CHAPTER 7 EXPLORING KEYWORDS

WordSmith has a function that identifies *keywords*. Keywords are words which are significantly more frequent in one corpus than in another (Hunston 2002: 68). The WordSmith manual describes keywords as follows:

> Keywords are those whose frequency is unusually high in comparison with some norm. Keywords provide a useful way to characterise a text or a genre. Keywords usually give a reasonably good clue to what the text is about. Potential applications include: language teaching, forensic linguistics, stylistics, content analysis, *text retrieval*. (our emphasis).

By virtue of its domain-specificity, a keyword is therefore a unigram term, i.e., a term comprising only one word. Keywords are always established through corpus-evidence.

In this section, we explore the potential of keywords to reduce the gap between the termbases and the corpora, that is, reduce the number of termbase terms that are infrequent in the corpus, and increase the number of termbase terms that are frequent in the corpus.

7.1 Potential significance and related research

The potential of keywords to indicate domain-specificity is acknowledged in the literature. McEnery et al describe a keyword as "a word that is more frequent in a text or corpus under study, the analysis corpus, than it is in some (larger) reference corpus, where the difference in frequency is statistically significant" (2012: 245). For Baker, keywords are a "measure of saliency" (2006: 125) of the corpus, of its uniqueness when compared to general language (p. 147).

Research has also been carried out investigating the potential of keywords for term extraction. Drouin (2003) developed a hybrid term extraction method using keywords and proved its effectiveness, particularly for identifying domain-specific single-word terms, using a reference corpus and analysis corpora that are significantly smaller than those used in the current research. What are called keywords in WordSmith and the current research are

called *domain-specific lexical units* by Drouin. Drouin applies a certain frequency threshold to the keywords, and further restricts them to nouns and adjectives[81]; the final retained units which are subsequently used to find MWTs are called *specialised lexical pivots* (SLP). Drouin then attempts to apply an SLP restriction to all the words in a retained MWT. This proves to be overly restrictive, and he concludes by recommending the use of a simple raw frequency threshold for identifying MWTs from candidates that are extracted based on the SLP as the search term (or pivot).

Chung (2003) also used the corpus-comparison approach for single-word term extraction, using an anatomy text as the domain corpus. Anatomy, as it turns out, has a higher proportion of single-word terms than other domains, so this approach if successful could be very effective in this domain. She concludes that this method is a reasonably simple, valid, and practical way of identifying (single-word) terms, which is also reliable and easy to replicate (p. 242). The main weakness, she acknowledges, is that it does not identify MWTs. This is the challenge that we explore using concordances.

In later research, Drouin et al (2005) refined his previous work, this time comparing two different approaches: use of the general reference corpus and use of a larger domain specific corpus, for identifying domain-specific single-word terms based on relative frequency. They conclude that a balance of the two approaches produces the best results.

Kit and Liu (2008) conducted research in single-word termhood using a similar approach. Based on the observation that a true term is more peculiar to its own subject field than to a general domain, they calculate the termhood of mono-term candidates by their rank difference in a domain-specific corpus and a reference corpus (p.211). This is essentially how WordSmith calculates keywords. For evaluation purposes, they tested and validated the method using a legal corpus and the BNC as the reference corpus, and they used an existing dictionary of legal terms as the gold standard against which they compared the output of the ranked terms. Their experiments achieved a high precision rate. They also discovered that combining keyword ranking with raw frequency improves the identification of corpus-

81 Drouin used part-of-speech tagged corpora

specific terms, compared to using the keyword ranking alone. As future work, Kit and Liu envision extending the corpus-comparison methodology as applied to mono-word termhood measurement to facilitate automatic extraction of MWTs, focusing on the role of highly-ranked mono-word terms in forming MWTs (p. 222).

Anick proposed the "lexical dispersion hypothesis": "a word's lexical dispersion -- the number of different compounds that a word appears in within a given document set -- can be used as a diagnostic for automatically identifying key concepts of that document set" (2001: 34). He developed and tested an algorithm for computing lexical dispersion which was successfully used for the purpose of interactive refinement of search queries. His approach therefore leveraged keywords to identify domain-specific MWTs.

Bowker and Pearson recommend producing concordances of keywords as a means of finding multi-word term candidates (2002: 150). This is precisely what we aim to explore.

