6.2.3 Pattern fragments
The examples below show the above selected restriction features in combinations, the first step towards a formalised description. They are not fully elaborated patterns; they just illustrate part of the formalisation.

(20) *tage kørekort*
Vcoll = **VP{no_b-pass}, N(obj) {sing.indef.}**, which prevents the generation of the following ungrammatical sentence

(21) **Kørekortet blev taget af ham i går*
(lit.: The driving test was passed by him yesterday)

but allows for the grammatical, impersonal sentence

(22) *Ved denne køreskole kan kørekort tages på en uge.*
'At this driving school driving licences can be purchased in a week.'

(23) *tage æren for [ngt]*
Vcoll = VP{no_pass}, N(obj){sing.def.}, which prevents the generation of the following ungrammatical sentence

(24) **En ære for det velgennemførte projekt blev taget af ham*
(lit.: a credit for the well accomplished project was taken by him)

but allows for well-formed sentences, like

(25) *Han tog æren for det velgennemførte projekt*
'He took the credit for the well accomplished project'.

6.3 Approaching a generalised encoding strategy: Verbal collocation types, their prevalent features related to description levels

A tentative overview in Figure 2 (below) illustrates a few types of bound word combinations with the focus on selected subtypes of collocations (Vcoll). The table represents a simplified illustration and it is not claimed to

be exact e.g. in pointing to the specific object of the conceptual model. It is only intended to give an idea about how to record information on prevalent linguistic properties of complex lexical items related to the layers of description.

The choices presented in the table are still subject to changes because of the stepwise development of the strategy. The linguistic features mentioned in the table are somewhat broadened compared to the aspects presented in section 6.1. Although the line of that section is followed in dealing only with surface variation of collocations, the table may need some comments.

The selection of types and features is not claimed to be exhaustive; for the sake of clarity only a few characteristic linguistic features are listed together with information on their allowed/restricted variability, e.g. the word order feature is not included in the table. In general, a variation of a feature is regarded as 'allowed' only if it applies to the collocation without loss of its lexico-syntactic and semantic cohesion. Continuity is a property of the internal structure regarding potential insertion of modifying elements, and it depends on the type and degree of internal cohesion. Morpho-syntactic properties are in general well suited for formalised description, as shown in the table. The degree of semantic cohesion regards the transparency of the meaning; this is probably the most difficult feature of a bound word combination to cope with because it is difficult to 'measure'.

The last column headed 'Birth level' points to the layer where the word combination type can be described as a lexical unit with a particular linguistic behaviour i.e. by means of the features belonging to that layer. The note on 'treatment facilities' points to some tentative choices as the conceptual model comprises several descriptive devices (complex and structured objects) that can appropriately be used in different combinations to describe just the same linguistic information content.

It is obvious that fully frozen, i.e. invariant bound word combinations can easily be treated like single word lemmas, i.e. as morphological units. Verbal collocations can be represented as specific syntactic units i.e. descriptions of single-word lemmas that occur in specific constrained constructions. Finally, an appropriate analysis leading to a satisfying treatment of semantic units is still pending.

Examples referred to in the table in Figure 2, below:

(26) *for alle tilfældes skyld* (lit. for all eventualities' sake)
'just in case'

(27) *elektronisk motorvej* (lit. electronic highway)
'information highway'
(28) *det hvide snit* (lit. the white cut)
'lobotomy'

(29) *stille træskoene* (lit. leave the clogs)
'kick the bucket';
(30) *finde sted* (lit. find place)
'take place'

(31) *stille* <+ > *spørgsmål, krav* (lit. set <+> question)
free insertion of regularly compatible elements like determiner, numeral, adjective
' to demand'

(32) *begå* <+ > *fejl/lovbrud* (lit.commit mistake/violation of law)
free insertion; the noun is restricted to semantic classes: law and mistake
'to make mistakes' / 'to break the law'

(33) *stille sin/den værste sult/ nysgerrighed* (lit: satisfy one's/ the worst hunger/curiosity)
the noun component is restricted: lexically to closed sets; morphosyntactically: a determinative element is required
'satisfy one's hunger/take the edge of the appetite' /
'satisfy one's curiosity'

(34) *tage kørekort* (lit. take driving licence)
the noun component is restricted wrt number and definiteness, *blive*-passive is not allowed.
'take driving lessons/ pass one's driving test'

(35) *gå ned* (lit:go down) 'break down; terminate'
(36) *sige til* (lit: say to) 'say the word'.

