Quantifying domain-specificity: the occurrence of financial terms in a general corpus

Gisle Andersen

Unifob AKSIS

University of Bergen

Summary

In this paper, I investigate the transfer of terminology from professional discourse to language in general from a corpus-linguistic point of view. Domain-specificity can be approached from a qualitative and a quantitative angle, and the current paper has its main focus on the former approach. Frequency of use outside the specialist field can be seen as a pivotal indicator of terminology transfer. This factor is a reflection of other underlying factors, such as commercial value and public interest. The paper introduces corpus-based methods for the empirical study of transfer of terminology and presents preliminary results from the domain of business and finance. Specifically, the inventory of the KB-N termbase and that of the Norwegian newspaper corpus (NNC) are compared. It is shown that a common vocabulary can, but need not entail a conceptual overlap of the two datasets. Furthermore, the study focuses on differences between multiword and monoword terms, showing that the former type represents a high degree of ambiguity while the latter type represents meanings more specific to the domain. Moreover, it is shown that low-frequency items common in KB-N and NNC are generally a more reliable indicator of conceptual overlap between the two datasets. Finally, as can be expected, the occurrence of KB-N terminology is considerably higher in the business and finance newspaper than in the other parts of the newspaper corpus.

In any scientific field, the linguistic items which make up its inventory of scientific terms have a potential for being used outside the professional discourse and for entering into the general language. A decade or so ago, the notion of a liquid crystal display was confined to actors in electronics and related businesses/technologies, but now, the LCD screen has become a household name, alongside the device itself. Two years ago, knowledge of subprime lending was confined to banks and financial institutions, while today reports of the subprime crisis are disseminated globally and have also reached our shores, the term being a linguistic innovation and an anglicism in Norwegian.

My current research interest is to investigate the transfer of terminology from professional discourse in academic and scientific contexts to popularised discourse in the general public domain. What are the factors that determine whether or not such transfer takes place? The characteristics of the scientific field in question? Linguistic complexity? (While *DNA* rings a bell to most adult language users, *deoxyribonucleic acid* hardly would). Commercial viability? The answers may be many and complex, but it is clear that the role of *global communication* and the role of the *mass media* should not be underestimated. Nationally and locally, newspapers serve a crucial function in disseminating scientific results that are relevant to their readerships, commercially, technically or otherwise.

The current paper can be seen as an attempt to shed light on one aspect of terminology transfer from professional discourse to language in general. I argue in favour of a corpus-based approach to the above questions, claiming that *frequency of use* outside the specialist

field can be seen as a pivotal indicator of terminology transfer. This factor is a reflection of other underlying factors, such as commercial value, public interest and so on. Rather than arriving at major conclusions, my modest ambition is to introduce corpus-based methods for the empirical study of transfer of terminology, as well as presenting preliminary results from one domain, that of business and finance. More specifically, I investigate the degree to which the inventory of the KB-N termbase is used in a general written corpus, the Norwegian newspaper corpus.

Data and method

The empirical basis of this study consists of a termbase containing terms, definitions and other metadescriptions, and a corpus of written Norwegian. These are described in turn in this section.

The KB-N termbase and corpus

KB-N (Knowledge Bank of Norway) is a result of a three-year project funded by the Research Council of Norway. The resources developed in the project consist of a parallel corpus and a bilingual termbase for economic-administrative domains. It is primarily the termbase which has been consulted in the current study. The termbase is produced on the basis of an English-Norwegian parallel corpus of translated texts from the business and finance domains. Advanced computational term extraction procedures were applied in the construction phase, and the manual work of the terminologist involved checking of machine-retrieved term candidates. A more thorough description can be found in Øvsthus et al. (2005) and on http://mora.rente.nhh.no/projects/kbn/.

The compilation of the KB-N termbase is an ongoing pursuit; therefore, the figures presented in the current section are likely to change. At the date of extraction, the key content of the termbase can be summarised as shown in Table 1.

Description	n	Example
Total entries in bilingual database	8472	
- Entries containing English and Norwegian term	7865	vekst | growth
- Entries containing English term only	286	- | autonomy
- Entries containing Norwegian term only	320	depresiere | -
- Dummy entry containing no term	1	- | -

Table 1: Key figures for KB-N termbase

The termbase consists of 8063 unique Norwegian terms, of which 116 terms occur more than once with different meanings according to subfields.