The difference between the past research and the current approach is that we are not performing automatic term extraction (ATE), but are investigating the potential of keywords as an aid in human term identification. With today's highly performant search tools and concordancing software, a terminologist can very quickly identify productive MWTs if provided with reliable domain-specific keywords, at least, that is our hypothesis. Given the availability of such technologies, any productivity-related benefits of automating subsequent stages of the selection process, i.e., those stages that are performed by ATE tools that leverage keywords, may be minimal, and furthermore, outweighed by reduced precision and recall compared to human evaluation. Of course, this is purely speculative until an empirical study is conducted to determine the effectiveness of mining multi-word terms from keyword-based concordances versus addressing the noise and silence of automatically-extracted term candidates. The current research aims to demonstrate the feasibility of the former, not to establish its superiority over ATE.

In our study, we are interested in explaining the gap between the company's termbase and the corresponding corpus. If keywords reveal what the text is about, then quite likely they

are very significant terms for a termbase. We are interested in keywords as potential terms themselves (unigrams) and as building blocks for MWTs (bigrams and trigrams). Determining how well keywords are represented in the termbase, and how prominent they are in the corpus, might help explain the gap between the two.

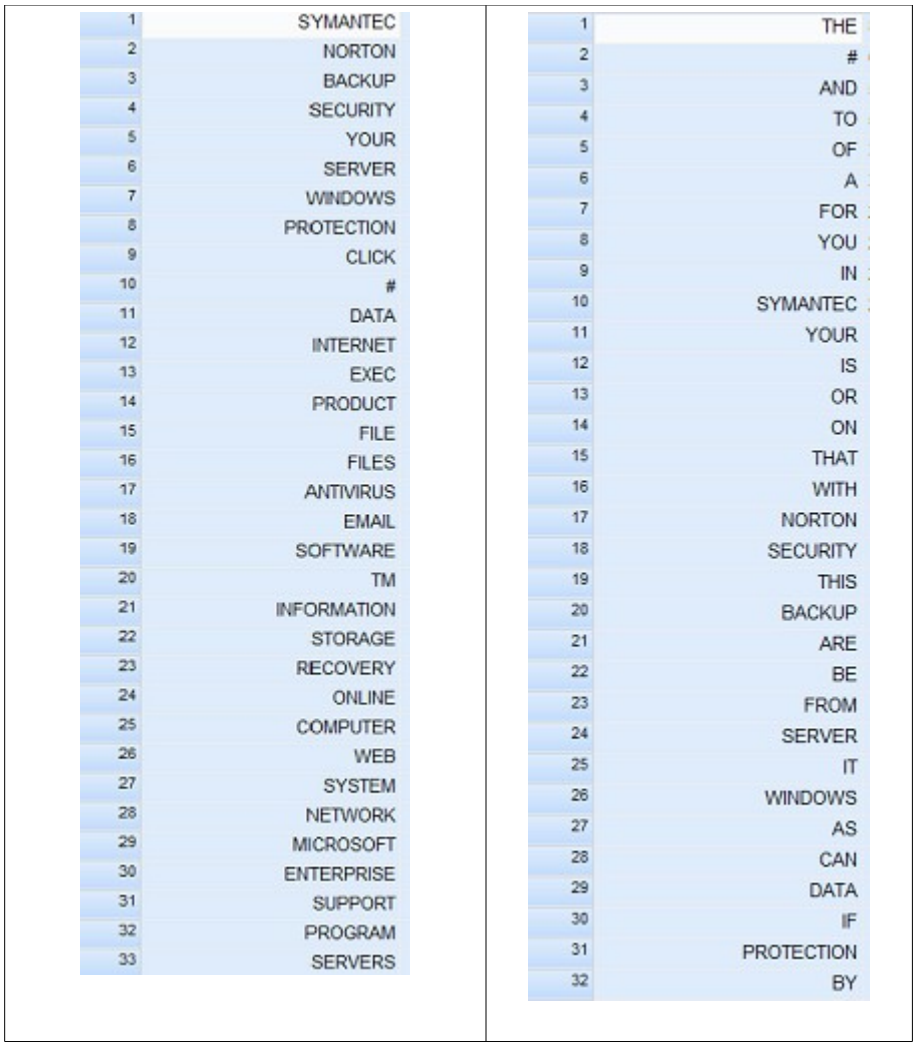
7.2 Keyword identification

The keyword function in WordSmith calculates a keyness measure by comparing a word list from the corpus under study (the so-called small word list) to a word list from a reference corpus (the so-called large word list). The keywords are calculated by comparing the relative frequency of the same words in the two lists. More specifically, the keyness of a lexical item is computed[82] based on:

- its frequency in the small word list
- the number of running words in the corpus under study
- its frequency in the large word list
- the number of running words in the reference corpus.

