Published in LSP & Professional Communication Volume 2, Number 2, October 2002 - ISSN 1601-1929

Danish Verbs as Knowledge Probes in Corpus-based Terminology Work

Lotte Weilgaard Christensen
University of Southern Denmark – Kolding
Denmark

1. Background

The aim of this article is to present the first results of a project on the retrieval of terminological information from machine-readable Danish corpora in connection with practical terminology work. Works by Ahmad (1994), Bowker (1996), and Meyer & Mackintosh (1996) about linguistic signals, and especially verbs, inspired me to study the use of such signals to extract terminological information from a Danish corpus. Ahmad uses the term 'knowledge probe' to refer to lexical phrases and verbs which often occur together with terminological data in authentic texts and may thus be used as search patterns for identifying and extracting terminological information. I find his choice of denotation appropriate for terminology work, especially in connection with concept-related information, which can be defined as knowledge-rich information. In order to provide a focussed and fine-grained approach, a valency theory of verbs called the Pronominal Approach (PA) was implemented in my studies. The objective of my studies has been to examine whether the valency patterns of a number of verbs are suitable as knowledge probes and thus also as search patterns in corpus-based terminology work. This means that my studies have not only a theoretical aim, where I combine theories of terminology and verb valency, but also a most practical aim as I hope to provide the terminologist with an efficient search method when (s)he uses a corpus-analysis tool in practical terminology work.

As Danish is only spoken by a small language community, there is a lack of adequate commercial language technology tools for tagging, lemmatising, and parsing. Consequently, machine-readable corpora in Danish are normally raw texts without any kind of tags. Having no commercial language technology tools such as taggers and lemmatisers to support our search strategies, we have been forced to find other solutions. Thus, my purpose is to outline a linguistic approach suitable

for identifying and extracting terminological data without applying the above-mentioned tools.

The corpus consists of Danish texts within the subject area of hydraulics. The choice of one single subject area guarantees subject specific studies. Moreover, this approach is comparable with the terminologist's approach to a new subject area and is useful for evaluating the suitability of the method. Further, by concentrating on one subject area, a basis will be created for subsequently introducing other areas in order to conduct a comparative study.

Together with other similar studies to be carried out, this first study is to form the basis of a list or, as I should prefer to call it, a catalogue of important Danish verbs as knowledge probes. Generally speaking, it will appear from the catalogue which verbs normally occur together with synonyms, which ones occur together with definitions or parts of definitions, etc., and which will therefore be useful search patterns for synonyms, definitions, or parts of definitions. Thus, the catalogue offers the terminologist a search method enabling him or her to extract different terminological information categories. In this article I focus on the retrieval of definitions as well as information that ought to be part of definitions. Therefore, I shall refer to the verbs indicating definitorial information as definitorial verbs.

On the basis of experience, a terminologist will know what verbs indicate the occurrence of definitions. Being a terminologist myself, I will nevertheless register all verbs used in this very corpus. My systematic studies thus make it possible to list all verbs that provide terminological information. As to Danish, there is a great need for a complete list of relevant verbs and other identifiers because, compared with German texts for example, Danish texts include rather few explicitly classifying/categorising information.

The listed verbs, in this case the definitorial verbs, were compared with their entries in a valency dictionary of Danish verbs. This dictionary is called the Odense Valency Dictionary (OVD) and is based on the PA mentioned above. My argument for including the valency dictionary in the studies and not just compiling a simple list of verbs is to complete the list with the verbal valency patterns. According to my hypothesis the verbal valency patterns may be used as search patterns filtering out irrelevant occurrences (noise) and thus improving search results (precision). I shall illustrate the method in section 2, so for now I shall just outline the idea behind the method. In technical texts, the Danish verb *adskille* may have the meaning *take apart* or the meaning *differ from.* The latter, being the terminological reading, indicates the presence of coordinate concepts in text. This reading includes the preposition *fra* (from) in its valency pattern. A search for the verb and the preposition *fra* (from) of the terminological reading will exclude all or at least most occurrences meaning *take apart*. In what follows the term ‘reading’ is used for meaning as well as dictionary entry.

Section 2 includes a description of the general principles behind the PA and the OVD. Section 3 deals with the corpora used. In section 4 I shall give an outline of how I implement the PA in my studies of the verbs. In section 5 I shall sum up the more specific results of the studies; particularly to what extent the PA offers an efficient and suitable search method for practical terminology work based on machine-readable corpora. Also, a few linguistic observations will be presented. Finally, in section 6 I shall consider more general results.

2. The Pronominal Approach (PA) and the Odense Valency Dictionary (OVD)

From 1993-98 researchers at my institution, the Southern Denmark Business School – now the University of Southern Denmark, participated in the UDOG project (Udforskning af Dansk Ordforråd og Grammatik), which stands for the exploration of Danish vocabulary and grammar. This project was a research programme under the auspices of the Danish National Research Council for the Humanities (Statens Humanistiske Forskningsråd).