The Norwegian Newspaper Corpus

The Norwegian Newspaper Corpus (henceforth NNC) is a self-expanding web-based corpus of Norwegian newspaper texts. The collection of this large, dynamic corpus began in 1998 and it is still ongoing. On a daily basis, a web mirroring tool retrieves recently published texts from a set of remote web sites, capturing the entire internet version of ten major

Norwegian newspapers. A set of own-developed tools is used for further processing and annotation of the texts. The system automatically selects the relevant text, ignoring advertisements, navigation menus, metatext, html code and so on. Further, the texts are annotated with word class and other morphosyntactic information. The newspapers in the corpus include both national and regional newspapers:

- Adresseavisen (AA), Trondheim
- Aftenposten (AP), Oslo
- Bergens Tidende (BT), Bergen
- Dagsavisen (DA), Oslo
- Dagbladet (DB), Oslo
- Dagens Næringsliv,(DN) Oslo
- Fædrelandsvennen (FV), Kristiansand
- Nordlys (NL), Tromsø
- Stavanger Aftenblad (SA), Stavanger
- Verdens Gang (VG), Oslo

One of the newspapers, *Dagens næringsliv*, is a business and finance newspaper, while the other nine are more general in content and coverage. As of June 2008, the corpus consists of 656 million words. More specific information about NNC can be found at http://avis.uib.no/.

The basic method used in the current investigation involves machine-based identification of entries in the termbase that also occur in the corpus. This was mainly done by means of computational scripts that were developed for this particular purpose by my colleague Knut Hofland and myself, and to a lesser extent by means of the corpus' web-based search interface.

Terms and domain-specificity

The newspapers in the corpus are, by Norwegian standards, large newspapers that cover a wide range of topics, including news from the business and finance sector. All newspapers report regularly, for example, developments in the stock exchange, foreign currencies, the real estate market and so on, as part of the general account of news events. Given this, we can expect that the corpus vocabulary contains words which are included in the termbase. And, indeed, as will become evident, there is considerable overlap between the linguistic inventory of the KB-N termbase and the linguistic items that occur in NNC.

Observing that items like *bostyrer* (trustee), and *fast eiendom* (real property) occurs in the general corpus, what does this tell us about the domain-specificity of the terms in question? A term which is used consistently with one precise scientific meaning, agreed upon and understood by experts in a particular field, and not used in other contexts is maximally specific to that domain. Occurrences outside the domain, in non-professional contexts, can, but need not, be a signal that terminology transfer from professional discourse to the general language is taking place. Some terms are never used outside their original domain, other terms occur with a popularised meaning that is less precise and less technical than the one used by specialists in the field, yet other terms occur frequently and exclusively in the technical sense agreed upon by specialists.

At the current stage, the corpus-linguistic approach I am proposing involves merely a context-unspecific word-level comparison. Naturally, this only gives a crude picture of the nature of the overlapping content and the degree of domain-specificity of the terms investigated. Due to the general ambiguity of language, a formal overlap in lexical content does not entail a semantic overlap in the concepts denoted by the words in the termbase. Therefore, the overall finding that there is overlap in linguistic material between the two datasets raises more questions than it answers. Do the NNC tokens represent the same *concept* as the KB-N term does? Does the KB-N term represent a more *technical sense* used by specialists, as opposed to the more general journalistic use found in the corpus? Do the corpus tokens represent *lexical ambiguities* that are not found in a well-defined termbase such as KB-N? What about the multi-disciplinary nature of this termbase (Kristiansen 2005) – how can we know which sub-domain is actually manifested in the corpus examples? These issues require considerable attention and cannot be resolved within the confines of a single paper, but they can be seen as interesting topics for further study. At the current stage, what can be concluded is that both qualitative and quantitative considerations are required in order to describe the nature of terminological use in general corpora.

Arguably, then, domain-specificity can be construed as a continuum. Given the presence of one business and finance newspaper (DN) as a constituent in the corpus, it is sensible to distinguish four different groups of words.

Group I *Common vocabulary* of words occurring in the KB-N termbase and in the corpus generally
Group II KB-N terms which occur with a *considerably higher frequency in DN* than in the rest of the corpus
Group III KB-N terms which occur *exclusively in DN* but not in the rest of the corpus
Group IV Words *unique to KB-N* and non-occurrent in the corpus.

The categories represent different degrees of domain-specificity, ranging from low to high.

Searching for terminology in a general corpus

When searching for KB-N terms which occur in NNC, it was practical to distinguish four different categories which were treated separately:

- untruncated monoword terms (mono-lexical terms); e.g. *bostyre* (board of trustees)
  truncated monoword terms; e.g. *bostyre.**
- untruncated multiword terms (poly-lexical terms); e.g. *børsnotert selskap* (listed company)
- truncated multiword terms; e.g. *børsnotert.* selskap.**

Corpus tokens belonging to the first and third categories were retrieved by means of exact (untruncated) pattern matching of Norwegian headwords in the termbase with words in the corpus. Notably, this distinction is not concurrent with the more general, corpus-linguistic disctinction of lemma vs. word form, since some headwords in the termbase are not lemma forms but inflected forms, examples being the plural forms *avviklingsutgifter* (decommissioning costs) and *balanseførte utgifter* (capitalised costs), and the definite form *bokføringsloven* (the Bookkeeping Act). Tokens belonging to the second and fourth categories were retrieved by means of truncated pattern matching. Hence, the number of

tokens retrieved was generally larger than for categories 1 and 3. For an item such as *bostyre.**, the forms retrieved is shown in Table 2 (relative frequencies per billion words).