The following screen captures show the difference between a keyword list and a simple word frequency list, generated from the Symantec corpus:

82 The statistical calculation uses the chi-square test of significance and the Log Likelihood test.



1	SYMANTEC
2	NORTON
3	BACKUP
4	SECURITY
5	YOUR
6	SERVER
7	WINDOWS
8	PROTECTION
9	CLICK
10	#
11	DATA
12	INTERNET
13	EXEC
14	PRODUCT
15	FILE
16	FILES
17	ANTIVIRUS
18	EMAIL
19	SOFTWARE
20	TM
21	INFORMATION
22	STORAGE
23	RECOVERY
24	ONLINE
25	COMPUTER
26	WEB
27	SYSTEM
28	NETWORK
29	MICROSOFT
30	ENTERPRISE
31	SUPPORT
32	PROGRAM
33	SERVERS

1	THE
2	#
3	AND
4	TO
5	OF
6	A
7	FOR
8	YOU
9	IN
10	SYMANTEC
11	YOUR
12	IS
13	OR
14	ON
15	THAT
16	WITH
17	NORTON
18	SECURITY
19	THIS
20	BACKUP
21	ARE
22	BE
23	FROM
24	SERVER
25	IT
26	WINDOWS
27	AS
28	CAN
29	DATA
30	IF
31	PROTECTION
32	BY

Figure 24: Keywords vs words

To identify keywords, we first produced a word list for each of the three corpora. We then obtained a word list for the British National Corpus (BCN) and for the American National Corpus (ANC)[83]. The BNC word list contains 512,588 words (types), and the ANC word list contains 155,329 words, quite a difference in size. Since all three companies in our study are American-based, it would make most sense to use the ANC word list in our key-word calculation, although we felt that, statistically, regional differences in the two Englishes are unlikely to have a significant impact on the result. Nevertheless, to ensure that the size of the word list would not impact the results, we decided to compute keywords using both lists separately and compare the results.

83 These lists are provided with WordSmith as standard reference word lists for the KeyWord function.

As expected, there was no significant difference in the keywords produced using either the BNC or the ANC word list. The same keywords were produced in each case, the only difference being an occasional repositioning of a keyword in the list, for instance, in the ANC output a keyword might appear in position five and in the BNC in position six. There were few cases of this and the ranking proximity meant that the overall keyword list in the high-ranking region was identical. For this reason, we show only the results using the ANC reference word list in the following sections.

When considering keywords it is necessary to use sets that are statistically comparable between the three companies. For instance, if we decide to study the top ten keywords for Minitab, how many keywords do we need to study for the other two companies to ensure comparable results? This factor is determined based on corpus size, using the smallest corpus as a baseline.

	Corpus size	**Factor**	**Number of keywords to study**
Minitab	3,973,265	1	10
SAS	22,136,564	5.57	56
Symantec	19,808,928	5.01	50

Table 59: Normalised sets of keywords for investigation

The following screen capture shows the top keywords for Minitab in WordSmith:

N	Key word	Freq.	%	Texts	RC. Freq.	RC. %	Keyness
1	ENDOFTEXT	26,126	0.66	100	0		81,234.55
2	#	240,845	6.06	100	442,359	2.99	74,483.44
3	DATA	38,525	0.97	100	13,054	0.09	67,688.80
4	MINITAB	17,133	0.43	100	4		53,168.58
5	CHOOSE	16,081	0.40	100	824		43,773.83
6	PLOT	15,339	0.39	100	682		42,344.53
7	CLICK	14,491	0.36	99	2,139	0.01	33,274.94
8	VALUE	15,550	0.39	100	3,385	0.02	32,142.53
9	ENTER	10,996	0.28	100	561		29,934.82
10	CHART	10,123	0.25	96	173		29,768.83
11	MODEL	15,049	0.38	100	4,122	0.03	28,757.17
12	USE	19,951	0.50	100	10,052	0.07	28,509.96
13	VALUES	13,801	0.35	100	3,152	0.02	28,090.30
14	SAMPLE	12,694	0.32	100	2,341	0.02	27,542.72
15	DESIGN	12,268	0.31	99	2,435	0.02	26,068.34
16	TEST	14,125	0.36	100	4,831	0.03	24,659.36
17	COLUMN	10,603	0.27	100	1,485	0.01	24,633.70
18	K	11,549	0.29	97	2,469	0.02	23,998.52
19	PROCESS	12,553	0.32	100	3,439	0.02	23,980.47
20	GRAPH	8,100	0.20	99	128		23,897.60
21	DISTRIBUTION	10,903	0.27	100	2,084	0.01	23,417.62
22	VARIABLES	9,138	0.23	100	807		23,165.37
23	DIALOG	7,485	0.19	100	11		23,085.33
24	BOX	9,745	0.25	100	1,369		22,623.30
25	EACH	18,577	0.47	100	12,502	0.08	21,773.51
26	VARIABLE	8,038	0.20	100	719		20,333.65
27	EXAMPLE	12,657	0.32	100	5,537	0.04	19,589.63
28	REGRESSION	7,596	0.19	98	649		19,353.99
29	RESIDUALS	6,236	0.16	86	32		18,978.76
30	FACTORS	6,088	0.15	98	1		18,886.42
31	RESPONSE	10,802	0.27	99	3,748	0.03	18,731.85
32	WORKSHEET	5,824	0.15	98	10		17,942.64
33	NUMBER	13,909	0.35	100	8,690	0.06	17,226.17