One aim of our sub project was to make a systematic valency description of Danish verbs for human and machine-readable dictionaries. Another aim was to compile a human and machine-readable valency dictionary of Danish verbs. The result is the Odense Valency Dictionary mentioned above. One further aim was to test whether the valency descriptions of the PA could in fact be used for language technology tools (natural language processing (NLP)). We focussed on machine translation, and tests were performed on the Metal system, owned then by the German electronics company Siemens of Munich, as the Danish modules of this system were developed at our institution.

As a basis for the systematic description the PA was chosen because it is based on a closed word class, namely the pronouns (Kirchmeier-Andersen 1997b: 69). Below the practical use of the PA is illustrated by means of the verb *bestå*. A basic assumption of the PA is the constant relation of proportionality between the pronouns and the nominal constituents, e.g. between the *woman* and *she* as well as between the *machine* and *it*. Moreover, the PA offers a sense distinction between the readings based on syntactico-semantic features provided by the pronominal paradigms of the language and a number of other distributional tests (Schøsler & Kirchmeier-Andersen 1997: 36).

In the OVD the first 3 syntactico-semantic features are derived directly from the pronouns established in the different argument slots as constituents, whereas the features 4-7 are derived indirectly by means of other distributional tests (Schøsler & van Durme 1996: 45). For my purpose I focus on the first 3 features:

1. **Syntactic forms:** noun phrase, prepositional phrase, adverbial phrase, sentence (finite, non-finite)

2. **Syntactic functions:** subject, object, prepositional objects, valency-bound adverbials
3. **Semantic features:** human, concrete, abstract, proposition, countability, manner, direction, etc.

4. Number of arguments (including optional arguments)
5. Type of auxiliary verb, type of passive and verbal aspect
6. Use of preliminary subject: *det* (it) and use of "existentials": *der* (there)
7. Linking phenomena, e.g. *jeg knækker grenen* (I break the branch) and *grenen knækker* (the branch breaks) (Kirchmeier-Andersen 1997b: 116)

In the following I illustrate the PA by means of the verb *bestå*, which has four readings in the OVD. During the coding process, the lexicographer has to decide for each argument slot (constituent) which pronouns are valid. The pronouns found will provide the input for a pronominal sentence, which is identical with the valency pattern of the verb studied. In addition to the pronominal example, the OVD includes a 'self-made' example in a declarative clause in the present tense. As will appear from the readings below, according to the PA the argument slots are defined from a surface syntactic point of view. Thus, there is one entry for each syntactic variation of a verb. This fact is very important for my approach. At present the OVD contains 3,300 different entries corresponding to 1,600 bare forms (lemmas).

Pronominal example	*'Self-made' example*
1) han/denne her/det består af ham/denne her/det (he/this one/it verb X him/this one/it)	bordet består af en bordplade og 4 ben (the table consists of a table top and 4 legs)
2) det består (så længe)* (it verb X (so long))	virksomheden består (i 50 år) (the firm has existed (for 50 years))
3) han består det (he verb X it)	eksaminanden består eksamenen (the candidate passes the examination)
4) det består i det/at+kompletivsætning/ at-infinitiv (it verb X in it/that+completive clause/ a Danish to+infinitive)	problemet består i svigtende service (the problem is due to lack of service)

* Parentheses indicate an optional argument

The first reading is the definitorial reading indicating a partitive definition. This reading takes two obligatory arguments: a subject and a prepositional object with *af* (of) which is the preposition selected by the verb (Jensen 1985: 71-72), cf. item 1 **Syntactic forms** and item 2 **Syntactic functions** above. From the pronominal slot fillers *han/denne her/det* (he/this one/it) for the subject and *ham/denne her/det* (him/this one/it) for the prepositional object it can be derived that the subject as well as the prepositional object may be realised with the semantic features human *(han/ham)* (he/him) or concrete *(denne her)* (this one) or abstract *(det)* (it), cf. item 3 **Semantic features** above. The important thing is, however, that by means of a

search pattern combining the verb *bestå* and the preposition *af* occurrences of the definitorial reading can be retrieved, and at the same time occurrences of the three other readings will be excluded.

As I already mentioned, the PA defines the argument slots from a surface syntactic point of view. Therefore I suggest that valency patterns derived by this method are especially suitable as search patterns for a focussed retrieval of terminological information categories from machine-readable corpora consisting of raw texts, which are e.g. not POS (part-of-speech) tagged. In POS-tagged corpora grammatical patterns such as ADJ+NOUN or NOUN+VERB may be used as search patterns (Meyer 2001: 290, Pearson 1998: 125-128).

3. Corpora

For my studies a technical corpus within the subject area of hydraulics was compiled. Different genres are included in the corpus, which at present consists of about 110,000 words.

In my opinion this size is adequate to illustrate the suitability of the PA as a search method for terminological data, but I am fully aware of the shortcomings in connection with a practical terminology project. During the collection of the corpus texts, I very soon had to realise that many texts of the chosen subject area are characterised by many graphics and formulae, which, as a result of the lack of appropriate software tools, reduces the range of texts suitable for the corpus.

