Formalised Representation of Collocations in a Danish Computational Lexicon

Anna BRAASCH and Sussi OLSEN, København, Denmark

Abstract

In natural language processing one of the most interesting tasks is the treatment of complex lexical units. In this paper we focus on a specific collocation type and deal with a formalised, pattern-based description of collocations for a Danish computational lexicon, an extension of the PAROLE lexicon. We first describe our methodological approach to collocations from the NLP point of view. Secondly, we analyse a selection of frequent collocations found in our corpus and discuss a few selected morphological and syntactic constraints that apply to verbal collocations (our term). Published dictionaries are used to verify our findings. We then sum up the constraint types to which the selected collocations are subject and organise them in formalised patterns. Finally, we discuss a productive convergence of interests between traditional and computational lexicography in formalised representations.

1 Introduction

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 fully compositionally by using general rules of the language. (This is discussed from the lexicographer's point of view in e.g. [Moon, 1992] and [Heid, 1998]). Complex lexical units are coherent and (more or less) lexicalised building blocks of the language and belong as such to the vocabulary. The lexicalisation of word combinations is a process of step-by-step progression that is influenced by different factors. The process results in a large number of cohesion stages that can be classified along various axes, see e.g. in [Benson et al., 1986] and [Alexander, 1992].

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

- what kinds of word combinations should be in the dictionary
- what is their proper position in the macro- and microstructure of the dictionary
- what sorts of linguistic information should be encoded with the different word combination types.

In the following, we use the term 'dictionary' for published lexical data collections for humans, and the term 'lexicon' for computational lexical data collections for machine use, e.g. for natural language processing, henceforth NLP.

In NLP, non-compositionality is a crucial, but up till now a less studied problem for the lexicon. Generally, NLP systems are based on linguistic rules and regular patterns that describe the predictable and systematic behaviour of language; supplementary non-predictable behaviours and arbitrary choices are treated as exceptions to these rules. Linguistic information represented

in a lexicon must be very detailed, unambiguous, explicit, exhaustive and formalised. Therefore, e.g. for machine translation, the lexicographer has to consider some additional questions originating from the specific requirements of computational applications. For example it is obvious that word sense disambiguation is crucial especially to machine translation as well as to ontology-based information extraction. To this end, information on the variety and frequency of co-occurrences is essential. In order to discover which words co-occur and how frequent a collocation is, it is important to work both with existing dictionaries and large corpora, as discussed in [Fontenelle 1992a].

2 Approach and method

Our approach to the phenomenon of collocation is highly influenced by the design and practical development of a Danish NLP lexicon in the STO project. [1] This lexicon project is an extension of the Danish PAROLE [2] lexicon the structure and model of which will be briefly described in section 4. The aim of the project is to develop a large-scale Danish lexicon for language technology applications using the Danish PAROLE-lexicon consisting of 20,000 general language entries as the point of departure. The STO 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[3] -project. These will result in approx. 100,000 semantic readings (meanings).

Well aware that a considerable number of lexical units in a text are recurring bound word combinations, we regard encoding such word combinations, including collocations, in the lexicon as one of the most important tasks in order to extend the lexical and linguistic coverage. With the exception of subcategorisation information in the form of valency patterns (cf. section 3), these have until now not been incorporated into the STO lexicon.

Our investigation into recurrent bound word combinations is based on two Danish corpora. The first and largest one comprises 20 million tokens from newspaper texts; the second one is a corpus of 4 million 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. We are aware that our main corpus is not well balanced and that the predominance of newspaper text might result in a biased concordance.

