Natural Language Engineering **13** (1): 75–90. © 2006 Cambridge University Press
doi:10.1017/S1351324906004256 First published online 9 June 2006 Printed in the United Kingdom

Using shallow linguistic analysis to improve search on Danish compounds

BOLETTE SANDFORD PEDERSEN

Center for Sprogteknologi, University of Copenhagen,
Njalsgade 80, DK-2300 S, Denmark
e-mail: bolette@cst.dk

(*Received 13 May 2005, revised 2 January 2006*)

Abstract

In this paper we focus on a specific search-related query expansion topic, namely search on Danish compounds and expansion to some of their synonymous phrases. Compounds constitute a specific issue in search, in particular in languages where they are written in one word, as is the case for Danish and the other Scandinavian languages. For such languages, expansion of the query compound into separate lemmas is a way of finding the often frequent alternative synonymous phrases in which the content of a compound can also be expressed. However, it is crucial to note that the number of irrelevant hits is generally very high when using this expansion strategy. The aim of this paper is therefore to examine how we can obtain better search results on split compounds, partly by looking at the internal structure of the original compound, partly by analyzing the context in which the split compound occurs. In this context, we pursue two hypotheses: (1) that some categories of compounds are more likely to have synonymous 'split' counterparts than others; and (2) that search results where both the search words (obtained by splitting the compound) occur in the same noun phrase, are more likely to contain a synonymous phrase to the original compound query. The search results from 410 enhanced compound queries are used as a test bed for our experiments. On these search results, we perform a shallow linguistic analysis and introduce a new, linguistically based threshold for retrieved hits. The results obtained by using this strategy demonstrate that compound splitting *combined with* a shallow linguistic analysis focusing on the argument structure of the compound head as well as on the recognition of NPs, can improve search by substantially bringing down the number of irrelevant hits.

1 Introduction

The results and conclusions drawn from linguistically-based search experiments carried out during the last decade vary widely. Well known are those experiments done with WordNet (Voorhees 1994a, 1994b; Smeaton & Quigley 1996) and EuroWordNet (Gonzales, Verdejo, Peters & Calzolari 1998), where synonymy and other semantic relations such as hyponymy have been applied for enhancing queries. Loupy & El-Bèze (2002) and others also claim that the use of linguistic resources in general – including also other resources than semantic ones – improves search significantly. The TREC reports on natural language information retrieval are generally more neutral in their conclusions, stating that natural language processing

https://doi.org/10.1017/S1351324906004256 Published online by Cambridge University Press

seem to provide only quite modest results at its current stages (see, for instance, Strzalkowski, Gurthrie, Karlgren, Leistensnider, Lin, Perez-Carballo, Straszheim, Wang & Wilding 1996 Strzalkowski, Lin & Perez-Carballo 1997; and Strzalkowski, Stein, Bowden Wise, Perez-Carballo, Tapanainen, Jarvinen, Voutilainen and Karlgren 1998). Again, other investigations have led to the conclusion that query expansion based on linguistic resources tend to degrade results (cf. Allan 2000). In spite of such diverging opinions, there seems to a be general agreement on the fact that the quality and granularity of the applied resources obviously influence the results radically, and that a careful judgment of expansion strategies is necessary to achieve reliable results (cf. among others Loukachevitch & Dobrov 2004). Turning to the Danish scene, recent results provide promising results for linguistically based querying tuned to Danish text material using tagging, lemmatization and NP recognition for query and text analysis (cf. Paggio et al. 2003). These experiments include the application of the Danish SIMPLE semantic lexicon for query expansion based on synonymy and hyponymy, as are also reported on in Pedersen & Paggio (2004) and Andreasen, Jensen, Nilsson, Paggio, Pedersen and Thomsen (2004).

In this paper we focus on a specific search-related topic, namely search on Danish *compounds* and expansion to some of their synonymous phrases. Compounds constitute a specific issue in search, in particular in languages where they are written in one word, as is the case for Danish and the other Scandinavian languages. For such languages, expansion of the query compound into separate lemmas is a way of finding the often frequent alternative synonymous phrases in which the content of a compound can also be expressed. In other words, if we search for *byrådsmedlem* (`towncouncilFUGE`[1]`member`, 'town councillor'), search results containing *medlem af byrådet* (`member of towncouncil-THE`[2]) are presumably relevant, since the two expressions are synonymous.

Recent experiments for Swedish – a language which has the same characteristics as Danish regarding compounding – show that compound splitting improves search results substantially, mainly due to the fact that many queries containing compounds simply do not produce *any* search results without query expansio (cf. Chen & Gey (2003), Karlgren (2005) and Dalianis (2005)). In spite of these improvements in search, it is crucial to note that Dalianis gets 23% non-relevant hits when expanding to split compounds. The same scenario is reported on for Danish by Ankiro – a Danish company which develops language-based search engines for Danish, and which also performs compound splitting in several of their systems in order to improve search results (www.ankiro.dk).