As definitions are often found in textbooks, it is no mere coincidence that more than half of the corpus comes from this text genre. The textbooks implemented in the corpus are made up so that they represent different professional levels. However, during the collection it appeared that in addition to didactic and informative sections several of the textbooks also contain directive sections.

Almost 20,000 words derive from other types of informative text genres, i.e. encyclopaedias. Finally, documentation from the Danish enterprise Danfoss produced by the enterprise's hydraulics sales department is included. This documentation exceeds 20,000 words and contains text genres such as brochures, articles from periodicals, etc.

The OVD had some shortcomings in the valency descriptions of the verbs studied. Some of the verbs, i.e. verbs already studied as well as verbs to be studied in future, are not registered in the dictionary. As to other verbs, not all readings are entered, and in connection with a last group of verbs some of the readings are not completely described.

Owing to the shortcomings of the OVD and the limited size of the hydraulics corpus, it was necessary to consult another corpus. I had the possibility to access a corpus consisting of popular science texts from a project working on the coming

dictionary Den Danske Ordbog (DDO), which stands for the Danish Dictionary, a six-volume dictionary, which will be published in 2002/3. The DDO corpus consists of 6 million words. The DDO corpus has become more important for my studies than I had expected. Thus, in some cases I had to consult it systematically for a full description of the relevant, terminological verbs.

4. Studies of the chosen verbs

I studied 16 verbs. In my opinion this number is sufficient to test the suitability of the PA for my purpose. About half of the verbs are generally known as knowledge probes for terminological information, e.g. *definere* (define), *omfatte* (include), whereas the other half are not quite as obvious for the purpose, e.g *gælde for* (apply to), *skelne* (distinguish). These verbs were expected to indicate intensional definitions with the closest superordinate concept and distinguishing characteristics, or extensional definitions consisting of a superordinate concept and its subordinate concepts, or partitive definitions providing a superordinate concept and partitive concepts, i.e. concepts entering into a partitive relationship. The occurrences of the verbs in the hydraulics corpus were compared with their readings in the OVD.

In connection with each verb and its readings the following information was registered:

- the valency pattern of the readings defined by a pronominal example
- a corpus example or a 'self-made' example
- the valency-bound arguments
- readings realised in the hydraulics corpus
- the number of occurrences of each reading in the hydraulics corpus
- the terminologically relevant readings
- terminological information categories derivable by the relevant readings
- the suitability and quality of the retrieved data as input for terminological information
- proposals for search patterns though such patterns may depend on the query tool used
- concluding remarks related to each verb

The concluding remarks include observations about the ability of the search patterns to eliminate irrelevant occurrences; this indicates whether they are strong search patterns or not.

What follows is a summing-up of the main results of the studies of each of the 16 verbs. The verbs studied and the terminological information categories derived by

the relevant readings are listed in Table 1, section 5. More details can be found in Weilgaard Christensen (2000a, section 4, p. 247).

The verb *adskille* has five readings, two of them suitable as knowledge probes, i.e. reading 2 R2 and reading 3 R3, which are related readings. The second reading R2 *adskille sig fra* (differ from) takes three obligatory arguments: a subject, a reflexive object *sig* (oneself), and a prepositional object with *fra* (from), the preposition selected by the verb. In this reading, the subject and the object indicate coordinate concepts.

The reading R3 *adskille sig (fra) ved* (differ (from) by) takes three obligatory arguments and one optional argument indicated by parentheses: a subject, a reflexive object *sig* (oneself), (a prepositional object with the selected preposition *fra* (from)), and an adverb with the selected preposition *ved* (by). Provided that the optional argument is realised, this reading is partly identical with R2 above and thus provides information about coordinate concepts. Moreover, the adverb with the preposition *ved* (by) gives information about how the subject and the prepositional object differ from each other, i.e. it indicates a distinguishing characteristic.

The verb *bestå* has one terminologically relevant reading R1 (consist of), cf. above section 2.

The verb *definere* has only one reading. This reading takes two obligatory arguments and one optional: a subject, an object, and (an object complement with the selected prepositions *som* (by) or *ved* (by)). Occurrences with the preposition *som* indicate operational or intensional definitions, whereas the preposition *ved* is followd by an operational definition. The verb *definere* is terminologically interesting only in cases where the optional argument is realised, since it is irrelevant simply to be informed that something has been defined.

The verb *dele* has six readings. Two of them are useful for the extracting of terminological information. The reading *dele med* (divide by) R5 consists of two obligatory arguments and one optional argument: a subject, (an object), and a prepositional object with the selected preposition *med* (by). As was the case with *definere ved*, this reading, too, is suitable for the retrieval of operational definitions expressed by the prepositional phrase.