In order to verify our findings, we also wanted to trace candidate collocations identified in our corpora in published dictionaries. To this end we chose three dictionaries of various types. For Danish, unfortunately there does not exist any collocational dictionary as such, unlike the situation for English. The only dictionary of a related type is Dansk Sprogbrug (‘Danish Usage’), henceforth DS, which is a ‘Dictionary of style and constructions’. We used the 1st edition (1976); however, a 2nd (only slightly modified) edition was published in 1995. The second monolingual dictionary used is Nudansk Ordbog, NDO [Politiken 1999], a medium-size dictionary of current Danish. Being the only one on the market, it is used both by native speakers and language learners. The third one is the largest Danish-English dictionary, Dansk-Engelsk Ordbog, henceforth D-E [Vinterberg/Bodelsen 1991]. The last two dictionaries are also published

on CD-ROM. These dictionaries were created for rather different purposes, but all of them rely on the native language competence of the users. Although all of them incorporate a number of collocations, none of them provides thorough and systematic descriptions that would be sufficient for language learners to avoid defective or inappropriate production of collocations in Danish.

This is not the right place to discuss this topic in more detail, therefore we just mention a few general observations that are valid for all dictionaries discussed:

- an explicit or clear classification of word combinations according to their fixedness, etc. is lacking
- the principle of selection of the keyword(s) is not always clear to the user
- word combinations including collocations are represented in canonical form or as an illustrative usage example of the keyword sense
- in case of more than one single presentation of a collocation in the macrostructure the information given in these presentations is not always identical, this is confusing for the user
- the sense-oriented ordering of collocations within the article in question (i.e. in the microstructure) seems in some cases to be contra-intuitive (NDO and D-E).
- explicit information given about linguistic properties specific to collocations is sparse, but is supported by examples which require human understanding of analogies, etc.

This illustrates the lack of and need for a quantitatively and qualitatively reliable dictionary of collocations - although we also found a large amount of valuable information in the electronically searchable dictionaries, which make related information given in different places easily accessible.

3 Collocations vs. valency structures from an NLP point of view

In this presentation we restrict ourselves to a subtype of recurring word combinations, i.e. to collocations, and we deal in particular with verbal collocations. We use the term ‘verbal collocation’ for all collocation types that function as a verbal phrase (VP) in a sentence. a verbal collocation consists of a verb (V) and another immediate constituent, which can be a noun (N), a noun phrase (NP), a prepositional phrase (PP) or an adverbial phrase (ADVP). This is somewhat different from the usual terminology, which [HEID 1998] uses, e.g. noun-verb (N + V) collocations, but it seems to be appropriate for the specifications for an NLP lexicon. The reasons for using the term ‘verbal collocation’ are the following:

- a sequential analysis working on surface text operates from the left to the right and the canonical and straightforward word order of these collocations is *V* + (N/NP/PP) as opposed to the German word order, which is N/NP + *V* (e.g. the Danish equivalent of German *‘in Gang kommen’* is *‘komme i gang’)*

- in an automatic sentence analysis (parsing) it is important to recognise a phrase as a potential collocation at an early stage, thus a look-up of the verb in the lexicon will flag this
- the primary focus of the lexicon is on morphological and syntactic features, thus the recognition of the semantically dominant constituent (base) in the recognition process would not be a trivial matter

Further investigations into semantic properties and types will probably lead to additional classification criteria and an appropriately updated description of collocations.

A preliminary definition of a bound word combination was formulated as follows: a frequently co-occurring word combination of two or more components showing a certain degree of structural and meaning cohesion. We deal especially with word combinations having a more specific cohesion than valency constructions. The term 'valency' is used in our approach not only for the number of arguments for which a particular verb subcategorises, but more generally for subcategorisation requirements of verbs, nouns and adjectives.

A distinction between syntactically subcategorised or valency structures (described by patterns) and lexical collocations is not always clear cut, because subcategorisation is often further specified wrt the selection of a particular co-occurring lexical item. A valency structure contains one content word (verb, noun or adjective) and a governed grammatical structure (such as a prepositional phrase, infinitive, finite or infinite clause). Collocations consist of two main immediate constituents that are (groups of) content words (nouns, verbs, adjectives and adverbs). The nominal constituent typically carries the meaning and is the semantically fixed part ('base'), whereas the verbal one has a weakened meaning ('collocate') which modifies the semantic aspect of the collocation and provides it with properties of verbal inflection.