Figure 25: Minitab keywords

To choose ten keywords for our study, we need to consider the nature of the above candidates and eliminate several that are not suitable. For instance, the top keyword, *endoftext*, is a string of text that we had artificially inserted into the corpus as a marker to assist in the merging of files during the corpus preparation process. The character # is not particularly interesting as a keyword, nor is *Minitab*, as the name of the company. The candidates *choose*, *click*, *enter*, and *use* are predominantly verbs; we considered verbs in Section 6.6. At the moment we wish to focus on nouns since they make up the vast majority of terms in

any termbase. Note also that some of these keywords may have a very general meaning, such as *value* and *process*, and as such may not be productive in forming terminological compounds, i.e. MWTs having a domain-specific meaning. Note also that many of them may occur as both nouns *and* verbs. We will check these possibilities using concordances and if necessary we may adjust our keyword selection, looking further down the list for more domain-specific keywords such as *distribution* (position 21), *regression* (position 28), and *residuals* (position 29).

7.3 Keyword categorisation

WordSmith produces the top 500 keywords based on the *keyness* calculation. We decided to look at the keywords in three groups:

1. Top-ranking keywords by keyness
2. Mid- and low-ranking keywords by keyness
3. Keywords that do not occur, or are extremely rare, in the reference corpus

The top-ranking keywords are clearly of interest as these words are estimated to be the most representative of the corpus, statistically speaking. In the mid- and low- keyness range, we will find words that occur to a certain degree in both the company corpus and the reference corpus; the question to answer is if they have the same meaning. If not, we may observe cases of terminologisation, where a term is created by adding a domain-specific meaning to an existing (general) word. Keywords that do not occur, or are extremely rare in the reference corpus, are likely to evoke some distinctiveness relating to the company corpus. However, they may not occur frequently enough in the company corpus to obtain a high keyness value.

7.4 Frequency of keywords versus frequency of termbase terms

By virtue of their method of calculation, keywords occur relatively frequently in the corpus. It was interesting to note, however, how the frequency of keywords compares to the frequency of termbase terms. For this purpose, we reproduce below the table from Section

6.2.1.2 showing the average normalised frequency of the termbase terms.

	Minitab	SAS	Symantec
Number of termbase terms	1,777	4,195	6,441
Number of occurrences of the termbase terms in the corpus	355,072	2,283,004	2,368,703
Average frequency of terms	199.82	544.22	367.75
Normalisation factor	1.006	0.181	0.202
Average frequency, normalised	201.01	98.50	74.29

As this table shows, Minitab's termbase is most closely aligned to its corpus, based on the average occurrence of its terms, followed by SAS and then Symantec. The following table shows this same information for the keywords:

	Minitab	SAS	Symantec
Number of keywords	10	56	50
Number of occurrences of the keywords in the corpus	156,829	2,570,264	1,869,640
Average per keyword	15,682	45,897	37,392
Corpus normalisation factor	1	0.18	0.2
Normalised average	15,682	8,261	7,478

Table 60: Frequency of keywords in the corpus

The fact that the average occurrence of keywords in Minitab's corpus is approximately double that of the other two companies is potentially explained by the smaller number of keywords; the resulting list of keywords favours keywords in the very top range, some of which are quite general in meaning, as previously noted.

These two tables support our hypothesis that keywords may help to close the gap between termbases and corpora. On average they occur much more frequently than the termbase terms. By their keyness measure, the likelihood that they convey domain-specific meanings, at least for many of their contextual uses, is high. If we can focus on those keywords that are most productive in generating domain-specific MWTs, we will make further

progress. That task involves identifying the collocates of the keywords. We begin with the keywords that are under-represented in the termbases.