Word form	abs. freq.	rel. freq.
bostyre	24	37
bostyreadvokater	1	2
bostyreaktører	1	2
bostyrearbeid	2	3
bostyrebehandling	1	2
bostyremedlem	1	2
bostyremedlemmene	1	2
bostyremedlemmer	1	2
bostyremøte	2	3
bostyremøtet	6	9
bostyrene	25	38
bostyrenes	5	8
bostyrer	2224	3402
bostyrerapport	4	6
bostyrerapporten	14	21
bostyrerapporter	2	3
bostyrere	46	70
bostyreren	404	618
bostyrerens	34	52
bostyrereren	2	3
bostyrerjobbene	2	3
bostyrerne	60	92
bostyreroppdrag	4	6
bostyrers	52	80
bostyres	2	3
bostyret	1091	1669
bostyrets	138	211
SUM	4149	6349

Table 2: Result of truncated search bostyre. (alphabetically sorted)*

We note that the truncated search retrieves the headword itself, as well as inflected forms (*bostyrere*, indefinite plural form) and compounds containing the headword as its leftmost component (*bostyrerrapport*, trustee report), and their inflections. A truncated search result of multiword terms is shown in Table 3.

Word form	abs. freq.	rel. freq
børsnotert selskap	805	1231
børsnotert selskape	2	3
børsnotert selskapene	3	5
børsnotert selskaper	2	3
børsnotert selskapet	10	15
børsnotert selskaps	4	6

børsnoterte selskap	163	249
børsnoterte selskapene	566	866
børsnoterte selskapenes	23	35
børsnoterte selskaper	1185	1812
børsnoterte selskaper1	2	3
børsnoterte selskapers	11	17
børsnoterte selskaperÈ	1	2
børsnoterte selskapet	344	526
børsnoterte selskapets	4	6
børsnotertere selskapene	1	2
børsnoterterte selskapene	1	2
SUM	3127	4783

*Table 3: Result of truncated search børsnotert.*selskap.* (alphabetically sorted)*

Note that the retrieval also captures some spelling and formatting errors; these have not been manually rectified or removed.

This simplified retrieval method raises many critical issues, some of which are dealt with below. On the whole, both the truncated and exact matching generally retrieves corpus tokens that are relevant to the task of investigating terminology use in a general corpus. Nevertheless, I have chosen to focus on the results of the exact pattern matching in the following account. This is because it is generally a more reliable method in terms of correct identification of term use, and because the truncated method requires more post-processing than was possible to perform at this time.

Overlapping inventory in KB-N and NNC

The general result of the exact pattern matching of the Norwegian inventory of KB-N with the corpus was as follows:

Category	NNC	KB-N	%
monoword terms, exact matching	3 358	4 909	68.4 %
multiword terms, exact matching	1 293	3 276	39.5 %
TOTAL	4 651	8 185	56.8 %

Table 4: Type frequencies of KB-N headwords in NNC

The table shows that a great amount of the linguistic forms that represent terminological headwords in KB-N are also found in the NNC. We note that more than half (56.8 %) of the headwords were found in the corpus, and that the monoword terms are represented to a higher degree (68.4 %) than the multiword terms (39.5).

So which linguistic forms are overlapping items in the two datasets? In the following, I describe the inventory based on the categorisation into Groups I-IV above.

Group I – Common vocabulary

Looking at the occurrence of monoword and multiword terms, the top of the frequency list of common vocabulary is as shown in Table 5.

Mono-words	abs. freq.	rel. freq.	Multiwords	abs. freq.	rel. freq.
mål	570094	871966	administrerende direktør	45735	69952
bruker	279002	426737	statistisk sentralbyrå	14399	22023
gode	237719	363594	offentlig sektor	9898	15139
penger	230188	352075	privat sektor	4333	6627
feil	126509	193497	økonomisk vekst	4194	6415
næringsliv	94337	144290	varer og tjenester	3042	4653
pris	88425	135247	over streken	3022	4622
kontroll	87596	133979	lover og regler	2664	4075
tilbud	85792	131220	den europeiske sentralbanken	2645	4046
marked	85335	130521	fast eiendom	2528	3867
børs	83942	128390	forskning og utvikling	2160	3304
styre	81778	125080	finansnæringens hovedorganisasjon	2010	3074
økonomi	70497	107826	egne aksjer	1954	2989
kultur	70473	107789	statlig eierskap	1923	2941
behov	70082	107191	offentlig støtte	1690	2585
resultat	64995	99411	økonomisk politikk	1547	2366
utvikling	62512	95613	økonomisk utvikling	1283	1962
eier	62298	95286	flytende rente	1141	1745
bank	60851	93072	rettet emisjon	1128	1725
familie	59768	91416	resultat per aksje	1125	1721
kurs	59032	90290	offentlig tjenestemann	1091	1669
tap	57493	87936	mangel på arbeidskraft	1071	1638
sjekk	55566	84989	full dekning	1055	1614
kjøper	52172	79798	økonomisk krise	1054	1612
selskap	50080	76598	offentlige myndigheter	1028	1572
omsetning	48954	74876	straffbar handling	973	1488
tjenester	48547	74253	politiske myndigheter	960	1468
eiendom	47644	72872	økonomiske konsekvenser	904	1383
handel	47010	71902	høy rente	883	1351
sikkerhet	46901	71736	internasjonal økonomi	870	1331