In the following, we will focus on verbal collocations consisting of the verb 'tage' - that can combine with a wide selection of words, leaving us with very heterogeneous material - and a noun phrase, which in some cases is preceded by a preposition. We outline a method to describe the heterogeneous constraints to which these collocation types are subject. The selection of a representative number of collocations was based on an exhaustive analysis of the concordances extracted from the two corpora.

At the current stage of the STO lexicon, the most important properties to be analysed are constraints on syntactic and lexical variability, which basically differentiate bound word combinations from free combinations. In this respect, it is also important to discern collocations 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 in (1) and (2) look very similar to instances of structures subcategorised for in (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 stranden / i sommerhuset / på indkøb*
'go to the beach /to the summer house /shopping'
(lit.: take to the beach/ in the summer house / on shopping)
(4) *tage [ngt.] i skuffen / fra skabet*
'take / get [smth.] from the drawer/ from the closet'

The head noun of the PP in (1) does not allow for any modification at all, neither determination nor the insertion of an attributive, and the only modification allowed for in (2) is an adverbial modification of the verb. The valency bound PP's in (3) and (4) allow for all kinds of modifications of the head noun.

4 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 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. One of the implications of this 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 and raising/control. The semantic level is not yet instantiated 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 printed dictionaries, information is only linearly accessible beginning from the top of the entry. Decisions regarding the point of representation of fixed expressions and collocations in the structure of the lexicon (as lemmas or sub-lemmas) are therefore in our context not of primary theoretical relevance, in contrast to the questions discussed in [Moon 1992] and [Heer Henriksen 1995].

4.1 Perspectives for the use of the PAROLE lexicons

Each NLP system is devoted to a particular application (e.g. machine translation) and thus it has its own requirements concerning lexical coverage, information content, organisation and structure of the lexicon. The aim of PAROLE was to develop a 'poly-functional' lexicon type, which would be rich, powerful and flexible enough to provide relevant lexical information potentially to any NLP system. The ideal picture of lexicon development is a very large and comprehensive lexical database from which customised lexicons for individual systems can be derived on demand. The PAROLE lexicon concept represents significant progress towards such a general

database, which is one of the good reasons for choosing the Danish PAROLE lexicon as the point of departure for further development. Another good reason for our choice is that PAROLE lexicons have been developed for 12 European languages, a fact that makes multilingual links a challenging perspective in the future. Furthermore 10,000 lemmas of the PAROLE lexicons have been supplied with detailed ontology-based semantic information within the framework of the SIMPLE project.

5 Towards a formalised representation of collocations in the PAROLE model

The complexity of the features of collocations and the possibilities for their various combinations make a strictly modular description as required by PAROLE very cumbersome. Therefore it is useful to develop a method based on extensive use of patterns in order to describe morphological and 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, which is expressed in terms of feature-value pairs. This is consistent with the method used for description of inflectional behaviours.

In the following section, we give a number of simplified examples in order to illustrate the 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 a too homogenous composition of the corpus.

5.1 Selected constraint types

Below, we focus on a few constraint types that affect subtypes of verbal collocations (Vcoll) in different ways. The following should be observed

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

In the following tables, we use [ ] to indicate an obligatory slot to be filled in e.g. with an object NP, <> to indicate an optional slot that can be filled in e.g. with a prepositional object PP and () to indicate a syntactic function of a constituent.