7.5 Keywords that are under-represented in the termbases

Identifying keywords that are under-represented in the termbase may be useful because we can examine the concordances of these keywords first. This may lead to the discovery of keywords that are justified in being under-represented in the termbase because they are not domain-specific or they are too general in meaning. It may also lead to discovery of keywords that are productive in forming MWTs and therefore should be in the termbase.

To determine whether a keyword is under-represented in the termbase, we calculate the number of occurrences in the corpus of the termbase terms containing the keyword, and we compare that number to the number of occurrences of the keyword alone. For instance, for Minitab, the keyword *data* occurs 53 times in the termbase in the form of various MWTs such as *frequency data* and *data point*. These 53 terms occur 5,373 times in the corpus. However, the keyword *data* occurs 38,525 terms in the corpus. This large difference suggests that productive MWTs containing *data* are missing from the termbase. While the frequency of the keyword, as a unigram, will always be larger than the frequency of any set of n-grams containing this keyword (in our case, the termbase terms containing the keyword), the larger the gap between the two figures, the more likely that the termbase is missing frequently-occurring n-grams containing the keyword. Note that in performing this calculation, if the keyword exists in the termbase as a unigram, it must be removed from the list of termbase terms before counting their frequencies. Otherwise, the frequency count will include *all* n-grams containing the keyword, even MWTs that are not in the termbase. For instance, *plot* exists in the Minitab termbase as a unigram as well as in 83 MWTs. It is necessary to remove *plot* from the list of termbase terms containing *plot* in the Minitab corpus, otherwise MWTs containing *plot* that are not in the termbase will also be counted.

In the following sections, we perform a case insensitive search of the termbase terms in the corpus, since, as the keyword itself is in lower case (and it would not be justified to capi-

talise it), using a case sensitive search of the keyword would produce unreliable results. Because of this, the actual frequencies of the termbase terms in the corpus are slightly lower than indicated, meaning that the gap between the termbase terms and the keywords indicated in the following tables is slightly larger than indicated.

7.5.1 Top-ranking keywords

7.5.1.1 Minitab

The ten keywords we chose to examine for Minitab are shown below. In this analysis, we include 459 terms marked with a constrained usage value in the set of termbase terms that we use for counting purposes, since these terms are permissible in the corpus even though their usage may be restricted to certain contexts.

Keyword	Keyword frequency	Termbase terms with keyword	Frequency of termbase terms	Termbase freq. as % of keyword frequency	Termbase terms in range A or absent	% of termbase terms in Range A or absent
data	38,525	53	5,373	**13.95**	15	28.30
plot	15,339	92	12,232	79.74	34	36.96
value	15,550	47	6,692	43.04	15	34.04
chart	10,123	52	8,365	82.63	12	23.07
model	15,049	30	2,813	**18.69**	10	33.33
sample	12,694	20	5,912	46.57	2	10.00
design	12,268	60	7,352	59.93	24	40.00
test	14,125	76	7,270	51.47	24	31.58
column	10,603	8	144	**1.36**	2	25.00
process	12,553	17	1,686	**13.43**	6	35.29
Total	**156,829**	**455**	**57,839**	**average: 41.08**	**144**	**average: 29.76**

Table 61: Top-ranking keywords for Minitab

Seven keywords (70 percent) of our sample are in the termbase as unigrams: *data*, *plot*, *value*, *sample, design, column* and *process*. Terms based on the keywords *data*, *model*, *column*, and *process*, present the largest gaps between the termbase and the corpus (the

cells with the grey background). These keywords may therefore have greater potential than others as nodes of important MWTs that are missing from the termbase. Furthermore, some of the MWTs containing these nodes that are already in the termbase may be infrequent in the corpus. Running a concordance on these keywords will identify high-frequency terms that are missing from the termbase, and low-frequency terms that are documented in the termbase and therefore may be of limited value.

On the other hand, the keywords *plot* and *chart* are comparatively well represented in the termbase and therefore a concordance search on these keywords is less likely to lead to as many new productive terms. However, note that 37 percent of the terms containing *plot* are in the low frequency range, meaning that this set of terms, while it already contains important members, also likely contains a significant amount of redundancy.