Table 5: Top of frequency-ranked list of common vocabulary (rel. freq. per billion words)

The leftmost list of *frequent monowords* shows that the common vocabulary represents general words that are not in any way specific to the domains of KB-N. They include general nouns like *mål* (aim, target), *feil* (error), *kontroll* (control), *behov* (need), *resultat* (result), *sjekk* (check) and *sikkerhet* (security). They also include words that are syntactically ambiguous between nominal and verbal readings, including *kjøper* (buyer/buys), *bruker* (user/uses) and *styre* (board/lead) or between nominal and adjectival readings, like *gode* (good (n/aj)). Furthermore, we find multiply semantically ambiguous nouns like *selskap*

(party/company) and *tap* (loss), which also happen to have senses within the business and finance domains. These are lexical items which belong to language in general, and we can assume that many/most of the corpus tokens represent other senses than that those denoted by the technical terms in KB-N. However, the frequent monowords also include words which more prototypically represent this domain, including items like *børs* (stock exchange), *omsetning* (trade/sales) and *handel* (trade). But generally, the frequent monowords that are common in KB-N and NNC cannot be viewed as a reliable indicator of conceptual overlap between the two datasets, since the conceptual information denoted by a technical term like *tap* (loss) is much more specific than the more general denotations of that noun.

On the other hand, the list of *frequent multiwords* is much more consistent in denoting more technical and specific senses within the business and finance domain. They include unambiguous terms like *administrerende direktør* (chief executive officer), *flytende rente* (floating rate of interest) and *rettet emisjon* (private placement). But also among these words we find phrases whose meanings are sensitive to the context of use, and where many corpus tokens presumably represent meanings outside the business and finance domains, such as the metaphorical phrases *over streken* (above the line) and *full dekning* (full coverage). Note that this category also represents proper nouns like *Den europeiske sentralbanken* (the European Central Bank) and *Statistisk sentralbyrå* (Statistics Norway, lower-cased as result of the retrieval process).

The results so far show an interesting difference between multiwords and monowords used as terms in Norwegian, in that the former type represents a high degree of ambiguity while the latter type represents meanings more specific to the domain.

Arguably, low-frequency words are just as interesting when comparing the datasets of KB-N and NNC. Table 6 gives a list of KB-N terms which are hapax words in NNC (occurring only once in the corpus).

Monowords	abs. frq.	rel.frq.	Multiwords	abs.frq.	rel.frq.
toppfirma	1	2	total faktorproduktivitet	1	2
tosatsstruktur	1	2	tradisjonell forretning	1	2
tradingportefølje	1	2	trykt annonse	1	2
transaksjonsbehandling	1	2	tvilsom regnskapsføring	1	2
transaksjonskonto	1	2	tvungent salg	1	2
transaksjonsutgifter	1	2	uavhengig tredjemann	1	2
transisjon	1	2	ubenyttet beløp	1	2
underenhet	1	2	ubenyttet rettighet	1	2
utbyttebeløp	1	2	uforutsett forpliktelse	1	2
utstederselskap	1	2	uforutsett risiko	1	2
utvalgsmetode	1	2	underliggende papir	1	2
utvinningsindustri	1	2	unikt merkenavn	1	2
valutagjeld	1	2	unikt salgsargument	1	2
valutakursstyring	1	2	uoppdagede feil	1	2
valutakurssystem	1	2	uoppsigelig leieavtale	1	2
valutaposisjon	1	2	usikre fordringer	1	2
valutastøtte	1	2	ustabil inflasjon	1	2
vareenhet	1	2	utvidelse i begge retninger	1	2

vekstmirakel	1	2	utøvende revisor	1	2
verdidisiplin	1	2	varer med mindre feil	1	2
verdinettverk	1	2	vekst i markedsandel	1	2
verdipapiroppgjør	1	2	velkjent produkt	1	2
verdipapirregister	1	2	velutviklet økonomi	1	2
verdipapirtilsyn	1	2	vilkårlig tilbud	1	2
verdiskapningsregnskap	1	2	volumvektede gjennomsnittspriser	1	2
verditall	1	2	vurdering av alternativer	1	2
verdsettingsmetode	1	2	økonometrisk analyse	1	2
videresalgspris	1	2	økonomisk disposisjon	1	2
volumbransje	1	2	økonomisk prognose	1	2
volumrabatt	1	2	økonomisk støtteprogram	1	2

Table 6: Bottom of frequency-ranked list of common vocabulary (rel. freq. per billion words)

At this end of the frequency scale, we find more consistently words which represent domain-specific meanings. This applies to both monowords like *tosatsstruktur* (dual rate structure) and *videresalgspris* (resale price) and multiwords like *usikre fordringer* (bad debts) and *utøvende revisor* (practising auditor).