The reading *dele i* (divide into) R6 takes three obligatory arguments: a subject, an object, and a prepositional object with the selected preposition *i* (into) and indicates an extensional definition. The object refers to a superordinate concept, and the prepositional phrase refers to the corresponding subordinate concepts.

The verb *forstå* has six readings. Only *forstå* (understand) R6 is useful as a knowledge probe. This reading takes three obligatory arguments: a subject, an object, and a prepositional object with the selected preposition *ved* (by) and

indicates an intensional definition. The prepositional phrase is identical with the concept to be defined, whereas the definition is expressed by the object.

The verb *gælde* has six readings. Only the reading *gælde for* R2 (apply to) is terminologically relevant. This reading takes two obligatory arguments: a subject and an adverb with the preposition *for* (to). The subject, often realised by means of a completive clause, expresses a common characteristic of a concept expressed in the prepositional phrase, cf. the example in section 5, *Authentic corpora examples versus recommendations for good definitions*.

The verb *inddele* has two readings, which are both useful as knowledge probes. The reading *inddele i* (divide into) R1 takes three obligatory arguments: a subject, an object, and a prepositional object with the selected preposition *i* (into), and like the verb *dele i* R6, *inddele i* indicates an extensional definition.

The reading *inddele (i) efter* (divide (into) by) R2 takes three obligatory arguments and one optional argument: a subject, an object, and an adverb with the selected preposition *efter* (by), (a prepositional object with the selected preposition *i* (into)). In contradistinction to the reading R1 above, the prepositional object may not, since it is an optional argument, be realised in R2. The adverb with the preposition *efter* (by) indicates a type of characteristic.

The verb *indeholde* (contain) has two readings. Both readings take two arguments: R1 takes a subject and a quantifier phrase, and R2 takes a subject and an object. The readings are suitable for the extracting of partitive definitions or at least some of the coordinate, partitive concepts of a partitive definition. Without going into detail I shall just briefly mention that the quantifier phrase may in some cases serve as a distinguishing characteristic.

The verb *indgå* has two readings. Only reading R1 *indgå i* (form part of) is relevant as a knowledge probe. The reading takes two arguments: a subject and a prepositional object with the selected preposition *i*. The hits found were to some extent irrelevant. The useful corpus examples, however, indicate partitive definitions, the subject denoting the partitive concepts and the prepositional phrase the superordinate concept.

The verb *mene* (mean) has five readings. Reading R3 is the only terminologically relevant reading. It takes three obligatory arguments: a subject, an object, and a prepositional object with the selected preposition *med* (by). The reading indicates part of an intensional definition. Reading R4 is also realised with a prepositional object with the preposition *med* (with). However, only R3 can be used in the passive. The hits found were less significant than I had expected.

The verb *omfatte* (include) has two readings. They both take two arguments: a subject and an object. R1 is suitable for the extracting of primarily extensional and secondarily partitive definitions. According to my studies, only R2 can appear in

the passive. Moreover, the subject of R2 is realised with constituents such as laws and regulations which have, in addition to the semantic feature concrete, the semantic feature abstract, cf. section 5, *LSP readings*.

The verb *opdele* has three readings. The readings R2 *opdele i* (divide into) and R3 *opdele (i) efter* (divide (into) by) correspond to R1 *inddele i* and R2 *inddele (i) efter*. They have identical valency patterns and are suitable for the retrieval of the same terminological information categories.

The verb *sammensætte* has four readings. R3 *sammensætte* (compose) is the terminological reading which takes three obligatory arguments: a subject, an object, and a prepositional object with the selected preposition *af* (of). The noun of the prepositional phrase mainly occurs in plural. In the passive form, combined with the auxilary verb *være* (be), *er sammensat af* (be composed of) is useful as a knowledge probe for partitive definitions, of which the partitive concepts are indicated by the prepositional phrase.

The verb *skelne* has five readings. Three of the readings are suitable for the retrieval of terminological information. R1 *skelne (i)mellem* (distinguish between) takes two obligatory arguments: a subject and a prepositional object with the selected preposition *(i)mellem* (between). The noun phrase in the prepositional phrase appears in plural (cf. section 5, *Other linguistic signals*, 'generic expressions'*)* or is realised by two coordinate nouns indicating coordinate concepts.

R3 *skelne fra* (distinguish from) takes three obligatory arguments: a subject, an object, and a prepositional object with the selected preposition *fra* (from). In R3, the object and the prepositional object may express coordinate concepts.

R5 *skelne på* (distinguish by) takes three obligatory arguments: a subject, an object, and an adverb with the selected preposition *på* (by). The adverb denotes a type of characteristic.

The verb *tælle* has four readings. R2 *tælle blandt/til* (belong to) is useful as a knowledge probe and takes three obligatory arguments: a subject, an object, and a prepositional object with the prepositions *blandt* (among) or *til* (to). The object refers to one or more subordinate concept(s), and the prepositional object refers to the corresponding superordinate concept and thus to part of an intensional definition. However, this observation still has to be tested on the basis on more examples.