5.1.1 Inflection

Definiteness of nouns in Danish is expressed in two ways: by a suffix or by a separate article. Number 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 subtype	**Number and Defin. of Obj.**	**Collocation example in canonical form**
V+N (obj)	Sing. indef.	*tage affære* 'intervene' (lit.: take affair)
		tage kørekort 'take driving lessons/to pass one's driving test' (lit.: take driving licence)
V+NP (obj +PP)	Sing. indef.	*tage bestik af [ngt]* 'take stock of [sth]'
	Sing. def.	*tage æren for [ngt/ngn]* 'take (the) credit for [sth]'
V+NP (obj + <PP>)	Sing. indef.	*tage stilling <til [ngt]>* 'make up one's mind about [sth]' (lit.: take attitude to sth)
	Sing. Indef/def	*tage ansvar/ansvaret <for [ngn/ngt*] > 'take/shoulder the responsibility for sth'
	Plu.indef.	*tage hensyn*[4] *<til [ngn/ngt]>* 'show consideration for someone'/ 'take sth into consideration'
V+NP (obj) NP: (poss.pron. obl.)	Sing. def.	*tage sin afsked* 'resign' (lit.: take one's resignation)

Table 1: Collocations and object number and definiteness

5.1.2 Passive transformation of the collocation

Danish has two ways of expressing passive: the *'-s'* passive and the '*blive*' passive' marked as *'s'* and *'b'*, respectively, for further description see [Allen et al., 1995]. If neither of the passive forms is applicable, the marking is No_pass. If both passive forms are possible there is no marking (Ø). The marking of passivisation restrictions below is not showed in detail (several combinations of passivisation restrictions are possible).

Vcoll subtype	**Passivisation of the collocation**	**Collocation example in canonical form**
V+N(obj)	Ø	*tage affære* 'intervene' (lit.: take affair)
	No_b_pass	*tage kørekort* 'take driving lessons/to pass one's driving test' (lit.: take driving licence)
V+NP(obj +PP)	Ø	*tage bestik af [ngt]* 'take stock of [sth]'
	No_pass	*tage æren for [ngt/ngn]* 'take (the) credit for [sth]'
V+NP(obj + <PP>)	Ø	*tage stilling <til [ngt]>* 'make up one's mind about [sth]' (lit.: take attitude to sth)
	Ø	*tage hensyn <til [ngn/ngt]>* 'show consideration for someone'/ 'take sth into consideration'
	No_pass	*tage ansvar/ansvaret <for [ngn/ngt*]> 'take/shoulder the responsibility for sth'
V+NP(obj) NP: poss.pron. +N (poss.pron. obl.	No_pass	*tage sin afsked* 'resign' (lit.: take one's resignation)
V+[NP(obj)]+PP	Ø	*tage [ngt] i brug* 'put [sth] into service' (lit.: take [sth] into use)
	Ø	*tage [ngt] i øjesyn* 'inspect [sth]' (lit.: take [sth] into eye's view)
V+PP	Ø	*tage til genmæle* 'reply' (lit.: take to reply)

Table 2: Collocation passivizability

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 possible (marking no_a) or semantically restricted (marking r_a) depending on the particular Vcoll-subtype. We expect to find some instances of collocations, which allow for unrestricted or at least very flexible sets of modifying adjectives but the material studied so far shows no instances of such a behaviour.

The marking r_a means that the insertion of adjectives is semantically highly restricted to a finite set of intensifying lexical items, e.g. *særlig* 'special', *stor* 'big', *afgørende* 'decisive'.

Vcoll subtype	Adjective insertion	Collocation example in canonical form
V+N(obj)	no_a	*tage affære* 'intervene' (lit.: take affair)
	no_a	*tage kørekort* 'take driving lessons/to pass one's driving test' (lit.: take driving licence)
V+NP(obj +PP)	no_a	*tage bestik af [ngt]* 'take stock of [sth]'
	no_a	*tage æren for [ngt/ngn]* 'take (the) credit for [sth]'
V+NP(obj + <PP>)	r_a	*tage stilling <til [ngt]>* 'make up one's mind about [sth]' (lit.: take attitude to sth)
	r_a	*tage hensyn <til [ngn/ngt]>* 'show consideration for someone'/ 'take sth into consideration'
	no_a	*tage ansvar/ansvaret <for [ngn/ngt]>* 'take/shoulder the responsibility for sth'
V+NP(obj) NP: poss.pron. +N (poss.pron. obl.)	no_a	*tage sin afsked* 'resign' (lit.: take one's resignation)
V+[NP(obj)]+PP	no_a	*tage [ngt] i* [5] *brug* 'put [sth] into service' (lit.: take [sth] into use)
	r_a	*tage [ngt] i* [2] *øjesyn* 'inspect [sth]' (lit.: take [sth] into eye's view)
V+PP	no_a	*tage til genmæle* 'reply' (lit.: take to reply)

Table 3: Adjective insertion

5.2 Formalised description

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.