An interesting finding, then, is the observation that low-frequency items common in KB-N and NNC are generally a more reliable indicator of conceptual overlap between the two datasets, since there is good reason to claim that at an item like *økonometrisk analyse* represents identical meanings in the two datasets.

Group II – KB-N terms which are considerably more frequent in DN than rest of NNC

I now turn to the description if the second category of words to be considered, namely those that occur in the business and finance newspaper *Dagens næringsliv*, as well as in the rest of the newspaper corpus, but with a considerably higher frequency in the former subset. This intermediate group is interesting because it represents core vocabulary of the business and finance domains, at the same time as its members show signs of emerging use outside domain-specific settings.

These words were retrieved by considering the relative frequency of each item in DN compared with the relative frequency in the other newspapers. The set consists of words where DN accounted for more than 50 per cent and less than 100 per cent of the use in the corpus. The token frequency of these terms is as shown in Table 7.

Category	NNC-DN	KB-N	%
monoword terms, exact matching	638	4 909	13,0 %
multiword terms, exact matching	224	3 276	6,9 %
TOTAL	862	8 185	10,6 %

Table 7: Type frequencies of KB-N headwords considerably more frequent in DN than rest of NNC

We note that the size of this group is limited compared to the common vocabulary discussed above; this category amounts to about ten per cent of the KB-N headwords. Again, the monowords account for a higher percentage than multiword terms.

The most frequent items in this category are shown in Table 8, which gives relative frequencies and the percentage of tokens that was found in DN.

Monowords	NNC-DN	DN pct.	Multiwords	NNC-DN	DN pct.
børs	1113658	70,8 %	den europeiske sentralbanken	37671	77,6 %
næringsliv	952555	57,7 %	egne aksjer	33240	89,0 %
portefølje	624654	97,6 %	resultat per aksje	21041	92,2 %
eiendom	578595	66,6 %	rettet emisjon	16237	77,3 %
fond	538128	78,7 %	internasjonal økonomi	8160	57,4 %
næring	480554	79,6 %	rentebærende gjeld	3666	61,3 %
råvarer	479602	86,7 %	konkurranseutsatt sektor	3521	57,6 %
fritid	476184	86,6 %	utestående aksjer	3438	83,8 %
annonsering	462060	97,6 %	fusjoner og oppkjøp	2775	67,6 %
driftsresultat	244171	72,4 %	norsk investorforum	2589	53,9 %
aksje	227002	76,9 %	aktiv forvaltning	2568	67,0 %
eierandel	182786	58,3 %	bokført verdi	2527	53,8 %
rentekutt	128940	72,8 %	urealisert tap	2444	79,0 %
bransje	119455	60,6 %	negativ avkastning	2320	67,6 %
analytiker	116204	80,5 %	innenlandsk etterspørsel	2133	75,2 %
børsmelding	110674	85,7 %	høy inflasjon	2092	53,7 %
pant	56124	56,3 %	stram pengepolitikk	2050	79,3 %
prisfall	55751	52,2 %	negativ kontroll	2050	52,3 %
investor	55192	53,9 %	ulovlig innsidehandel	2050	51,9 %
megler	52438	53,2 %	glidende gjennomsnitt	1967	98,4 %
aksjonær	47012	56,2 %	lange renter	1885	83,2 %
nettoresultat	46307	93,6 %	økonomiske utsikter	1781	54,4 %
rentemøte	43781	68,3 %	ekstraordinære kostnader	1698	67,5 %
grunnfondsbevis	43305	82,0 %	privat konsum	1698	64,6 %
inntjening	42393	52,0 %	finansiell stabilitet	1595	52,7 %
sammenstilling	40509	91,9 %	finansiell rådgiver	1553	66,3 %
børsnotering	38148	68,6 %	kortsiktige lån	1450	66,5 %
nasdaq-indeksen	33592	79,7 %	børsnoterte aksjer	1450	55,1 %
nøkkeltall	33198	86,8 %	ekspansiv finanspolitikk	1388	64,0 %
indeks	29532	77,0 %	bli børsnotert	1388	52,6 %

Table 8: Top of frequency-ranked list of words most frequent in DN (rel. freq. per billion words)

The table shows that this category consists almost exclusively of items which are firmly placed in the financial and business domains (*fritid* (leasure time) being one exception). The table also serves to give an overview of the subject matter of the financial newspaper, and of the characteristic inventory which distinguishes DN from the other, more general newspapers.

Group III – KB-N terms which occur only in Dagens næringsliv

I now turn to the description of the third category of KB-N words, namely those that occur exclusively in *Dagens næringsliv* and not in any other subset of NNC. These words are interesting in that they constitute an intermediate category of core vocabulary which is not strictly confined to professional discourse between specialists, but they appear in the discourse of a sector-specific newspaper. But at the same time they are not general enough to be used in the more mainstream newspapers. The token frequency of these terms is shown in Table 9.