The verb *udgøre* (constitute) has two readings. There are very significant similarities between the readings of the verbs *udgøre* and *indeholde*, cf. above.

5. Results of the studies

In the following section I shall discuss the suitability of the PA as the basis of a search method for the retrieval of terminological information. In Table 1 below I have listed the definitorial verbs studied and described as well as the terminological information categories that were extracted during the working process from the corpora applied. The X marks in the table indicate which of the verbs can be used for the retrieval of specific terminological information categories. Moreover, I shall point out some aspects related to verb diathesis, semantics, LSP readings, and other linguistic signals than valency patterns. Besides, I will briefly look at aspects related to authentic corpora examples versus recommendations for good defining practice.

Suitability of the PA

On the basis of the study of the 16 verbs that form the input for Table 1, I believe that the PA is very suitable as a search method for readings with a valency pattern consisting of prepositional objects or adverbs. These adverbs have to be realised as prepositional phrases in which the preposition can be used as part of the search pattern. The fact is that many of the definitorial verbs studied appeared with prepositions as part of their valency pattern. Only three of the verbs studied, namely *indeholde* (contain), *omfatte* (include), and *udgøre* (constitute), do not have a valency pattern consisting of a preposition. In connection with a comparable study of another group of verbs (Weilgaard Christensen: 2000b), this tendency was not quite as significant as in this case. Therefore, I did not expect many patterns including a preposition; I had expected more patterns including a direct object.

Thus, optional arguments that occur with prepositions, i.e. *han definerer det (som/ved det)* (he defines this (by/by this)), proved to be essential for the search results as the prepositions can help to eliminate terminologically uninteresting patterns and ensure a more focussed search. The same holds for free adverbs with a rather consistent structure and frequency. However, the present study does not give such examples. In this way optional arguments and free consistent adverbs become obligatory in my study and consequently from a terminological point of view. As a result, I wish to argue that as a matter of fact the surface of the syntactic structure defined by the PA offers a suitable search method for terminological concept information.

Comments on Table 1

As mentioned above, Table 1 contains the terminological information categories found during my study (horizontal row) and their knowledge probes, i.e. the terminologically relevant readings of the verbs derived by the PA (vertical column). During my study more information categories than originally supposed could be added. As an example, I had not expected the predominance of the operational definition when searching for the verb *definere* (define) in combination

with *som* (by) and especially *ved* (by). The same goes for the verb *dele* (divide) together with *med* (by). An operational definition means a kind of ostensive definition that defines a concept by referring to a method or an operation (Balzer 1979: 15, Quist et al. 1983: 35). It even turned out that some of the verbs could be used as knowledge probes to identify each of the categories 'type of characteristic' and 'characteristics'. As to the verbs indicating characteristics, it seems possible to subdivide them into knowledge probes for either common or distinguishing characteristics. However, this observation still has to be tested thoroughly on the basis of more verbs.

The verbs *adskille sig fra* (differ from) and *skelne mellem/fra* (distinguish between/from) are categorised as knowledge probes for search results leading to co-ordinate concepts. In the readings of these two verbs subordinate concepts as part of extensional definitions were retrieved. However, their valency patterns did not guarantee the appearance of a superordinate concept. This is the reason why a separate column for co-ordinate concepts figures in the table. Moreover, it is worth noticing that some readings do not guarantee that all subordinate or partitive concepts occur in the sentences containing the knowlegde probes. This was for instance the case in connection with the verb *indeholde* (contain). Similar observations apply to the verb *tælle blandt/til* (belong to) where one or more subordinate concepts are expressed by the object whereas the superordinate concept is expressed by the prepositional phrase, cf. the column 'subordinate ∏ superordinate concept'.

To a certain extent, the order in which the information categories in Table 1 are listed reflects which types and frequencies of the terminological information categories I had expected to be able to extract by means of the verbs studied. And in fact most of the marks are found in the first columns of Table 1. I think that the table provides interesting information about my working process. The process can be compared with a 'top-down' approach in which I started out with different types of definitions and was able, during the working process, to add more specific terminological information categories, e.g. that of type of characteristic. However, this is, of course, attributable to the choice of the verbs studied.

The Danish readings in English, cf. Table 1: *adskille sig fra* (differ from*), adskille sig (fra) ved* (differ (from) by), *bestå af* (consist of), *definere (som)* (define (by)), *definere (ved)* (define (by)), *dele med* (divide by), *dele i* (divide into), *forstå ved* (understand by), *gælde for* (apply to), *inddele i* (divide into), *inddele (i) efter* (divide (into) by), *indeholde* (contain), *indgå i* (form part of), *mene med* (mean by), *omfatte* (include), *opdele i* (divide into), *opdele (i) efter* (divide (into) by), *er sammensat af* (be composed of), *skelne (i)mellem* (distinguish between), *skelne fra* (distinguish from), *skelne på* (distinguish by), *tælle blandt/til* (belong to), *udgøre* (constitute).