- *tage kørekort*

  Vcoll = VP{no_b-pass}, N(obj) {sing.indef.}, which prevents the generation of the following ungrammatical sentence (noun definite, b_passive)

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

  but allows for the grammatical, impersonal sentence

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

- *tageæren for [ngt]*

  Vcoll = VP{no_pass}, N(obj){sing.def.}, which prevents the generation of the following ungrammatical sentence (noun indefinite, b_passive)

  (7) **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

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

Further properties that can be subject to constraints on the collocation are e.g. topicalisation, making sentences like the following ungrammatical

(9) **Æren for det velgennemførte projekt tog han*
'The credit for the well accomplished project he took'

and also pronominalisation of the object which prevents the following kind of sentences

(10) **Hvad angår æren, tog han den for det velgennemførte projekt*
'With respect to the credit, he took it for the well accomplished project'

Another aspect, which we have not dealt with here, is the choice of preposition following the noun in many Vcoll subtypes. The preposition of the collocation is often but not always the preposition which the noun subcategorises for. The noun *hensyn* 'consideration' normally appears with the complex preposition *over for* 'towards' but in the collocation *tage hensyn til* ('show consideration for someone'/ 'take sth into consideration') the preposition has changed. In our approach we simply bind the preposition to the collocational pattern, an approach which is appropriate and practical for NLP.

6 Lexicographic relevance of a formalised representation of collocations

The boundary between traditional lexicography i.e. the elaboration of dictionaries for human users and computational lexicography for NLP purposes is getting gradually less sharp. On the one hand, traditional lexicography makes increasing use of e.g. computational tools for corpus analysis, computer-aided editorial systems and electronic publishing media (CD-ROM and Internet). On the other hand, dictionaries elaborated electronically for human use make up valuable repositories of accumulated lexicographic experience and linguistic information, such as sense disambiguation and style marking, which can be reused in the construction of NLP lexicons. In the last two decades, much interesting research has been carried out in this field concerning the ability to exploit machine-readable lexical resources for NLP applications. A

comprehensive, general discussion of the topic can be found e.g. in [Boguraev/Briscoe 1989] and in [Heyn 1992]. The referenced literature deals mainly with English monolingual dictionaries; [Fontenelle 1992b] and [Heid 1994] concentrate on English and French. Less comprehensive basic investigations for Danish are documented e.g. in [Boje/Braaasch 1991] and [Braasch 1994].

We can think of an opposite direction of reuse, that is an exploitation of the material contained in a lexical database. Although this subject seems to be less frequently discussed, it is obvious that detailed information given in an unambiguous and explicit way readily could be utilised for other purposes than NLP. The STO lexicon will provide a more comprehensive description of collocations than can be found in existing Danish dictionaries.

More precisely the formalised description of each collocation contains

- its canonical form
- a list of features defined for the collocation type, instantiated by attribute-value pairs, such as phrase type, constituent structure, the syntactic function of each constituent, etc.
- a description of constraints on each of its constituents whenever observed (compared to the free and unconstrained occurrence of the constituent word)
- computational references (links) to the appropriate single word entries that are elements of the collocation. (These entries contain full inflectional patterns and syntactic subcategorisation patterns of the word, i.e. the description of their regular behaviours).

In our opinion, particularly lexicographers working on learner's dictionaries could take advantage of well-structured information that can be computationally derived from the lexical database. On the other hand, meaning descriptions and other semantic information - which make up an essential part of dictionaries - must be dealt with additionally.