On the basis of the findings concerning Danish and Swedish, we do not in this paper question the fact that compound splitting is a fruitful strategy in search. Query expansion seems to be a *precondition* for satisfactory search results – also in the light of the fact that both languages are becoming heavily influenced by English, meaning that misspelling of compounds (in two words) is becoming more and more frequent in texts. Rather we have asked ourselves the following questions: Does the internal structure of the compound tell us anything relevant about the likelihood of finding

[1] FUGE stands for linking element which can be either *-s-*, *-e-* or zero.
[2] THE stands for postponed determiner (*-en* or *-et*).

https://doi.org/10.1017/S1351324906004256 Published online by Cambridge University Press

relevant search results based on an expanded compound query? And can we by performing a *shallow linguistic analysis* discard a substantial part of the irrelevant hits found by splitting compounds? Although Scandinavian languages (as well as other Germanic languages) are specific in this context, similar phrase normalization problems have been investigated for English and have been treated also by using linguistic analysis, for instance as described in Strzalkowski *et al.* (1996, 1997). They use parsing techniques to locate so-called head+modifier pairs streams as a means to normalize English phrases like *weapon proliferation, proliferation of weapons,* and *proliferate weapons.* Although we limit ourselves to focus on the Danish compound problem, we pursue a similar approach in this work using a shallow noun phrase identification technique combining it, however, with detailed lexical information on predicate-argument structure.

Summing up, the aim of this paper is to examine how we can obtain better search results on split compounds, partly by looking at the internal structure of the original compound, partly by analyzing the context in which the split compounds occur[3].

2 Hypotheses and test scenario

2.1 Basic hypotheses

The two primary hypotheses for our experiments are:

- that some compounds are more likely to have synonymous 'split' counterparts than others; and
- that search results where both the search words (obtained by splitting a compound noun) occur in the same noun phrase, are more likely to contain a synonymous phrase to the original compound query.

An investigation of the first hypothesis calls for an analysis of the internal structure of the compound to see for instance if compounds where the head is derived from a verb act in a particular way and are eventually safer candidates for expansion than more lexicalized, non-predicative compounds. From several studies on compound nominalizations such as Levi (1978), Grimshaw (1990), Lapata (2002) and Ørsnes (1995), we know that verb-derived compounds follow certain phrase-like patterns inherited from the verb from which they are derived. It is therefore reasonable to assume that the content of verb-derived compounds is expressed also by several other phrase-like constructions – which can only be found by query expansion. Since syntactic valency features on nominalizations are encoded in the Danish computational lexicon, STO (Braasch & Pedersen (2002), Braasch & Olsen (2004))[4], this hypothesis can be pursued by associating these syntactic features to all compound heads.

[3] The experiments reported on have been carried out in the Danish research project, VID (VIden og Dokumenthåndtering med Sprogteknologi – Knowledge and Document Handling with Language Technology) in co-operation with five Danish companies, among others the technology provider Ankiro. The project was funded by the Danish Centre for IT research.

[4] STO – Sprogteknologisk Ordbase is a computational lexicon for Danish negotiated through European Language Resource Association (ELRA). The lexical database contains 65.000 entries with morphological and syntactic information, cf. www.cst.dk/sto.

https://doi.org/10.1017/S1351324906004256 Published online by Cambridge University Press

To test the second hypothesis it is necessary to look at the context of the split compound. This strategy can be seen as a variant to the classical closeness criterion: hits where the search words are close to each other are generally better than hits where they are far apart. However, the syntactically-based criterion opens for a more fine-grained analysis where hits with the two search words far apart – but still in the same noun phrase – can be given a high priority. This is desirable for instance in example 1 where the following hit can be found for the query *apoteksovertagelse* (pharmacyFUGEtakeover):

1) *Udgifter i forbindelse med overtagelse, nyanlæg eller flytning af et apotek*
   Expenses in relation to takeover, foundation or movement of a pharmacy

In example 1, the two search words are five words apart and yet belong to the same, complex NP. The hit is highly relevant to the query since *overtagelse … af et apotek* (takeover … of a pharmacy) is synonymous to *apoteksovertagelse*. In contrast, in example 2 the search words are found within the distance of four words, but they do not belong to the same NP. The hit is in fact less relevant since we are no longer dealing with a synonym phrase to *apoteksovertagelse*:

2) *Han lod ved overtagelsen foretage en optælling af apotekets varelager*
   He let at takeover-THE make a counting of pharmacy-THE-GEN[5] stock

In 2) the two query words are in the same sentence but not in the same NP. One could argue here that expanding on the genitive form (*apotekets*) could give rise to many irrelevant search results; however, other examples demonstrate that genitive forms are often substituting the non-head of the compound in a subjective genitive construction with deverbal nouns, cf. example 3 as a very relevant hit for the query *biblioteksdrift* (librabryFUGEworking):

3) *Bibliotekernes daglige drift*
   Libraries-THE-GEN daily working

In conclusion, the second hypothesis opens for a more nuanced analysis of the context in which the split compound occur. By focusing on the NP we believe it is possible to locate with greater precision the split compound occurrences which are actually synonyms to the original compound.

2.2 Split compounds used as queries in the test scenario

To develop a test bed for our experiments, a compound list of 1000 Danish compounds was generated by the firm Ankiro selecting randomly from a list of 60.000 compounds harvested from a text database, KommuneInformation, provided by the general Danish municipalities. KommuneInformation (www.kommuneinformation.

