

# Determinologized Usage Patterns as a Filter in Web-Based Term Extraction Systems

Jakob Halskov (jh.id@cbs.dk)

Dept. of Computational Linguistics, Copenhagen Business School

## Abstract

The paper describes how identifying recurrent usage patterns of terms in non-specialized contexts (determinologized usage) can act as part of a filtering device and increase the precision of term extraction systems using the Internet as a corpus. A pilot study compares salient collocations with the mother term *server* in four corpora containing texts on Information Technology of various levels of specialization. A number of strong collocations with *server*, which are exclusive to a large corpus of newswire, are identified as cases of determinologized usage, and it is shown that these collocations are stable through-out an eight-year time frame. It is speculated that the presence of determinologized usage patterns (DUPs), in addition to density measurements of knowledge patterns and knowledge rich contexts, may be a reliable indicator of the usefulness of a text to a terminologist.

## 1 Finding the right texts

Automating the compilation of specialized corpora by using the Internet has a number of obvious advantages. Keeping termbases up-to-date is an arduous task, especially for domains like Information Technology (IT), which are characterized by rapid term growth. As has been demonstrated in [2] bootstrapping specialized corpora and extracting terms from the Internet is a speedy means of achieving this goal. The web, however, is a very dirty and multifarious collection of texts. While the documents returned by the search engines may contain a number of the terms

specified in a query, the documents are ranked not by their usefulness to terminologists but by pragmatic principles like popularity ranking [3]. Even when documents are ranked purely by the vector-based similarity of the query with page content, there is no guarantee that termhood assumptions about strings, which hold for neat corpora of highly specialized discourse, are true in other cases. When large numbers of non-specialists start to use terms from a domain (the process known as determinologization), they form strong collocations, which, formally speaking, may resemble terms, while not functioning as such.

The problem of assessing the knowledge richness of a document can be approached in various ways. [14] has demonstrated how recurrent phrases can be used as text-type discriminators, and [1] outline a corpus-building web application for terminologists which reorders the query results returned by search engines based on the density of so-called knowledge patterns (KP) and knowledge-rich contexts (KRC). KPs indicate possible semantic relations between terms (for example *is a* or *is a kind of*) and KRCs are cases where a KP is surrounded by lexical units which have already been identified as terms. The authors demonstrate that reordering the URLs returned by Google via assessments of KP and KRC densities results in the retrieval of a corpus which is more knowledge-rich than a baseline corpus in both cases.

While the KP/KRC method filters out overly specialized texts where most semantic relations between domain specific concepts are implicitly stated, the method will conceivably overlook texts which are too general in the sense that they contain determinologized usage in the form of semantically vague col-

locations representing imprecise, or even, incorrect knowledge about the target domain. The following sections outline a first attempt at further enhancing a KP/KRC filter with a negative weighting scheme for cases of determinologized usage in web documents.

## 2 Determinologization

It is not surprising that conceptual fuzziness tends to occur when non-specialists use terminology in non-specialized communicative contexts. It seems intuitive that what is a term (representing a clear-cut concept) to one person may be a (possibly unknown) word representing a fuzzy category to another person who lacks the required specialist knowledge to decode the term fully and correctly. It also seems probable that traces of this conceptual fuzziness can be registered in linguistic usage.

This phenomenon, known as determinologization, has been defined as "the ways in which terminological usage and meaning can 'loosen' when a term captures the interest of the general public" [10, p.12], and the semantic changes caused by determinologization have been grouped into two types:

1. the original terminological sense is by and large preserved
2. the original terminological sense is diluted [9, p.202]

This distinction between preservation and dilution of a terminological concept largely corresponds to the distinction between sense modulation and sense selection in lexical semantics [5], but due to the standardization bias in terminology few studies have explored how the meaning of terms evolves [15] and how communicative context affects termhood [4, 11, 7]. Determinologized usage of the second type is much rarer than the first type and will be near impossible to detect by purely statistical means, so the findings outlined in the pilot study will pertain only to Meyer's first category.