Information categories / Readings of the verbs	Intensional definition	Extensional definition	Coordinate concepts	Partitive definition	Operational definition	Type of characteristic	Subordinate → superordinate concept	Common characteristic	Distinguishing characteristic
R2 adskille sig fra			X						
R3 adskille sig (fra) ved			X						X
R1 bestå af				X					
R1 definere (som)	X				X				
R1 definere (ved)					X				
R5 dele med					X				
R6 dele i		X							
R6 forstå ved	X								
R2 gælde for								X	
R1 inddele i		X							
R2 inddele (i) efter		X				X			
(R1) R2 indeholde				X					
R1 indgå i				X					
R3 mene med	X								
R1 omfatte		X		X					
R2 opdele i		X							
R3 opdele (i) efter		X				X			
R3 er sammensat af				X					
R1 skelne (i)mellem			X						
R3 skelne fra			X						
R5 skelne på						X			
R2 tælle blandt/til							X		
(R1) R2 udgøre				X					

Table 1. Definitorial verbs as knowledge probes for terminological information categories

Verb diathesis

My studies yielded interesting results as to verb diathesis. The terminologically relevant readings of the verbs *adskille sig fra* (differ from), *bestå af* (consist of), *indgå i* (form part of) and especially *omfatte* (include) only occur in active constructions as these readings cannot appear in the passive, whereas other verbs, such as *forstå* (understand), are mainly realised in the passive form. As to the verb *sammensætte* (be composed of), my observations showed that the passive form combined with the auxilary verb *være* (be) constitutes a strong search pattern. In connection with the verb *mene* (mean), it still has to be confirmed by more tests whether the inflectional passive, adding –s, may be a good search pattern to eliminate irrelevant data seen from a terminological perspective. (Danish verbs may form the passive in three ways: 1. Inflectional, adding –s, 2. Analytical, combining with the auxiliary *blive* (become) and 3. Analytical, combining with the auxiliary *være* (be). The first two types correspond roughly to the German Vorgangspassiv whereas the last corresponds to the German Zustandspassiv (Kirchmeier-Andersen 1997a: 12 and 1997b: 71)).

Semantics

It was rather surprising that according to the description in the valency dictionary OVD the valency pattern of the verb *definere* (define) takes an object and an object complement with abstract constituents, cf. section 2, item 3 Semantic features. But actually both corpora confirmed the described semantic restrictions. Only in the DDO a couple of occurrences were realised with concrete constituents, which are atypical according to my studies. The interesting thing is that the PA offers the semantic information that the verb *definere* (define) appears with syntactical objects and object complements representing immaterial objects in a terminological sense.

LSP readings

With regard to my findings related to the verb *omfatte* (include), there is reason to assume that some definitorial verbs tend to occur in texts on specific subject areas. Thus, one reading of the verb *omfatte* (R1) will only appear in technical documentation, whereas the other reading of the verb (R2) is likely to be used in legal documentation only. So there is evidence that it will be possible to list a group of definitorial verbs appearing in all or at least most subject areas, while a number of readings will mainly or only occur in certain subject areas. To illustrate this, verbs used as knowledge probes for partitive definitions will logically be found in technical documentation in particular.

Other linguistic signals

In the cases in which the readings derived by the PA did not provide a suitable search pattern, it was quite often possible to find other linguistic signals. Some of the verbs studied are combined with generic expressions denoting a hyponymous relation (generic-specific, e.g. X is a kind of) such as *arter* (kinds), *typer* (types), *serier* (series), *grupper* (groups) or with a colon; all of them may indicate

information about subordinate concepts. In the hydraulics corpus the identifiers mentioned often occurred close to the verb *omfatte* (include) – a verb without a strong valency pattern for our terminological aim. Unlike *omfatte*, for instance the verbs *dele med/i* (divide by/into), *inddele i/efter* (divide into/by), *opdele i/efter* (divide into/by), and *skelne mellem/fra/på* (distinguish between/from/by) have a strong valency pattern, providing a strong search pattern, as they need not be combined with further identifiers except the prepositions belonging to their valency pattern. In combination with the verb *definere* (define), colon turned out to be a usable identifier, whereas the adjectives *samme* (same) and *tilsvarende* (corresponding) seem to be good identifiers when appearing in sentences together with the verb *gælde for* (apply to).