[5] GEN stands for genitive-*s*.

https://doi.org/10.1017/S1351324906004256 Published online by Cambridge University Press

dk) is a large public database containing text material relevant to all Danish citizens such as government orders, laws, departmental circulars, as well as more general municipality information. When the experiment was run, the database contained 64 million running words. The text material is generally characterized by complex sentence structures influenced by the fact that it lies mostly within the juridical domain.

All compounds were automatically split using the firm's compound analyzer which applies a firm-specific computational lexicon and a set of exception rules. In cases of more than two lemmas in a compound, the coherent boundary was calculated in order to identify which two lemmas (of which one was itself a compound) to carry on with in search; as in *aktie/beskatningsregel* (`share/taxationFUGErule`) in contrast to *arbejdsgiver/medlem* (`workFUGEgiver/member`, 'employer member').

Due to a long training phase and a very large set of harvested compounds to match candidate compound components up against, the error percentage on compound splitting is less than 1%.

The query test set was modeled as a list containing (split) compounds in isolation – i.e. not with compounds as part of longer queries. The split compounds were used to search in the same database. Hits for split compounds have been defined as text excerpts where the two query words – in any inflectional form – were found within a distance of 10 words. For each query a limit has been set to a maximum of 200 hits.

2.3 Classification of the test compounds

Seen from a morphological perspective, the corpus material is rather homogenous in the sense that it contains only nominal compounds with slightly different internal structures:

- noun + noun, as in *båd+designer* (boat designer),
- noun + noun-noun compound, as in *aktie* + (*beskatnings+regel*) (`share + (taxFUGErule)`),
- noun-noun compound + noun as in (*arbejds* + *giver*)+*medlem* (`workFUGEgiver) + member`) 'employer member'.

Thus, compounds composed of other word classes than nouns are not treated in the experiment. However, nominal compounds can be further subdivided with regards to the word class that the head is derived from. Since this aspect plays a role in our hypotheses, we have classified the test material into three classes:

- **Verb-derived compounds**, i.e. compounds where the head is derived from a verb: *depotanbringelse* (`depositinvestment`), head derived from *anbringe* ('to invest'),
  råstofudvikling (`rawmaterialextraction`), head derived from *udvikle* ('to extract').
- **Other predicative compounds**, i.e. where the head is not derived from a verb but takes an argument anyhow:

https://doi.org/10.1017/S1351324906004256 Published online by Cambridge University Press

bistandspligt – pligt til (`aidFUGEduty` – 'duty to'). Includes also deadjectival compounds as in: *dækningsmulighed – mulighed for* (`coverFUGEpossibility` – 'possibility for').

- **Non-predicative compounds**, i.e. compounds where the head does not take arguments:
  forsikringstaksator (`insuranceFUGEvaluer`) and *reaktortank* (`reactortank`).

As mentioned previously, these three classes are all automatically identifiable when using a computational lexicon with syntactic information; we have therefore made use of the Danish STO lexicon. In this lexicon, deverbal nouns are encoded differently than other valency-bearing nouns; likewise, the encoding of zero-valent nouns (class three) differs from that of the other two classes.

Furthermore, we have analyzed verb-derived compounds in order to see which role the non-head plays in relation to the head. Again, this relates to our first hypothesis: do some role fillings make the compound act more regularly than others regarding synonymous alternatives? According to Grimshaw (1990) and others, the role of the non-head is influenced by the syntactic and semantic characteristics of the verb of which the head is derived. On Grimshaw's approach (1990:14–15), the following restrictions seem to apply to English deverbal compounds:

"when the head takes more than one internal argument, the least prominent must be inside the compound, and the prominent outside",

as is seen in examples 4a and b:

4 a) *cookie-baking by children* (*children bake cookies)* vs.
  b) **child-baking of cookies*

In these examples, *cookie* (typically considered to have the role 'theme') is considered less prominent than *child* ('agent') according to a hierarchy of semantic roles; therefore it can occur as the non-head in the compound, in contrast to *child*.

Danish, however, seems to be less restrictive in this respect, as argued by Ørsnes 1999:109, as seen in example 5:

5) *mediadækningen af damefodbold*

`mediacoveringTHE of ladyfootball`

where the agent role in the first compound is played by *media*: *medierne dækker damefodbold* (the media cover female football). Nevertheless, these exceptions to Grimshaw's principle are apparently rather rare; at least they do not occur in our material. In our documents, the agent role only occurs as non-head in compounds where the head is derived from an intransitive verb as in example 6:

6) *tandlægeindgreb* (*tandlægen griber ind*)

`dentistintervention` (the dentist intervenes)

In contrast, in compounds where the head is derived from a transitive verb, the non-head typically fills the 'theme' role, as in 7:

7) *apoteksovertagelse* (*nogen overtager et apotek*)

`pharmarcyFUGEtakeover` (somebody takes over a pharmacy)

https://doi.org/10.1017/S1351324906004256 Published online by Cambridge University Press

Finally, an adverbial role occurs in some cases, namely that of location, as in:

8) *gartneriproduktion* (*nogen producerer noget i gartneriet*)

   `marketgardenproduction` (somebody produces something in a market garden)

Non-heads playing thematic roles, e.g. the beneficiary role, also occur in predicative compounds of the non-event type:

9) *børnetilskud (tilskud for børn)*

   `childrenFUGEsubsidy` (subsidy for kids)

10) *chaufførsats (sats for chauffører)*

    chaufeurrate (rate for chauffeurs)

Compared to predicative compounds where theme is by far the most frequent role in the test material, the diversity of roles that can be filled out by the non-head in non-predicative compounds is overwhelming (see, for instance, Järborg (2003) for 14 different role fillings for Swedish, Mueller (1998) for 30 different categories in English; and Johnston & Busa (1999) and Søgaard (2004) who apply Pustejovsky's (1995) qualia roles (encoded on the compound head) as a means to calculate the thematic role of the non-head). This diversity of roles is also found in our test material, and we have chosen not to go into a deeper a study of these different roles since we have currently no reliable means to calculate them automatically[6].