## 3 Identifying determinologized usage patterns

The hypothesis of the pilot study presented in section 4 is that instances of determinologized usage can be identified by computing and comparing lexical profiles of mother terms in corpora representing discourse at various points on the scale from non-specialized to highly specialized. While the lexical profiles, or wordsketches [8], needed in lexicography are complex and involve grammatical relations between collocates, for instance SUBJ\_OF, OBJ\_OF, terminologists are primarily interested in noun phrases, and these can be examined by computing the association strength between positional co-occurrence pairs such as the list of word forms occurring to the immediate left of a mother term. The statistical approach to collocational analysis has been well researched both with a view to computational lexicography [13] and, more recently, to computational terminology [6].

### 3.1 Data and methodology

The domain of Information Technology (IT) was selected as a testing ground for the pilot study presented in this paper. With the possible exceptions of medicine and appliances, IT is a unique domain in that it has not merely captured the interest of the general public, but has even become an indispensable part of everyday life both in the workplace and at home. At the same time IT is a technical subject field characterized by concepts which require a high degree of determinacy, and the combination of these two factors make it ideal for an investigation such as the present.

In order to sample text representing IT related discourse at various levels of specialization, the following four corpora were compiled.

1. Patent applications filed to the US Patent and Trademark Office<sup>1</sup> under USPC class 709<sup>2</sup>

<sup>1</sup><http://www.uspto.gov>

<sup>2</sup>Electrical computers and digital processing systems: multicomputer data transferring

2. Proceedings of Conferences in Research and Practice in Information Technology<sup>3</sup>
3. The popular science journal, PcPlus<sup>4</sup>
4. New York Times newswire<sup>5</sup>

As summarized in table 1 the four corpora represent three different communicative settings, and although they are of vastly different sizes, the absolute number of occurrences of the mother term to be investigated in section 4 are comparable.

Intuitively one would expect to find examples of determinologized usage only in contexts where non-experts communicate with each other, ie. in the corpus of the *New York Times* newswire in the present case. However, our interpretation of co-occurrence patterns of mother terms in this corpus should acknowledge two basic facts:

1. since terms are specific to a domain we must expect their frequency of occurrence to be relatively low outside the domain in question<sup>6</sup>.
2. when lexical units, which function as terms in specialized discourse (e.g. bus, server, driver), occur in non-specialized contexts, the most frequent senses are likely to be the non-specialized ones.

The first problem is overcome by using a very large corpus, but the second problem is trickier. Ideally, the mother terms should be disambiguated (sense tagged) prior to the computation of co-occurrence patterns and extraction of salient collocations. Even unsupervised word sense disambiguation (WSD) is labor-intensive, however, and as evidenced by the pilot study in section 4 analysis of unannotated corpus data nevertheless yields interesting results.

## 4 Pilot study: *server*

Tables 2 and 3 list the top ten left and right collocates of the mother term *server* which are *exclusive*

<sup>3</sup><http://crpit.com/>

<sup>4</sup><http://www.pcplus.co.uk>

<sup>5</sup><http://ldc.upenn.edu>

<sup>6</sup>seven per million running words in the case of “server(s)” in the *New York Times*

to each of the four corpora. The collocation candidates are identified by extracting bigram data from the four raw corpora using the Ngram Statistics Package [12] and subsequently computing the statistical significance of the co-occurrence pairs using an implementation of Fisher’s exact test from Stefan Evert’s Utilities for Cooccurrence Statistics toolkit<sup>7</sup>. A significance cut-off point of  $-\log(p) > 6$  and a joint frequency threshold of  $f > 10$  are enforced to omit insignificant co-occurrences and one-off coinings from the comparison.

Disregarding the strongest exclusive left collocate of *server* in the newswire corpus (which is part of a repetitive phrase in the meta text headers of the corpus), the three collocations, *computer servers*, *internet servers* and *network servers* strike the eye. The collocation of *computer* with *server* does not represent a more specific, subordinate concept to that represented by the mother term. It is a semantically void domain label whose only purpose is to highlight the IT context of *server* and distinguish it from other contexts and senses (e.g. *tennis server*, *altar server*, *food server*) which readers who are not immersed in the IT domain might confuse it with. It is not surprising that such labels are not needed in domain specific texts. The collocations *internet server* and *network server* on the other hand are infrequent newswire variants of the term *web server* which ranks second in the list of left collocates common to all the corpora in table 4.