Category	NNC-DN	KB-N	%
monoword terms, exact matching	179	4 909	3,6 %
multiword terms, exact matching	85	3 276	2,6 %
TOTAL	264	8 185	3,2 %

Table 9: Type frequencies of KB-N headwords which occur exclusively in DN

The table shows that this is an even smaller group of words, accounting for only 3.2 per cent of the KB-N headwords. Again, the monoword terms account for a higher percentage than the multiword terms.

The frequency-ranked list of DN-unique words in the corpus is shown in Table 10.

Monowords	NNC-DN	Multiwords	NNC-DN
energiindeks	1346	varige konsumgoder	1056
produktmarked	1180	geografisk marked	518
lagerøkning	497	diskontert kontantstrøm	331
etableringshinder	476	innenlandsk kreditt	269
lagernivå	352	ekstraordinær post	207
konsumenttillit	331	funksjonell valuta	207
kupongrente	331	ikke-varige goder	186
trendvekst	311	effektiv avkastning	166
markedsutslag	290	nominell stabilitet	145
kreditteksponering	269	illikvid aksje	124
mindreavkastning	269	uro i valutamarkedet	124
insuffisiens	228	eksisterende kunde	104
kvotedisiplin	228	monetær økonomi	83
basispunkt	207	syklisk oppgang	83
kredittforening	207	msci world-indeksen	62
kpi-vekst	186	negativ goodwill	62
konkurssannsynlighet	124	økonomisk indikator	62
nynotering	124	ekstern forvalter	41
petrobufferporteføljen	124	faktisk volatilitet	41
tegningsrettsemisjon	124	innenlandsk kostnadsvekst	41
tilbudsforstyrrelse	124	innenlandsk prisvekst	41
honorarinntekt	104	internasjonalt aksjemarked	41
inntektseffekt	104	opprinnelige kostpris	41
konsernintern	104	samlet konsum	41

statskasseveksel	104	skattemessige avsetninger	41
innkjøpsordre	83	spekulativ boble	41
konkurransekursindeks	83	strukturelt underskudd	41
rapporteringsperiode	83	virkemidler i pengepolitikken	41
sluttbrukermarked	83	aktiv indeksforvaltning	21
tidsverdi	83	analyse av konkurranse	21

Table 10: Top of frequency-ranked list of words that occur only in DN (rel. freq. per billion words)

We note that this category consists of words with a much lower overall frequency than the previously discussed categories. Also, the items in this group clearly represent domain-specific concepts.

Group IV – KB-N terms non-occurrent in the newspaper corpus

Finally, I look at those headwords in KB-N that do not occur in the newspaper corpus, and therefore show no tendency of transfer from professional to more general discourse. It is not my intention to study these in detail. The overall statistics for this category is shown in Table 11.

Category	not in NNC	KB-N total	%
monoword terms	1 551	4 909	31.6 %
multiword terms	1 983	3 276	60.5 %
TOTAL	3 534	8 185	43.2 %

Table 11: KB-N headwords that are non-occurring in NNC

We observe that the majority of multiword terms in KB-N and about a third of the monowords are not represented in the general newspaper corpus. This means that these items constitute a core vocabulary of terms that are restricted to the domain-specific discourse. A randomised selection of words in this category is shown below.

> abandonering, balansedag, clearingmedlem, dagsrenter, ekvivalensenheter, flankestrategi, indeksendring, kampanjemedier, ledelsesmetoden, markedsføringsenhet, opsjonsprisingsmetode, produktutviklingsprosess, reklameraid, satsforskjell, totalgruppe, varelagerfinansiering, årsregnskapsperiode.

Given these findings, the members of this fourth set can be considered maximally domain-specific terms within the business and finance domain.

Distribution across newspapers

So far, I have only compared DN as opposed to the other newspapers in NNC. It is also of interest to consider the general distribution across the ten newspapers in the corpus, to see if other interesting differences emerge. The overall picture can be seen in Figures 1 and 2.