The 455 termbase terms containing these keywords constitute 20 percent of the 2,236 terms in the termbase (1,777 corpus valid plus 459 constrained), but their frequency amounts to less than 15 percent of the total frequency of all termbase terms in the corpus (401,390). Further, 32 percent of the termbase terms are infrequent (144/455). These findings suggest that the keyword-based termbase terms are not optimised, i.e., too many of them are in the low frequency range, and that a better selection could be made of terms containing *data.*

7.5.1.2 SAS

The 56 keywords we examined for SAS are shown in the table below. Keywords that evoke characteristically domain-specific concepts include *syntax*, *metadata*, *plot* and *node*.

Keyword	Keyword frequency	Termbase terms with keyword	Frequency of termbase terms	Termbase freq. as % of keyword frequency	Termbase terms in range A or absent	% of termbase terms in Range A or absent
data	257,330	204	150,795	58.60	104	50.98
statement	165,489	17	2,848	**1.72**	6	35.29
option	115,452	15	20,029	17.35	8	53.33
variable	78,923	107	30,752	38.96	45	42.06

Keyword	Keyword frequency	Termbase terms with keyword	Frequency of termbase terms	Termbase freq. as % of keyword frequency	Termbase terms in range A or absent	% of termbase terms in Range A or absent
value	97,607	47	12,758	13.07	24	51.06
set	107,959	47	95,533	88.49	18	38.30
page	99,728	23	789	**0.79**	**21**	**91.30**
output	75,564	21	6,914	9.15	7	33.33
procedure	63,804	13	986	**1.55**	**8**	**61.54**
model	72,788	58	8,405	11.55	36	62.07
table	74,384	85	17,148	23.05	56	65.88
syntax	38,808	5	197	**0.51**	**3**	**60.00**
server	40,301	69	21,475	53.29	34	49.28
type	54,098	31	5,619	10.39	14	45.16
input	36,051	18	2,750	7.63	8	44.44
function	37,329	32	1,737	**4.65**	**21**	**65.63**
format	30,353	34	3,433	11.31	21	61.76
method	36,775	41	2,753	7.49	32	78.05
metadata	23,112	28	10,926	47.27	15	53.57
parameter	21,673	11	1,057	**4.88**	**6**	**54.55**
view	30,104	20	3,539	11.76	8	40.00
plot	23,603	35	4,744	20.10	23	65.71
log	21,419	14	6,508	30.38	6	42.86
argument	15,040	3	1,538	10.23	1	33.33
attribute	19,335	19	2,534	13.11	10	52.63
analysis	32,009	25	5,366	16.76	15	60.00
box	28,317	12	20,691	73.07	4	33.33
macro	17,490	28	8,685	49.66	12	42.86
node	18,465	23	1,251	6.77	16	69.57
graphics	19,979	17	13,831	69.23	8	47.06
class	33,644	11	3,030	9.01	4	36.36
character	25,672	36	7,856	30.60	22	61.11
dialog	15,571	1	13,460	86.44	0	0.00
graph	17,165	9	994	5.79	4	44.44
matrix	16,834	7	433	**2.57**	**4**	**57.14**

Keyword	Keyword frequency	Termbase terms with keyword	Frequency of termbase terms	Termbase freq. as % of keyword frequency	Termbase terms in range A or absent	% of termbase terms in Range A or absent
error	19,466	17	3,952	20.30	7	41.18
properties	19,619	4	230	**1.17**	2	50.00
command	19,488	22	2,928	15.02	15	68.18
display	21,344	3	82	**0.38**	**3**	**100.00**
label	16,649	11	1,165	7.00	4	36.36
code	19,510	30	3,892	19.95	22	73.33
catalog	12,355	14	5,132	41.54	7	50.00
chart	15,107	45	4,561	30.19	32	71.11
reference	21,184	18	2,840	13.41	10	55.56
object	19,470	36	4,824	24.78	23	63.89
regression	13,354	9	2,984	22.35	5	55.56
statistics	15,990	2	206	**1.29**	1	50.00
axis	13,255	25	3,037	22.91	12	48.00
entry	16,431	36	6,325	38.49	18	50.00
menu	13,635	10	4,691	34.40	6	60.00
library	18,872	33	7,057	37.39	15	45.45
system	37,912	40	18,906	49.87	27	67.50
viewer	10,603	1	430	**4.06**	0	0.00
string	13,527	12	4,409	32.59	7	58.33
default	54,596	7	173	**0.32**	5	71.43
numeric	21,409	8	4,862	22.71	3	37.50
Total	**2,275,951**	**1,550**	**574,050**	**average: 22.98**	**848**	**average: 52.27**

Table 62: Top-ranking keywords for SAS

Eight keywords (14 percent) are in the termbase as unigrams: *attribute*, *box*, *class*, *graph*, *label*, *object*, *program*, and *viewer*. The 1,550 termbase terms containing these keywords constitute 37 percent of the 4,195 terms in the termbase, and their frequency represents 25 percent of the total frequency of all termbase terms in the corpus (2,283,004, see Section 7.4). This gap (37 versus 25) is proportionally similar to that of Minitab. Further, 55 percent

of the keyword-based termbase terms are infrequent (848/1,550). These findings suggest that this set of termbase terms is also under-optimised.