Further down the list we find other strong collocations like *pc server(s)*, *powerful server(s)* and *centralized server(s)* which are exclusive to the newswire corpus. While the *pc* in *pc server(s)* arguably functions as a domain label on par with the *computer* in *computer server(s)*, the collocates *powerful* and *centralized* result in fuzzy categories which are of no use to terminologists building a domain ontology. They can, however, be used to filter out text types which are too general and should not be included in a specialized corpus.

In order to assess whether the determinologized usage patterns (DUPs) identified in table 2 are truly valid also in a diachronic perspective, the newswire

<sup>7</sup><http://www.collocations.de>

Table 1: IT related corpora

	patents	popular science	proceedings	newspapers
tokens	2M	6.7M	6M	1.1Bn
no. textual units	162 applications	48 issues	1040 papers	?
time frame	2001-2004	2000-2004	2001-2005	1994-2002
communicative setting	expert-expert	expert-intermediate	expert-expert	intermediate-novice
“server(s)” in total	7457	4439	1682	7330
“server(s)” per million	3729	663	280	7

Table 2: exclusive left collocates of “server”

-log(p)	NEWSWIRE	-log(p)	PATENTS	-log(p)	POP. SCIENCE	-log(p)	PROC.
2967	cox _	235	TM _	345	SQL _	37	capability _
652	sql _	206	secondary _	189	FTP _	32	Pounamu _
481	computer _s	189	primary _	125	Exchange _	28	agent _
302	internet _s	176	directory _	121	your _		
302	network _s	87	management _	117	Active _		
211	unix _s	77	home _	108	POP3 _		
105	shadow _	65	License _	86	Commerce _		
102	computer _	63	producer _	79	2000 _		
99	universal _	60	said _	75	Windows _		
98	cyrano _	58	streaming _	66	DNS _s		
...	...	...	...	...	...		

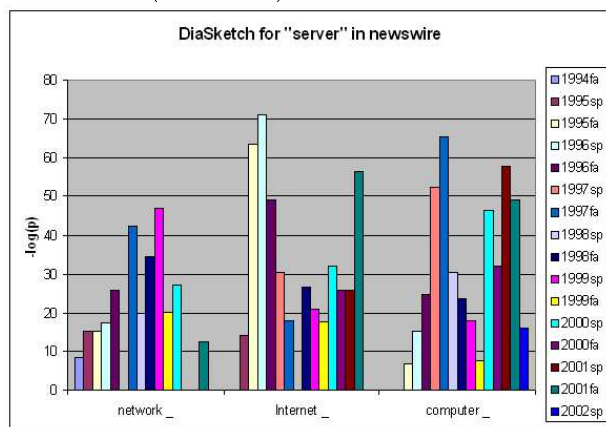
Table 3: exclusive right collocates of “server”

-log(p)	NEWSWIRE	-log(p)	PATENTS	-log(p)	POP. SCIENCE	-log(p)	PROC.
145	_ systems	264	_ blade	183	_ 2003	32	_ site
124	_ business	221	_ node	113	_ 2000	23	_ sites
116	_ farms	90	_ blades	20	_ running	10	_ model
95	_ computing	69	_ advertising	13	_ on	7	_ has
85	_ products	61	_ advertisement	12	_ name	6	_ which
74	_ networks	39	_ transmission	12	_ itself		
64	_ wan	30	_ push	12	_ will		
58	_s running	29	_ apparatus	11	_ configuration		
54	_s computers	26	_ sync	10	_ without		
40	_ product	24	_ process	9	_ or		
...	...	...	...	...	...		

Table 4: left collocates of “server” common to all corpora

collocation	$-\log(p)$ combined	$-\log(p)$ in newswire	$-\log(p)$ in pop. science	$-\log(p)$ in patents
client server	1677	1607	54	16
web server	1626	600	410	616
proxy server	662	88	94	480
web servers	592	334	91	168
mail server	542	203	288	51
application server	465	118	12	335
file server	283	165	11	107
central server	230	119	67	45
application servers	84	24	15	45
remote server	72	20	42	10
database server	59	28	20	11

Figure 1: DiaSketch of selected collocates of “server” in newswire (1994-2002)



corpus was split into 16 segments (each of six months’ duration) and what we might call a DiaSketch (Diachronic wordSketch) was computed for three of the main DUPs and visualized as figure 1. The diagram clearly shows that the three collocates, or DUPs, are strong through-out the eight-year time frame, although *network server(s)* seems to be waning.