3 Shallow linguistic analysis of the search results

We now turn to the automatic shallow analysis of the search results where the aim is to introduce a new, more critical threshold for retrieved hits on split compounds. The material contains hits on enhanced queries only, search on the original compound is not performed as a part of this experiment. However, since the compounds are harvested from the same text corpus, we know that in these cases all examined compounds would give hits also to the original query.

The search results have been tokenised and pos tagged, and NPs have been tagged using a recognizer written in the CASS formalism (cf. Abney (1996) and for the Danish implementation, Haltrup (2000)). Noun phrases including up to two prepositional phrases are recognised. Then, all search hits have been assigned a relevance score according to the following algorithm:

$Score = 10(I - N)$ where
I = number of found search words (in any morphological form) in the search hit and
N = number of NPs in the search hit which contains a search word (in any morphological form).

[6] Søgaard (2004) has experimented on Danish compounds by using the qualia encodings in the Danish SIMPLE lexicon. The coverage of this lexicon is however not broad enough to cover our material.

https://doi.org/10.1017/S1351324906004256 Published online by Cambridge University Press

Score for search words being in the same noun phrase	Enhanced query: opgørelse omsætning (‘account turnover’)
20.0	andelsforeninger som nævnt i § 16A , stk . 1 , og som ikke er medlemmer , skal hverken medregnes ved **[NP1 [NP [N <T>opgørelsen</T>]] [PRÆP af] [NP [N_GEN andelsforeningens] [N <T>omsætning</T>]]]** med henholdsvis medlemmer og ikkemedlemmer eller ved **[NP1 [NP [N <T>opgørelsen</T>]] [PRÆP af] [NP [PRON_DEMO den] [ADJ samlede] [N <T>omsætning</T>]]]**
10.0	vedrørende sine subsidiære påstand henvist til , at det ikke af Østre Landsretsdommen fremgår , at de af klageren udarbejdede **[NP1 [NP [N <T>opgørelser</T>]] [PRÆP over] [NP [V_PARTC_PAST mistet] [N <T>omsætning</T>]]]** i andet halvår af 1985 og første halvår af 1986 har dannet grundlag for rettens udmåling af erstatningen
0.0	**[NP [PRON_DEMO Denne] [N <T>opgørelse</T>]]** foretages på **[NP2 [NP1 [NP [N grundlag]] [PRÆP af] [NP [N <T>omsætningen</T>]]] [PRÆP i] [NP [PRON_DEMO det] [ADJ foregående] [N regnskabsår]]]** . Efter regnskabsårets udløb foretages endelig regulering af afgiftstilsvaret . For nyetablerede virksomheder sker den

Fig. 1. Search results on *opgørelse* and *omsætning* (‘account’ and ‘turnover’) enhanced from the original query *omsætningsopgørelse*.

As discussed and illustrated in section 2.1, the motivation behind the algorithm is based on the hypothesis that search results where both the search words (obtained by splitting a compound noun) occur in the same noun phrase, are more likely to contain a synonymous phrase to the original compound query. In other words, the algorithm ensures that a higher score is assigned to hits that fulfill this criterion. For illustration, let us consider the first hit in figure 1, where the two search words *omsætning* (turnover) and *opgørelse* (account) each occur twice. Following the algorithm, I is therefore set to 4. The number of NPs containing one or more search words in the hit is 2, so N is set to 2. In this case $I - N = 2$, amounting to a score of 20. In other words, it is implicitly calculated that the hit contains two NPs in which both search words occur (Jongejan *et al.* 2004)[7]. Likewise in the first hit in figure 2, four occurrences of the two search words *hud* (skin) and *reaktion* (reaction) are found, so I is again set to 4. However the number of NPs containing one or more search words in the hit is in this case three, so N is set to 3. $I - N = 1$ and a score of 10 is therefore assigned to the hit. Finally, if we consider the second hit in figure 2, we find three occurrences of the search words, so I is set to 3. The number of NPs containing one or more search words in the hit is now also three, so N is also set to 3. Therefore, the hit achieves a score of 0.