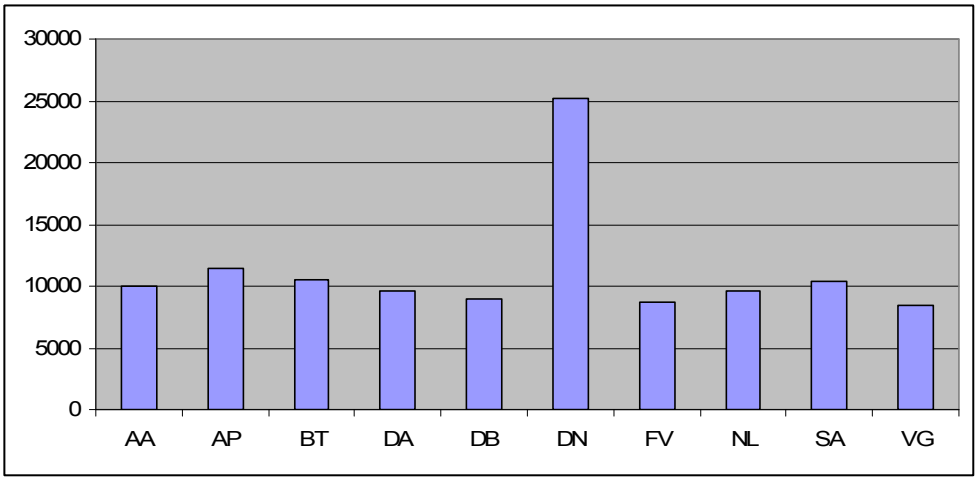

Figure 1: Distribution of untruncated monoword terms in the newspaper corpus (rel. freq. per million words)

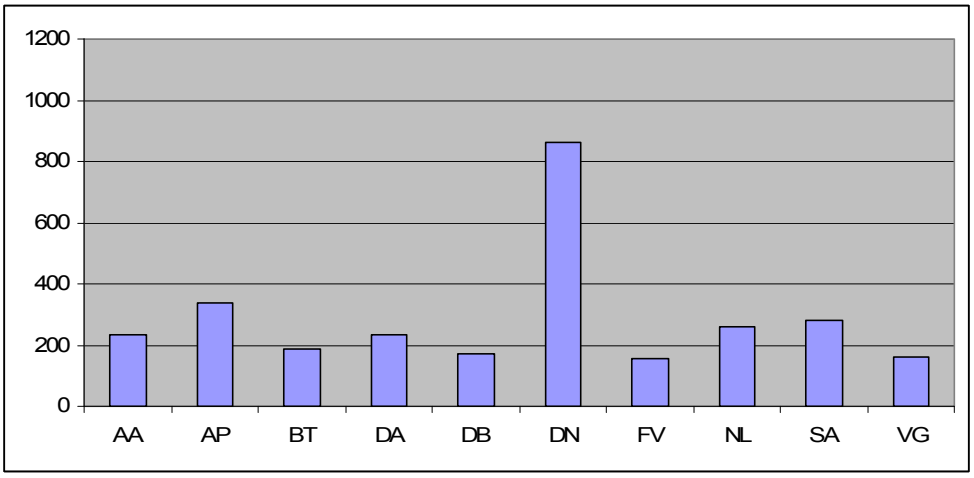

Figure 2: Distribution of untruncated multiword terms in general corpus (rel. freq. per million words)

As can be expected, the occurrence of KB-N terminology is considerably higher in DN than in the other newspapers. We observe that the headwords occur in all the other newspapers with remarkably similar frequencies; the frequency of monoword terms ranges from 8 500 to 11 400 tokens per million words, as opposed to 25 000 tokens per million words in DN. The variability is greater for the multiword terms, which range from 208 to 456 tokens per million words, as opposed to 1 134 tokens per million words in DN. For both term types, it can be seen that the prestigious and conservative *Aftenposten* has the second highest proportion of terms, while the tabloids *Dagbladet* and *Verdens Gang* have the lowest scores. Although not shown in any of the figures, these differences tend to increase if we consider the figures drawn from the truncated search. Then, monoword terms occur with a frequency in DN of 62 000 tokens as opposed to an average of 24 000 tokens per million words, while multiword terms have a frequency in DN of 1134 tokens as opposed to an average of 372 tokens per million words.

Examples of use

Having focused mainly on quantitative aspects so far, I now briefly discuss qualitative aspects of this empirical approach to terminology.

In principle, the occurrence of a termbase item in a corpus can signal reference to a concept with the same properties as denoted by a termbase entry. However, this need not be the case. Although clearcut cases exist, it may be difficult for the analyst to judge whether the author of the newspaper article has the professional skills and communicative intentions of such a kind that the use of the term in a non-scientific context represents the same concept as it does for the person whom the termbase was developed for (or by). Clearcut cases would, of course, be proper nouns like *Den europeiske sentralbanken* (the European Central Bank), where any author would be taken to refer to the one unique referent denoted by the term. But in many other cases, careful consideration is required in order to conclude that the authentic examples in a general, non-academic/non-technical corpus represent identical concepts as those denoted by the terms in a termbase.

There is therefore a legitimate need to go beyond the observations that can be extracted from word frequency lists and look into actual corpus usage, to see if a given term in corpus usage actually represents the meaning denoted in the termbase. This requires a large-scale investigation, but only a brief check of individual forms has been carried out so far. In the following, I examine some examples of more or less arbitrarily chosen terms from the extracted frequency lists above, with a view to comparing their use in the NNC corpus with the meanings recorded in the KB-N termbase. The aim is to get an impression of the extent to which the examples found in the corpus are relevant examples of the meanings associated with the termbase entries. I start with the low-frequency items.

As shown above, a term such as *velkjent produkt* (well-known product) is a low-frequency multiword term which is represented once in the corpus.