7.5.1.3 Symantec

The 50 keywords we examined for Symantec are shown below. Some are very specific to Symantec's industry (computer security), such as *antivirus*, *ghost*, *firewall*, and *threat*.

Keyword	Keyword frequency	Termbase terms with keyword	Frequency of termbase terms	Termbase freq. as % of keyword frequency	Termbase terms in range A or absent	% of termbase terms in Range A or absent
backup	108,484	206	94,950	87.52	123	59.71
security	112,406	226	75,410	67.09	140	61.95
server	82,299	258	44,911	54.57	182	70.54
protection	76,461	158	42,963	56.19	98	62.03
desktop	12,767	23	3,759	29.44	18	78.26
file	49,457	97	10,990	22.22	69	71.13
antivirus	42,973	93	37,220	86.61	66	70.97
software	52,466	73	9,050	17.25	53	72.60
email	42,785	49	4,733	11.06	35	71.43
storage	39,597	94	17,091	43.16	56	59.57
recovery	40,437	77	24,499	60.59	42	54.55
internet	48,035	47	23,318	48.54	30	63.83
computer	48,657	14	1,573	**3.23**	**9**	64.28
system	64,745	97	26,697	41.23	63	64.95
support	52,982	39	14,656	27.66	19	48.72
enterprise	33,727	108	19,248	57.07	80	74.07
management	46,603	181	23,273	49.94	128	70.72
data	77,793	89	25,072	32.23	61	68.54
network	38,134	105	11,936	31.30	64	60.95
disk	29,447	90	17,194	58.39	64	71.11
service	46,118	263	27,621	59.89	205	77.95
online	33,527	66	13,425	40.04	37	56.06

Keyword	Keyword frequency	Termbase terms with keyword	Frequency of termbase terms	Termbase freq. as % of keyword frequency	Termbase terms in range A or absent	% of termbase terms in Range A or absent
business	43,034	21	4,080	9.48	12	57.14
media	35,244	43	16,306	46.27	22	51.16
restore	24,804	31	3,616	14.58	18	58.06
remote	24,250	66	11,953	49.29	37	56.06
web	31,115	67	14,854	47.74	49	73.13
access	31,625	56	7,559	23.90	45	80.36
pc	23,143	38	5,151	22.26	28	73.68
installation	22,053	22	1,108	**5.02**	**13**	59.09
threat	22,825	35	3,481	15.25	24	68.58
solution	26,683	83	9,289	34.81	57	68.67
download	20,369	13	1,158	**5.69**	**5**	38.46
agent	23,388	195	15,598	66.69	160	82.05
virus	24,286	40	7,135	29.38	25	62.50
option	21,354	35	5,674	26.57	20	57.14
partner	21,050	27	8,875	42.16	13	48.15
customer	20,571	8	3,691	17.94	5	62.50
product	27,821	31	7,122	25.60	19	61.29
license	19,370	38	10,460	54.00	24	63.16
subscription	18,455	7	562	**3.05**	**3**	42.86
update	18,069	29	3,758	20.80	18	62.07
client	19,894	98	8,458	42.52	79	80.61
ghost	17,801	46	10,147	57.00	27	58.70
application	21,253	60	1,808	**8.51**	**51**	85.00
firewall	15,878	42	9,774	61.56	30	71.43
folder	14,253	12	1,085	7.61	8	66.67
device	16,427	49	3,125	19.02	40	81.63
protect	18,662	9	1,789	9.59	7	77.78
program	32,090	46	10,412	32.45	32	69.57
Total	**1,835,667**	**3,700**	**757,617**	**average: 35.68**	**2,513**	**average: 65.42**