Authentic corpora examples versus recommendations for good definitions

On the symposium *Analysing LSP genres* arranged by the Aarhus School of Business (Denmark) in 1997, Margaret Rogers stated in her paper *Genre and Terminology:*

> "Even a strict evaluation and selection of documentation cannot change the fact that a terminology – descriptive or prescriptive – is an attempt to represent the system of the specialist lexicon **whereas text – evaluated or not evaluated – is language in use**." (Rogers 2000: 4) (my bold)

Margaret Roger's statement corresponds to my observations. As an example, verbs used as knowledge probes for intensional definitions may lead to sentences with incomplete definitions only consisting of the closest superordinate concept, but not the distinguishing characteristics. Similar observations go for sentences found by knowledge probes where you expect extensional or partitive definitions. In some cases only a few subordinate or partitive concepts occur so that you will have to find the missing concepts in other sentences of the corpus to complete the definition. In other cases the subordinate concepts found are unexpectedly not only mentioned but also examplarily defined by an intensional definition. A comparable study of another group of verbs (Weilgaard Christensen 2000b: 69) and this study documented that many of the corpus examples retrieved begin with demonstrative pronouns. That implies that the terms and the definitions or explanations very often appear in different sentences. The same observation is made by Pearson (1998: 151), who uses the term "complex formal defining expositives" about such appearances. I therefore conclude that it would be illusory to expect corpus-based terminological work to offer exemplary definitions and terminological information which can be entered directly into a terminological data base. So to take up the comparison with the 'top-down' approach again, it becomes evident that in connection with 'language in use' a 'bottom-up' approach cannot be ignored. In fact, you often have to combine information gleaned from different sentences and passages of texts in order to end up with an adequate definition. In such cases, the knowledge probes may provide invaluable assistance to the terminologist. To put it in Bowker's words (1996: 35): "They [these knowledge probes] help the terminographers to piece together the conceptual structure of a subject field and to

clarify the meaning of a term and its characteristics". Consequently, there is still a lot of terminological work to be done.

The following examples illustrate how searches for the verb *gælde for* (apply to) and *adskille* (differ from) supplement each other and provide the assistance described by Bowker:

> For komponentfittings gælder, [at dette er nipler for indskruning i komponenterne (ventiler, filtre, pumper m.m.)] (common characteristic)(Component fittings are nipples to be fitted onto components by screwing (valves, filters, pumps, etc.))
>
> *Fittings for komponenter adskiller sig først og fremmest fra hinanden [ved den måde, de tætner mod komponenten.] (distinguishing characteristic)*
> (The major difference in component fittings is the sealing method
> against the component.)

6. Concluding remarks

PA

Having demonstrated the usability of the PA, I would like to add some further aspects.

In very detailed, comparative studies from our participation in the UDOG project it was confirmed that out of all readings of a verb only a number of them will be realised in LSP texts (Weilgaard Christensen & Christoffersen 1999). Consequently, some of the readings dealt with in my studies are not very likely to appear in LSP texts, e.g. colloquial readings. That means that my studies may leave the impression that we should expect more noise from irrelevant data than is the case. Therefore it is quite obvious that only LSP corpora should be consulted when a list of relevant, verbal knowledge probes is compiled. In this way only readings occurring in LSP documents will be included in the studies. By employing this method, it is possible to achieve exact results about which readings lead to noise and consequently have to be eliminated. This would then again improve search strategies.

In my studies I concentrated on verbs without including term candidates in the search strings. Of course, the search pattern derived by the PA combined with the term candidate guarantees far better search results. Anyhow, my approach involved term candidates in another way because the information retrieved only presented data about term candidates that occurred with more detailed information in the corpus and were not just mentioned. Moreover, my method guarantees the retrieval of terminological information in sentences in which the term candidate is not mentioned but represented by a pronoun or the term of the superordinate concept or even a combination of these. On the basis of the 16 verbs examined, the PA turned out to offer strong search patterns. Therefore, I believe that the PA is an obvious point of departure for compiling a catalogue of specific, terminological knowledge

probes. Further, a catalogue as the one described will indicate whether the corpus applied is knowledge-rich in a terminological sense, and whether it is suitable for knowledge extraction in practical terminology work. If many knowledge probes figure in a word list generated from a corpus, the corpus will probably form a good basis for terminology work and not just for term extraction.

In her book *Terms in Context* (1998) Jennifer Pearson criticises some researchers for not using authentic data in their descriptions of definitions in LSP texts. I should therefore draw attention to the implementation of corpora and thus authentic data in my work. Moreover, the PA has made it possible to examine the data systematically and make observations about definitorial information 'in language use' that can, of course, only be retrieved by systematic studies of authentic data. Therefore, the PA and the authentic corpus data have formed and in future will form an important symbiosis in my studies.

7. References

Ahmad, K. & M. Rogers (1994*). The analysis of text corpora for the creation of advanced terminology databases*, Talk 2 on the 10th anniversary of the Southern Denmark Business School, Kolding

Balzer, W. & A. Kamlah (1979*). Aspekte der physikalischen Begriffsbildung: theoretische Begriffe und operationale Definitionen*. Braunschweig: Friedr. Vieweg. ISBN 3-528-08440-5

Bowker, Lynne (1996). Towards a corpus-based approach to terminography. In: *TERMINOLOGY* VOL. 3(1). Amsterdam/Philadelphia: John Benjamins Publishing Co. ISSN 0929-9971

Daugaard, Jan (1995). Valency – The Pronominal Approach applied to Danish, Russian, and Chinese. In: *Odense Working Papers in Language and Communication*, No. 8, March 1995. Ed. Daugaard, Jan, Institute of Language and Communication, Odense University, Odense. ISSN 0906-7612