> [1] Hedda Heyerdahl Braathen, prisbelønt designer og medeier i designbyrået accendo produktutvikling, har valgt å rette søkelyset mot et **velkjent produkt** i bybildet. (DB010909)

We note from the example that it is taken from an extract describing business and product development. Hence, the corpus token concerns subject matter that, at least superficially, belongs to the KB-N domain, and there appears to be good reason to argue that the termbase entry and the corpus token denote the same concept in this case.

The term *tosatsstruktur* (dual rate structure) is a low-frequency monoword term which occurs once in the corpus.

> [2] Utvalget foreslår at en på noe sikt legger om skattleggingen til et prinsipp om verdsettelse til markedsverdi med et bunnfradrag, men foreslår at en strammer inn noen ordninger også på kort sikt. Forenkling av fradragsstruktur for lavinntekter: Lønnsfradraget oppheves. **Tosatsstruktur** i minstefradraget: Nedre grense om lag 20 000 kroner. (DB030206)

This corpus extract is directly reporting departmental plans for new taxation regulations. It is clear that the corpus token is from the same domain and represents the same concept as in the termbase.

A high frequency multiword term such as *over streken* (above the line, 3022 tokens), on the other hand, has a much vaguer meaning and is metaphorically constructed. In KB-N, this is defined as follows:

> [3] A rather vague term meaning that part of the profit and loss account/income statement above the measure of earnings on which earnings per share are based. Extraordinary items are normally not above the line. Contrasted with below the line. (KB-N)

It is clear that tokens of *over streken* in the corpus do not represent this specific sense, but the more general metaphorical sense of 'going too far' or 'commiting an unacceptable act'. In addition, it commonly denotes a physical sense of (especially) the football crossing the goal line. But it also occurs in some domain-specific contexts:

> [4] Nasdaq faller noe tilbake etter et kraftig åpningshopp, men er fortsatt **over streken** med en oppgang på 1,25 prosent. (DN010315)

But even in domain-specific settings, one cannot be sure that it represents the sense denoted by the definition.

The term *rettet emisjon* (private placement) is a high-frequency multiword with a more specific sense.

> [5] Den amerikanske PC-produsenten Dell kjøper 5 prosent av den norske Internett-kometen Fast Search & Transfer, gjennom en **rettet emisjon** som priser Fast til over 5 milliarder kroner. (AP990803)

In fact, all 1128 tokens of this multiword appear to represent uses from business news reports, and the corpus examples generally fit well with the senses represented in the KB-N termbase (sales/trade/trading/turnover).

A high-frequency monoword like *resultat* (result) has a vague meaning. As can be expected, the corpus data include many tokens outside the domain, as well as tokens where the word is used in the sense denoted by the KB-N definition:.

> [6] Selskapet har et negativt **resultat** på 10,1 millioner kroner. (DB0915)

On the other hand, a high-frequency monoword like *bostyrer* (trustee) is an example of a term with a specific sense relating to the business and finance sector.

> [7] Dagestad er **bostyrer** i Kristoffersens personlige konkursbo. (AP0217)

Here there is no reason to argue that the corpus examples represent other senses than those in the KB-N termbase.

Concluding remarks

In sum, the examples show that one must exercise caution when using corpus data in exploring degrees of domain-specificity, as much linguistic material with term status may serve many other functions in the general language. Nevertheless, I have argued that a corpus-based investigation of term usage it is nevertheless relevant in order to shed light on

the degree to which domain-specific terms can be observed in actual use in authentic datasets like a newspaper corpus.

We have seen, first, that low-frequency words are more liable to a domain-specific interpretation than high-frequency words. Second, multiword items are more liable to a domain-specific interpretation than monoword items. Third, we have seen that term material can represent very varying degrees of ambiguity, as there is also a major divide between, on the one hand, words like *resultat* (result) with a general and sometimes metaphorical sense with many alternative senses outside the domain, and, on the other hand, items like *bostyrer* (trustee), with a specific sense which seem consistently to denote the termbase concept, even in contexts outside the professional discourse.

References

Kristiansen, Marita (2005) Disciplinary autonomy and concept relations in electronic knowledge bases. A theoretical approach to KB-N −a knowledge base for economic-administrative domains. *SYNAPS − Fagspråk. Kommunikasjon. Kulturkunnskap,* 17(2005). 1-7.

Kunnskapsbank for norsk økonomisk-administrativt fagdomene. http://mora.rente.nhh.no/projects/kbn/.

Norsk aviskorpus. http://avis.uib.no/.

Øvsthus, K[ari]/ Innselset, K[ai] / Brekke, M[agnar] /Kristiansen, M[arita] (2005) Developing Automatic Term Extraction. Automatic Domain Specific Term Extraction for Norwegian. In: Madsen, Bodil Nistrup / Thomsen, Hanne Erdmann (eds.). *Terminology and Content Development. TKE 2005. 7th International Conference on Terminology and Knowledge Engineering*. [s.l.]: Litera. 337-348.