Feature / Type	Structure: Continuity	Lexical selection: Stability	Morpho-syntactic: Variability	Semantic: Transparency	'Birth' level of unit & treatment facilities	Ex. No.
Fully frozen expressions	**Yes**	**Total** ***(variation is not possible)***	***None***	**Cohesion, some transparency**	**Morphological unit**	**26**
Multi-word terms	**Yes**	**Total** ***(variation is not possible)***	***Only gram. agreement allowed***	**Full or partial cohesion**	**Morphological unit**	**27 28**
(Further types...)						
Fixed collocation V+N	**No**	**Total** ***N: Restricted to one single item***	**Partially frozen;** ***- insertion N is invariable*** **+V inflection …**	**Full cohesion**	**Morphological unit** ***List of components 'Rcompos', Restrictions on components…***	**29 30**
Collocation Light verb constr. (a) V+N	**No**	***N: Restricted to a few items***	**+V, N inflection +passive, +insertion +negation ….**	**Partial cohesion (verb meaning: -prototypical)**	**Syntactic unit** *Construction w. Synt label and function of 'Positions'… Linked Syntax & Semantics:e.g. 'Composition' list of allowed selection + indices 'RefLex' (lexical reference)*	**31**
Collocation type: Light verb constr. (b) V+N	**No**	***N: selection restricted by semantic type***	**+V, N inflection +passive, +insertion, +negation …..**	**Partial cohesion (verb meaning: - prototypical)**	**Semantic unit** *Unit of meaning with a core word + prototypical collocate (list)….*	**32**
Collocation type: Vcoll = V+ NP	**No**	***N: restricted to an enumerable lexical set***	***NP: restricted wrt number and definiteness*** **+V inflection +passive, +insertion, +negation …..**	**High cohesion (verb meaning: incorporated)**	**Syntactic unit?** *Restrictions on 'Positions'… <u>difficult to treat large lex.sets and high cohesion Move to semantics?</u>*	**33**
Collocation type: Vcoll = V + N	**No**	***N: restricted to a single item***	***N: restricted wrt .number and definitenes*** **+V inflection** ***-_b_passive, - insertion,*** **+negation …..**	**Partial cohesion (verb meaning: - prototypical)**	**Syntactic unit** *Restrictions on 'Positions' …*	**34**
Phrasal verb (a) V+Adv	**No**	**Total**	**+V inflection +passive, +insertion + negation ….**	**Full cohesion (all components)**	**Syntactic or semantic unit?** ***Different strategies used p.t.***	**35**
Phrasal verb (b) V+Adv	**No**	**Total**	**+V inflection +passive, +insertion + negation …..**	**Partial cohesion (verb meaning: - prototypical)**	**Semantic unit?** ***Strategies in discussion***	**36**
(Further types...)						

Figure 2: Selected bound word combinations and characteristic combinations of linguistic features

7. Conclusion

In this paper we focussed on the extension of an existing lexicon within the framework of the STO-project, considering the lexical coverage (the number of the lexical items) and the linguistic coverage (the types of lexical items). To achieve the best possible 'cost/benefit ratio' with respect to the extension, verbal collocation types were chosen as the first to be dealt with.

The approach presented brings together results of linguistic analysis, computational methods and application requirements. The general strategy we opted for was firstly, to subdivide information on complex linguistic features into many parts in accordance with the layers of description, secondly to formalise the information pieces in accordance with the descriptive language and finally, to link them coherently together through the layers. This strategy, developed in details for the encoding of verbal collocations, can be applied to further types of complex lexical items since it is adapted to a conceptual model that allows for complex and structured descriptions.

The selection and linguistic analysis of further frequent types of bound word combinations are still lacking and so is the establishment of practical encoding routines. Moreover, the quantitative and qualitative impact of the extension methods on the lexicon needs to be verified.

In a wider context, STO is the first national follow-up of the PAROLE-project but probably other national groups will follow. Therefore, it is important to be consistent with the PAROLE-model and descriptive methods in order to

ensure that the nationally produced lexicons remain compatible with each other. Multilingual linking of the lexicons for NLP applications will be an actual and challenging perspective.

8. Selected References

Alexander, Richard J. (1992). Fixed expressions, idioms and phraseology in recent English learner's dictionaries. In EURALEX '92 Proceedings, I-II,. Tampere (pp. 35-42).

Allan, R., P. Holmes. T. Lundskær-Nielsen (1995). Danish, A Comprehensive Grammar, Routledge, London and New York.

Bahns, J. (1996). Kollokationen als lexikographisches Problem, Niemeyer, Tübingen.

Benson, M., E. Benson, R. Ilson (1986). The BBI Combinatory Dictionary of English. A Guide to Word Combinations, Benjamins, Amsterdam, Philadelphia.

Blom, B. (1998). A statistical and structural approach to extracting collocations likely to be of relevance in relation to an LSP sub-domain text. In Nodalida '98 Proceedings.

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.

Calzolari, N., U. Heid, H. Khachadourian, J. McNaught, B. Menon, N. Modiano (1994). EAGLES LEXICON. Report on Architecture.

Christ. O. (1993) The Xkwic User Manual, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Cowie, A.P, R. Mackin, I. R. McCaig (1983). Oxford Dictionary of Current Idiomatic English. Vol. 2, Oxford University Press, Oxford.

Cruse, D. A. (1986). Lexical Semantics, Cambridge University Press, Cambridge.

Heer Henriksen, Berit (1995). Korpusbaserede relationsoplysninger og lemmatisering af flerordsforbindelser. In Nordiske studier i leksikografi III. (pp. 195-203), Reykjavik.

Heid, U. (1998). Towards a corpus-based dictionary of German noun-verb collocations. In Euralex '98 Proceedings, Université de Liège.

LE-PAROLE (1998). Report on the Syntactic Layer. Internal Report, Erli, Paris.

LE-PAROLE (1998). Danish Lexicon Documentation. Internal report, Center for Sprogteknologi, Copenhagen.

Moon, Rosamund (1992). Fixed expressions in native-speaker dictionaries, in EURALEX '92 Proceedings, I-II. Tampere (pp. 493-502.)

Navarretta, Costanza (1997). Encoding Danish Verbs in the PAROLE Model. In R. Mitkov, N. Nicolov & N. Nicolov (Eds.), Proceedings of Recent Advances in Natural Language Processing, Tzigov Chark, Bulgaria.

Pollard, Carl & Sag, Ivan A. (1994). Head-Driven Phrase Structure Grammar, The University of Chicago Press, Chicago & London.

Sinclair, John (1991). Corpus, Concordance, Collocation, Oxford University Press, Oxford.