According to this algorithm the threshold has been set to 10; search results with a score below 10 are thus considered ‘not retrieved’, whereas results with a score of 10

[7] Actually, in cases where the same search word occurs twice within the same NP, the weight of 10 is also (erroneously) assigned even if the other search word is outside the NP. This phenomenon occurred only once in our material, and has thus not really affected the results negatively.

https://doi.org/10.1017/S1351324906004256 Published online by Cambridge University Press

Score for search words being in the same noun phrase	Enhanced query: reaktion hud (‘reaction skin’)
10.0	pågældende organer . Tilpasningsreaktioner (f. eks . indvandring af makrofager i lungevævet , leverhypertrofi og enzyminduktion , **[NP1 [NP [ADJ hyperplastiske] [N <T>reaktioner</T>]] [PRÆP på] [NP [ADJ irriterende] [N stoffer]]]**) . Lokale **[NP1 [NP [N <T>reaktioner</T>]] [PRÆP i] [NP [N <T>huden</T>]]]** på grund af gentagen dermal anvendelse af et stof , som bedre klassificeres med R38 »Irriterer **[NP [N <T>huden</T>]]**«. Hvor der
0.0	eller **[NP1 [NP [V_PARTC_PAST gentagen] [N berøring]] [PRÆP af] [NP [N <T>huden</T>]]]** eller slimhinderne . 8 . Sensibiliserende : Stoffer og produkter , som ved indånding eller **[NP1 [NP [N optagelse]] [PRÆP gennem] [NP [N <T>huden</T>]]]** kan fremkalde **[NP [N <T>reaktioner</T>]]** , således at der ved yderligere eksponering af stoffet eller produktet fremkommer karakteristiske symptomer . 9 . Kræftfremkaldende : Stoffer og produkter

Fig. 2. Search results on *reaktion* and *hud* (‘reaction’ and ‘skin’) enhanced from the original query *hudreaktion.*

or over are conceived as retrieved documents. In conclusion, hits where both search words are in the same NP are preferred to hits where they are not in the same NP.

4 Evaluation

4.1 Methodology

As stated in the introduction, the aim of this paper is *not* to evaluate a ‘split compound’ strategy in search, compared to a non-split strategy. In contrast, we aim at finding ways of bringing down the number of irrelevant hits when expanding the query with split compounds.

The manually evaluated material amounts to 410 randomly picked compound queries, resulting in 3.367 hits. However, all in all, only one forth of the evaluated enhanced queries produced hits.

To calculate the proportion of retrieved material that is actually relevant (precision), as well as the proportion of relevant material actually retrieved in answer to a query (recall), we have as mentioned applied the strategy of defining retrieved hits as search results with a score of 10 or more. According to standards, precision is calculated on the basis of the intersection of the relevant hits and the retrieved hits (as mentioned corresponding in our context to the hits that have passed the threshold) divided with the number of retrieved hits. This related to columns E and G in figure 3. Recall on the contrary is calculated on the basis of the intersection of the relevant hits and the retrieved hits, divided with the number of relevant hits, relating to the columns E, G, and H in figure 3.

Relevant hits are defined on a manual judgment basis: if the search result was considered by the evaluator to contain a synonym to the search query, the retrieved document was tagged as relevant; in other words we apply a rather strict relevance

https://doi.org/10.1017/S1351324906004256 Published online by Cambridge University Press

A: compound	B: valency information on head derived from STO	C: for verb-derived compounds: semantic role of non-head	D: number of search results	E: number of retrieved hits (score over 9)	F: number of not retrieved hits (score below 9)	G: number of not relevant in retrieved	H: number of relevant in not retrieved	I: specific characteristics of retrieved	J: reason for not relevant in retrieved	K: reason for relevant in not retrieved	L: precision ((E-G) / E)	M: recall (E-G) / (E+H-G)
acceptinstrument 'accept instrument'			0	0	0	0	0				-	-
adgangsproblem 'admittance problem'	*that*-clause		39	3	36	0	2	Preposition: *omkring*		Complex NP: perfektum participium	1	0,6
administrationsopgørelse 'administration statement'	Deverbal: Di valent, oblique=*over* or *af*	theme	89	1	88	0	2	Preposition: *over*		In the same VP	1	0,3
adsorptionskapacitet 'absorbation capacity'	Genitive or oblique=*på*		0	0	0	0	0				-	-
aerosoldannelse 'aerosol formation'	Deverbal: Divalent, oblique=*af*	theme	6	3	3	0	3	Preposition: *af*		Complex coordinated NP	0,5	0,5

Fig. 3. Excerpt from result table.

criterion (even if synonymy can be difficult enough to judge) which proved the best strategy, since the compound queries were run without context.

4.2 Results on verb-derived compounds

The evaluated test set contains approx. 15% verb-derived compounds (registered as 'deverbal' in column B), and out of these approx. two thirds have the role 'theme' as non-head. For all compounds with 'theme' as non-head, synonymous 'of' constructions were very frequent in the search results, as for *kirkegårdsudvidelse* (`graveyardFUGEextension`): *udvidelsen af kirkegårde* (`extensionTHE of graveyards`). Also several of them display the objective genitive construction: *kirkegårdens udvidelse* (`the graveyards' extension`).

The rest of the group was not quite as systematic regarding synonymous constructions, however, to most of them corresponded synonymous NPs with other prepositions than *af* (of), as for instance in *depotanbringelse* (`depositplacement`): *anbringelse i depot* ('placement in a deposit') corresponding in this case to a locative role of the non-head.