## 5 Conclusions

Based on the findings from the simple pilot study of the mother term *server*, it seems plausible that DUPs like *computer server(s)*, *pc server(s)* and *powerful server(s)* may identify overly general texts which are not suitable for terminology work such as term extraction and ontology building. As internal text-linguistic features, DUPs would combine well with measurements of KP and KRC densities as conducted in [1]. While genre specific collocates can be used to filter out non-specialized newswire text, they can also be used to identify various specialized text types like patent applications (*said server* and *server apparatus*) and popular science (*your server*). Finally, a list of mother term collocates which are common to all<sup>8</sup> corpora, specialized as well as non-specialized (cf. table 4) may provide a terminologist with a good overview of the fundamental subordinate concepts of the one represented by the mother term.

The main sources of error in the pilot study are the following:

1. no lemmatization
2. no WSD of the non-specialized newswire corpus
3. no dispersion measure

<sup>8</sup>the corpus of conference proceedings was left out of this analysis because it has the lowest number of occurrences of “server(s)”

The computation of collocational strength on raw word forms rather than lemmas is both an advantage and a disadvantage. Clearly, collocations like *computer server(s)* in table 2 and *web server(s)* in table 4 would rank even higher if the singular and plural forms were conflated, but on the other hand it is an interesting fact that the plural forms are much more predominant in the non-specialized newswire text than in the specialized text types. Presumably, the explanation is that non-specialized text does not discuss the properties of an individual server but talks about the tendencies and effects of IT upon society in general. As for the WSD issue, it is conceivable that part of the strength of the collocation *powerful server*, for instance, stems from other, non-IT senses of server, but a quick glance at newswire concordances for server reassures us that the frequency of the IT sense far and away exceeds the other possible senses. Although a dispersion measure might adjust the figures slightly, the size and composition of the corpora should render the effect negligible.

## 6 Further work and perspectives

In order to further expand the list of DUPs discovered in the pilot study for a subfield of the IT domain, newswire bigram data containing other mother terms could be produced and analyzed. Although a computationally expensive operation, adding newswire trigrams to the analysis will identify even more DUPs. Once a comprehensive DUP catalogue has been compiled, its value as a filtering mechanism can be evaluated by compiling a URL list of documents containing IT-related search words and running the TerminoWeb application outlined in [1] on this corpus with and without the DUP filter.

## References

- [1] Akakpo Agbago and Caroline Barrière. Corpus construction for terminology. In *Proceedings of Corpus Linguistics 2005*, 2005.
- [2] Marco Baroni. Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*, 2004.
- [3] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th World Wide Web Conference*, 1998.
- [4] María Teresa Cabré. Elements for a theory of terminology: Towards an alternative paradigm. *Terminology*, 6(1):35–57, 2000.
- [5] D. A. Cruse. *Lexical semantics*. Cambridge University Press, 1986.
- [6] Lee Gillam, Mariam Tariq, and Khurshid Ahmad. Terminology and the construction of ontology. *Terminology*, 11:1:55–81, 2005.
- [7] Kyo Kageura. *The Dynamics of Terminology - a descriptive theory of term formation and terminological growth*. John Benjamins, 2002.
- [8] Adam Kilgarriff. Word sketch: Extraction and display of significant collocations for lexicography. In *Proceedings of ACL 2001*, pages 32–38, 2001.
- [9] Ingrid Meyer. *L'étirement du sens terminologique: aperçu du phénomène de la détermination*, chapter Le Sens en Terminologie, pages 198–217. 2000.
- [10] Ingrid Meyer. When terms move into our everyday lives: An overview of de-terminologization. *Terminology*, 6(1):111–138, 2000.
- [11] Jennifer Pearson. *Terms in Context*. John Benjamins, 1998.
- [12] Ted Pedersen and Satanjeev Banerjee. The design, implementation and use of the ngram statistics package. In *Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics*, 2003.
- [13] Frank Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1), 1993.

- [14] Michael Stubbs and Isabel Barth. Using recurrent phrases as text-type discriminators: A quantitative method and some findings. *Functions of Language*, 10:1:61–104, 2003.
- [15] Rita Temmerman. *Towards New Ways of Terminology Description*. John Benjamins, 2000.