Jensen, Per Anker (1985). *Principper for grammatisk analyse*. Copenhagen: Nyt Nordisk Forlag Arnold Busck. ISBN 87-17-03451-5

Kirchmeier-Andersen, Sabine (1997a). Linguistic Reflections on the Part-of Relation. In: *Nr. 22 OMNIS Workshop, November 1996*. Ed.: Hansen, S. L., Copenhagen Business School, Copenhagen. ISSN: 0902-0039

Kirchmeier-Andersen, Sabine (1997b). *Lexicon, Valency and the Pronominal Approach, An Application of the Pronominal Approach to Danish Verbs and Nouns*, Ph.D. Dissertation submitted at the University of Odense, April 1997

Meyer, Ingrid & K. Mackintosh (1996). Refining the terminographer's concept-analysis methods: How can phraseology help? In: *TERMINOLOGY* VOL. 3(1). Amsterdam /Philadelphia: John Benjamins Publishing Co. ISSN 0929-9971

Meyer, Ingrid (2001): Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. In: *Recent Advances in Computational Terminology*. Ed. Bourigault, Didier & Christian Jacquemin & Marie-Claude L'Homme, Amsterdam/Philadelphia : John Benjamins Publishing Company. ISBN 90 272 4984 9

Pearson, Jennifer (1998). *Terms in Context*. Studies in Corpus Linguistics. Amsterdam /Philadelphia: John Benjamins Publishing Company. ISBN 90 272 2269

Quist, A. H. M. & Agnes Horsager & Lotte Weilgaard Christensen (1983). En teoretisk redegørelse for eksplikationsbegrebet med udgangspunkt i definitionsbegrebet for at påvise tendenser til at anvende eksplikationer inden for et driftøkonomisk delområde. Unpublished MA-thesis. Aarhus School of Business, Aarhus

Rogers, Margaret (2000). Genre and Terminology. *Analysing Professional Genres*. Ed. Trosborg, Anna, Amsterdam /Philadelphia: John Benjamins Publishing Company. ISBN90 272 5089 8

Schøsler, Lene & Karen van Durme (1996). The Odense Valency Dictionary: An introduction. UDOG SHF 15-9018, Report No. 4. In: *Odense Working Papers in Language and Communication*, No. 13, September 1996. Ed. Karl-Heinz Pogner, Institute of Language and Communication, Odense University, Odense. ISSN 0906-7612

Schøsler, Lene & Sabine Kirchmeier-Andersen (1997). Studies in Valency II. The Pronominal Approach Applied to Danish. *RASK Supplement Vol. 5*. Odense University Press 1997, Odense. ISSN 1395-7236

Weilgaard Christensen, Lotte & Ellen Christoffersen (1999). Verbal Valency in Technical Texts - Some Characteristics. In: *UDOG Prefinal Report*. Ed. Sandford Pedersen, Bolette, Centre for Language Technology, Copenhagen

Weilgaard Christensen, Lotte (2000a). Danske verber som knowledge probes i terminologisk korpusarbejde. In: *I terminologins tjänst, Festskrift för Heribert Picht på 60-årsdagen*. Eds.: Nuopponen, Anita & Bertha Toft & Johan Myking, Proceedings of the University of Vaasa, reports 59, Vaasa. ISBN 951-683-857-X

Weilgaard Christensen, Lotte (2000b): Ekstraktion af terminologiske data vha. Den Pronominelle Metode. In: *Nordterm '99, Schæffergården 13.-15. juni 1999, Nordterm 10*. Ed. Larsen, Lianne, DANTERMcentret, Copenhagen

***

Published in LSP & Professional Communication Volume 2, Number 2, October 2002 - ISSN 1601-1929

ABSTRACT

Danish Verbs as Knowledge Probes in Corpus-based Terminology Work

Lotte Weilgaard Christensen
University of Southern Denmark – Kolding
Denmark

The aim of this article is to study the suitability of a number of Danish verbs as linguistic signals or rather knowledge probes for the retrieval of definitions in a Danish corpus on hydraulics. Moreover, this study will form the basis of a catalogue of knowledge probes for the retrieval of terminological information from machine-readable corpora in practical terminology work. The choice of one single subject area makes my study comparable with a terminologist's work, which, I believe, is very important for the evaluation of the method. A valency theory called the Pronominal Approach will be implemented to provide a more focussed approach than would be possible with a simple list of verbs. The basis of my method is to examine whether the valency patterns of a number of verbs are suitable as knowledge probes and thus also as search patterns in corpus-based terminology work. Further, I shall discuss how and to what extent the Pronominal Approach offers an efficient and suitable search method for practical terminology work based on machine-readable corpora consisting of raw texts that are neither lemmatised nor tagged. Finally, I look briefly at some aspects related to verb diathesis, semantics, LSP readings, and other linguistic signals than valency patterns as well as the importance of applying authentic corpora examples.

***