For the whole group of verb-derived compounds the evaluation results were the following: precision: 0.91 and recall: 0.61. Precision is quite satisfactory, only relatively few of the search results that had passed the threshold were irrelevant to the query. Recall, however, needs a closer examination, since a relatively large

https://doi.org/10.1017/S1351324906004256 Published online by Cambridge University Press

number of relevant hits had not passed the threshold. Two factors seem to play a role, of which the first, however, is the most important one:

- errors in the NP analysis: complex NPs are not always recognized by the NP analyzser, and such NPs are relatively frequent in the texts. Especially complex coordinated NPs cause problems; also relative clauses are not treated by the system as part of an NP.
- copula constructions where the subject predicate denotes a property: whole sentences in terms of copula constructions have not been analyzed as relevant by the system even if they are actually synonymous to the query, as in: *badevandskvalitet – badevandet er af en bestemt kvalitet* (`bathwaterFUGEquality` – the bath water is of a certain quality).

Point two apparently contradicts our second hypothesis according to which search words resulting from the splitting of compounds have to be in the same NP. However, these copula constructions are semantically very close to NPs, since we are actually dealing with a subtype of support verb constructions (Gross 1981)[8] where the main element is the noun (*kvalitet*) and not the verb, and this explains why these search results are in several cases relevant.

Point one, in contrast, shows a weakness of the automatic analysis: it is not trivial for a shallow mechanism to recognize complex NPs since several semantic factors play a role, as in the resolution of PP attachment and coordination.

An example of a defective NP analysis where the coordinated items are not registered as being in the same NP, is given in example 11:

11) Original compound: *arveerhvervelse*. Hit on split compound: [NP [N <T> *Erhvervelse* <T>]] *og afståelse af en fordring eller en kontrakt ved gave*, [NP [N <T>*arv*<T>]] *eller arveforskud sidestilles i denne lov med køb henholdsvis salg*.
`inheritanceFUGEacquisition`: [NP [N <T>Acquisition<T>]] and surrender of a demand or a contract in relation to a gift, [NP [N <T>inheritance <T>]] or advancement are placed on an equal footing with buying and selling in this law.

As an additional experiment, we have manually revised the cases where the NP recogniser looks over NP structures, in order to recalculate recall on a correct NP recognition. With such a correction we get a recall of 0.73, which is a clear improvement, and which shows that our hypothesis generally holds quite well for event nominals: Search results where both search words occur in the same NP do to a very large extent contain synonymous constructions to the initial compound and are therefore relevant hits.

[8] For an account on Danish support verb constructions, see Pedersen (1989); for a recent, extensive account on the topic, see Gross & Pontonx (2004).

https://doi.org/10.1017/S1351324906004256 Published online by Cambridge University Press

4.3 Results on other predicative compounds

Also deadjectivals (i.e. compounds where the head is derived from an adjective) and other predicative compounds follow a predictable pattern. They constitute around 10% of the evaluated material. Actually, since the heads in these compounds subcategorize for a specific preposition, they show even better evaluation results than event nominals, namely a precision of 0.94 and a recall of 0.70. The NPs found in the search results are thus very predictable and follow the encodings in STO, as in examples 12 and 13:

12) *bistandspligt – pligt til bistand* (STO encoding on *pligt*: preposition=*til*)
`aidFUGEduty – duty to aid`

13) *tyveririsiko – risiko for tyveri* (STO encoding on *risiko*: preposition=*for*)
`theftrisk – risk of theft`

An additional explanation to the higher recall is that complex coordinated NPs do not occur to the same extent with these compounds, probably because they are not derived from verbs.

4.4 Results on non-predicative compounds

Around three fourths of the evaluated queries are constituted by non-predicative compounds. For one third of these the split query strategy produced no results at all. We see two possible reasons for these results:

- Some compounds are lexicalised to an extent where they simply do not occur in split form, as seems to be the explanation for: *aftrækskanal* (ventilationFUGEshaft), *autovaskehal* (carwashFUGEhall 'car wash'), *beboelseskompleks* (residentialFUGEquarter), and *blodbudding* (bloodpudding). These are mostly compounds referring to concrete entities.
- Other compounds are relatively rare and maybe constructed on the fly in a very specific context. This could be a plausible explanation for the lack of hits for compounds like *bevistema* (evidencetheme), *blandingshonorar* (mixedFUGEfee), and *feriegirosystem* (vacationgirosystem).

For the rest of the non-predicative compounds we get a precision of 0.78 and a recall of 0.52. As can be seen, these results are not as good as for the other two categories. Generally, this group seems to be less predictable and much more heterogeneous. In other words, some queries give very good results, *husholdningsaffald* (householdFUGErubbish) (precision:1.0, recall 0.8), whereas others give very bad results, e.g. *ledelsesstilling* (managingFUGEposition) (precision: 0.33, recall: 0.12).

4.5 Summing up results

In figure 4 we sum up the test results for the three categories of compounds as well as the average precision and recall for all the evaluated categories. For comparison, the figure also contains the precision obtained when applying no threshold, i.e. when

https://doi.org/10.1017/S1351324906004256 Published online by Cambridge University Press

	Precision	Recall
Verb-derived compounds, applying the threshold	0.91	0.61
Other predicative compounds, applying the threshold	0.94	0.70
Non-predicative compounds, applying the threshold	0.78	0.52
All evaluated material, applying the threshold	0.81	0.54
All evaluated material without applying the threshold[9]	0.40	-

Fig. 4. Search results on enhanced compound queries.

searching on split compounds without any further linguistic restrictions. It should be noted again that the relevance criterion applied is very strict; only when a hit is considered to contain a fully synonymous phrase to the original compound, the hit is categorized as relevant.[9]

5 Conclusions and future work

In conclusion, both hypotheses presented in section 1 seem to hold. Some compounds are more likely to have synonymous 'split' counterparts than others, namely compounds where the head takes arguments, as is the case for verb-derived compounds or other predicative compounds. For almost all compounds in these two groups, the split compound strategy produces search results, and they generally follow a predictable pattern relating to the subcategorisation structure of the compound head. In contrast, non-predicative compounds in many cases do not produce search results whatsoever to a split query, and when they do, the number of irrelevant hits is higher. Furthermore, this group proved to be much more heterogeneous and less predictable regarding patterns of context.