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). For reasons of 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 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 bound word combinations are still outstanding and a design of practical encoding routines as well. 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 also 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 to each other. Multilingual linking of the lexicons for NLP applications is an actual and challenging perspective.

Notes

[1]SprogTeknologisk Ordbog ('Language Technology Lexicon'), i.e. a Danish lexicon for NLP applications. A project initiated by Centre for Language Technology, Copenhagen.

[2]LE-PAROLE-project (Preparatory Action for linguistic Resources Organisation for Language Engineering) 1996-1998 developed NLP lexicons for 12 European languages.

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

[4]The number of the noun in this example can only be inferred from an attributive adjective which agrees in number.

[5]It should be noticed that the object noun of these examples is an obligatory complement subcategorised for by the collocation as a whole. This implies that the canonical form of these collocations is discontinuous. The modified noun is the head noun of the PP.

References

[1] Alexander, Richard J. (1992). "Fixed expressions, idioms and phraseology in recent English learner's dictionaries". In *EURALEX '92 Proceedings*, I-II. Tampere.

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

[3] Bahns, J. (1996). *Kollokationen als lexikographisches Problem*, Niemeyer, Tübingen.

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

[5] 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.*

[6] Boguraev, B. & T. Briscoe (eds.) (1989). *Computational Lexicography for Natural Language Processing*, Longman, London and New York.
Boje, F. & A. Braasch (1991). 'Hvad får man skudt i skoene? Flerordsenheder i aktive ordbøger for mennesker og maskiner'. In R. Vatvedt Fjeld (Ed.) *Nordiske studier i leksikografi*, Oslo.

[7] Braasch, A. (1994). "How far do Printed Dictionaries and MT-Lexicons Share Information?" *In Studies in machine translation and natural language processing, Vol.8., Lexical Issues in Machine Translation* (eds. Alberto, P. & P. Bennet), EC, Luxembourg.

[8] 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.

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

[10] Christ. O. (1993) *The Xkwic User Manual*, IMS, Universität Stuttgart.

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

[12] Cruse, D. A. (1986). *Lexical Semantics,* Cambridge University Press, Cambridge.

[13] Fontenelle, Thierry. (1992a). "Collocation acquisition from a corpus or from a dictionary: a comparison". In *EURALEX '92 Proceedings II*, Tampere.

[14] Fontenelle, Thierry. (1992b). "Co-occurrence Knowledge, Support Verbs and Machine Readable Dictionaries". In *Papers in Computational Lexicography, COMPLEX '92.* Budapest.

[15] Heer Henriksen, Berit (1995). "Korpusbaserede relationsoplysninger og lemmatisering af flerordsforbindelser". In *Nordiske studier i leksikografi III.* Reykjavik.

[16] Heid, Ulrich. (1998). "Towards a corpus-based dictionary of German noun-verb collocations". In *Euralex '98 Proceedings,* Université de Liège.

[17] Heyn, Matthias (1992). Zur Wiederverwendung maschinenlesbarer Wörterbücher. Lexicographica Series Maior 45. Niemeyer, Tübingen.

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

[19] LE-PAROLE (1998). *Danish Lexicon Documentation.* Internal report, CST, Copenhagen.

[20] Moon, Rosamund (1992). "Fixed expressions in native-speaker dictionaries", in *EURALEX '92 Proceedings, I-II.* Tampere

[21] 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.

[22] Sinclair, John (1991). *Corpus, Concordance, Collocation,* Oxford University Press, Oxford.

Dictionaries:

Dansk Sprogbrug. En stil- og konstruktionsordbog af Erik Bruun (1978). Gyldendal, København.

Politikens Nudansk Ordbog med etymologi (1999). Politikens Forlag A/S, Denmark

Dansk-Engelsk Ordbog, Vinterberg, H., C.A. Bodelsen (1991). Gyldendal, København.