These results relate directly to our second hypothesis, according to which search results where both the search words (obtained by splitting the compound) are in the same noun phrase, are more likely to contain a synonymous phrase to the original compound query. Again, the first two groups clearly demonstrate that this is the case: precision is high for these two groups meaning that very few of the documents passing the NP threshold, are irrelevant. Only one sentence construction systematically contradicts our hypothesis, namely the support verb constructions expressed in terms of copula constructions. Documents with such copula constructions containing the compound non-head as the subject and the head as the predicate, and where the head denotes a property, are generally synonymous to the compound. As long as high weight is assigned only to NPs, these examples influence recall negatively; therefore

[9] Precision without a threshold is calculated by taking the intersection of all the originally retrieved hits (column D in figure 3) and the relevant hits (subtracting all the non-relevant hits from the total number of hits) and dividing it by the number of originally retrieved hits.

https://doi.org/10.1017/S1351324906004256 Published online by Cambridge University Press

further research should investigate the possibility of incorporating an analysis of such support verb constructions as possible synonyms to compounds.

Non-predicative compounds show a much more blurred picture, both precision and recall being lower and the results in general being very heterogeneous. One serious defect of the shallow linguistic analysis generally influenced the results negatively: the NP recognition in several cases proved to be unsuitable to the document types which include many juridical texts with very complex NPs. In future work we will experiment with an extended NP recognizer capable of recognizing complex coordinated NPs in order to see if recall can be improved without aggravating precision too much.

In spite of the reservations on the technical level of the applied language tools, the experiments performed above clearly demonstrate that compound splitting combined with a shallow linguistic analysis focusing on subcategorisation information and on the recognition of NPs, can improve search by bringing down the number of irrelevant results. When we compare with the precision obtained where no threshold is applied, we see that all three categories of compounds present much better results when applying a linguistically based threshold.

Another interesting aspect which we would like to go deeper into in future work, is how query expansion to more specific concepts in combination with our NP-based approach can further improve search especially by improving recall. During our experiments, we observed that the non-head of the query compound frequently occurs in a new specialized compound, the document still however being relevant to the original query. Use of a thesaurus or a compound analyzer also on the retrieved documents may have a positive impact on the results. Consider the following examples: *valutaomregning – omregning af lånevaluta* ('currency conversion' – 'conversion of loan currency'), *biblioteksdrift – drift af folkeskolebiblioteker* ('library working' – 'the working of public libraries'),and*besætningsudskiftning – udskiftning af svinebesætning* ('stock renewal' – 'renewal of pig stock'). It is not unlikely that such an expansion in particular would improve search on non-predicative compounds since it seems that lexicalized compounds like for instance *blodbudding* ('blood pudding') is only split when one wants to further specify the non-head, as in *budding af lammeblod* ('pudding of lamb blood').

Acknowledgements

Thanks to Steen Bøhm Andersen, Ankiro, for providing the split compounds list and the corresponding search results, and to Bart Jongejan, University of Copenhagen, for developing and implementing the weighting algorithm. Also thanks to the rest of the VID partners as well as to Patrizia Paggio for comments and suggestions.

References

Abney, S. (1996) Partial parsing via finite-state cascades. *J. Natural Lang. Eng.* **2**(4): 337–344.

Andreasen, T. Jensen, P. A., Nilsson, J. F., Paggio, P., Pedersen, B. S. and Thomsen, H. E. (2004) Content-based text querying with ontological descriptiors. *Database & Knowl. Eng. J.* **48**: 199–219.

https://doi.org/10.1017/S1351324906004256 Published online by Cambridge University Press

Allan, J. (2000) Natural Language Processing for Information Retrieval. *Tutorial notes presented at the NAACL/ANLPLanguage Technology Joint Conference*, Washington.

Braasch, A. and. Pedersen, B. S. (2002) Recent work in the Danish Computational Lexicon Project "STO". *EURALEX Proceedings 2002*, pp. 301–314. Center for Sprogteknologi, København.

Braasch, A. and Olsen, S. (2004) STO: A Danish Lexicon Resource – Ready for Applications. *Proceedings of LREC-2004,* **4**: 1079–1082.

Chen, A. and Gey, F. (2003) Combining Query Translation and Document Translation in Cross Language Retrieval. *Proceedings of CLEF Conference*, pp. 108–121.

Dalianis, H. (2005) Improving search engine retrieval using a compound splitter for Swedish. *Proceedings of Nodalida*, Joensuu, Finland, May 20–21.

de Loupy, C. and El-Bèze, M. (2002) Managing Synonymy and Polysemy in a Document Retrieval System Using WordNet. *Proceedings from the Workshop on Using Semantics for Information retrieval and Filtering*, pp. 20–27. LREC 2002, Las Palmas de Gran Canaria, Spain.

Gonzales J., Verdejo, F., Peters, C. and Calzolari, N. (1998) Applying EuroWordNet to Cross-lingual Text Retrieval. *Computers & the Humanities*, **31**: 185–207.

Grimshaw (1990) *Argument Structure*. MIT Press, Cambridge, London.

Gross, M. (1981) Les bases empiriques de la notion deprédicat sémantique. In: *Formes Syntaxiques et Prédicat Sémantiques*, vol. 63 of *Langages,* pp. 7–52. Larousse, Paris.

Gross, G. and de Pontonx, S. (eds.) (2004) Les verbes supports: nouvel état des lieux. *Lingvisticae Investigationes, Fascicule spécial*, **27**: 2.

Haltrup, D. (2000) NP-genkendelse i OntoQuery. Technical Report, Center for Sprogteknologi, Københavns Universitet.

Johnston, M. and Busa, F. (1999) Qualia structure and the compositional interpretation of compounds. In: E. Viegas (ed.), *Breadth and Depth of Semantics Lexicons*. Dordrecht: Kluwer Academic, pp. 167–187.

Jongejan, B., Pedersen, B. S. and Navarretta, C. (2004) Automatisk analyse af Zaccos og Ankiros tekstmateriale. *VID rapport no. 3*, Center for Sprogteknologi, University of Copenhagen.

Järborg, J. (2003) Semantisk uppmärkning. Metoder, problem og resultat. *Research Reports from the Department of Swedish,* Göteborg Universitet.

Mueller, E. T. (1998) Evaluating Representations of Common Sense. Talk for *AAAI-1998 Panel*, http://www.signiform.com/compoundnoun/classes.htm

Karlgren, J. (2005) Compound terms and their constituent elements in information retrieval In: S. Werner (ed.) *Proceedings from Nodalida*, Joensuu, Finland, pp. 105–11.

Lapata, M. (2002) The Disambiguation of Nominalzations. *Computational Linguistics*, **28**(3): 357–388.

Levi, J. (1978) *The Syntax and Semantics of Complex Nominals*. Academic Press. NY.

Loukachevitch, N. and Dobrov, B. (2004) Ontological Types of Associative Relations in Information-Retrieval Thesauri and Automatic Query Expansion. *Proceedings from Ontolex 2004- Ontologies and Lexical Resources in Distributed Environments*, pp. 24-29. Lisbon.

Ørsnes, B. (1995) *The Derivation and Compounding of Event Nominals in Modern Danish – an HPSG Approach with an Implementation in Prolog*. PhD dissertation, University of Copenhagen.

Paggio, P., Pedersen, P. S. and Haltrup, D. (2003) Applying language technology to ontology-based querying – The OntoQuery Project. *Applied Artificial Intelligence J.*: Artificial Intelligence for Cultural Heritage and Digital Libraries, **17**(8–9): 817–833.

Pedersen, B. and Paggio, P. (2004) The Danish SIMPLE Lexicon and its Application in Content-based Querying. *Nordic J. Linguistics*, **27**(1): 97–127.

Pedersen, S. N. (1989) The Treatment of Support Verbs and Predicative Nouns in Danish. In: J. Pind (ed.), *Proceedings of the Nordic Conference on Computational Linguistics*, pp. 208–218. University of Rejkjavik, Iceland.

https://doi.org/10.1017/S1351324906004256 Published online by Cambridge University Press

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

Smeaton, A. and Quigley, I. (1996) Experiments on Using Semantic Distances between Words in Image Caption Retrieval. *Proceedings of the 19th International Conference on Research and Development in IR*, pp. 174–180. Zurich, Switzerland.

Strzalkowski, T., Gurthrie, L., Karlgren, J., Leistensnider, J., Lin, F., Perez-Carballo, J., Straszheim, T., Wang, J. and Wilding, J. (1996) Natural Language Information Retrieval: TREC-5 Report, pp. 291–314. *The Fifth Text Retrieval Conference*, NIST Special Publication.

Strzalkowski, T., Lin, F. and Perez-Carballo, J. (1997) Natural Language Information Retrieval: TREC-6 Report, *The Sixth Text Retrieval Conference*, pp. 347–366. NIST Special Publication.

Strzalkowski, T., Stein, G., Bowden Wise, G., Perez-Carballo, J., Tapanainen, P., Jarvinen, T., Voutilainen, A. and Karlgren, J. (1998) Natural Language Information Retrieval: TREC-7 Report, *The Seventh Text Retrieval Conference*, pp. 217–226. NIST Special Publication.

Søgaard, A. (2004) A compound matrix. *Proceedings of The 11th International Conference on Head-Driven Phrase Structure Grammar*, pp. 444–456. CSLI Publications, Stanford.

Voorhees, E. (1993) Using WordNet to disambiguate word senses for text retrieval. In: Korfhage, R., Rasmussen, E. and Willett, P. (eds.), *Proceedings of the 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 171–180. Pittsburgh.

Voorhees, E. (1994) Query expansion using lexical-semantic relations. In: W. B. Croft and C. J. van Rijsbergen (eds.), *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 61–69.

https://doi.org/10.1017/S1351324906004256 Published online by Cambridge University Press