Table 63: Top-ranking keywords for Symantec

Ten keywords (20 percent of our sample) are in the termbase as unigrams: *backup*, *desktop, download*, *file*, *internet*, *network*, *online*, *restore*, *threat*, and *virus*. The 3,700 termbase terms containing these keywords constitute 57 percent of the 6,441 terms in the termbase, and their frequency in the corpus is 34 percent of the total frequency of all termbase terms in the corpus (2,237,761). The figure of 57 percent suggests that Symantec's termbase overall is more keyword based than the other two (22 and 37 percent). However, one has to keep in mind that the Symantec termbase is more heavily laden with product names, and some of the most frequent keywords, such as *antivirus*, *protection*, and *server* are found in product names. The difference (57 versus 34) is proportionally similar to that of the other two companies, also suggesting that a large number of these terms occur infrequently or not at all. Sixty-eight percent of the keyword-based termbase terms are infrequent (2,513/ 3,700). Replacing such under-optimised terms with more productive ones would vastly improve the correspondence between the corpus and the termbase.

7.5.1.4 Summary

Seventy percent of the high-ranking keywords are present in Minitab's termbase as unigrams, whereas for Symantec and SAS this percentage is much lower: 20 and 14 respectively. If, as we believe, these keywords are highly productive in forming MWTs that occur frequently in the corpus, having them in the termbase as unigrams would be very beneficial particularly for the company's translation activities. We will attempt to empirically validate this assumption later through corpus evidence.

The frequency of the keyword-based termbase terms as a percentage of the frequency of the corresponding keywords is an indicator of the level of keyword-based correspondence between the termbase and the corpus; the closer the termbase terms come to the keywords with respect to corpus frequency, the greater the number of important and frequent terms based on the keywords in the termbase. In this regard, Minitab performs the best, with the termbase terms accounting for 42 percent of the keyword occurrences on average, followed by Symantec with 36 percent and SAS with 23 percent. (As previously noted, Symantec's higher percentage compared to SAS could be attributed to a bias towards product names.)

Groups of keyword-based termbase terms that occur the least frequently in the corpus have been marked in bold. These terms will be subject to further analysis in Section 7.7.1 where we will identify better collocates by using raw statistics and the Dice relationship measure.

7.5.2 Mid- and low-ranking keywords

Keywords in the mid- and low-ranking area by keyness measure occur in greater proportion in the company corpus compared to the reference corpus, but the difference is not as great as for the high-ranking keywords. We are interested in discovering whether there are terms in this group that are polysemous, having a domain-specific meaning in the company corpus that differs from the meaning in the reference corpus. In this sense, they may be valid terms just as high-ranking keywords are, but their lower keyness is explained by semantic differences between the company corpus and the reference corpus. For instance, the terms *worm* and *cloud* likely have different meanings in Symantec compared to the reference corpus. Indeed, previous research confirms that borrowing words from the general lexicon is a frequent method of neologisation in the computing field (Lam Kam-mei 2001: xi), giving rise to such terms as *bookmark*, *folder*, and *mouse.* In his corpus-based study of computer science texts, Tong Sai-tao observed that words which lie in the middle frequency range are general vocabulary items which take on a specialised meaning (1993). We consider several examples of these types of terms in the following sections.

7.5.2.1 Minitab

Keyword	Frequency of the keyword	Termbase terms with keyword	Frequency of the termbase terms	Termbase frequency as % of keyword frequency
bar	1,646	4	926	56.26
smoothing	576	10	331	**57.47**

Table 64: Mid- and low-ranking keywords for Minitab

The word *smoothing* is not present in the termbase as a unigram. The word *bar* is present,

marked as a general lexicon word. They are both present in the form of several MWTs, such as *resistant smoothing* and *bar chart*. When we have a higher number of termbase terms without a corresponding increase in correspondence rate, such as in the case of *smoothing*, we may suspect that some of the termbase terms are infrequent, and some important ones are missing. Later, we will test this assumption through examining the concordances.

7.5.2.2 SAS

Keyword	Frequency of the keyword	Termbase terms with keyword	Frequency of the termbase terms	Termbase frequency as % of keyword frequency
cube	6,897	4	1,572	22.79
wizard	7,265	2	878	12.09
bar	9,307	12	4,174	45.85
workspace	4,634	6	546	11.78
filter	6,896	8	87	**1.26**
block	8,095	20	1,185	14.64
scatter	4,524	2	1,947	43.04
key	9,857	31	2,174	22.06
miner	2,865	0	0	0.00
locale	2,412	2	69	**2.86**
companion	3,030	0	0	0.00
Total	**65,782**		**12,632**	**Average: 15.94**

Table 65: Mid- and low-ranking keywords for SAS

The keywords *miner* and *companion* are not found at all in the termbase, either as unigrams or in any MWTs. These keywords occur in the corpus almost exclusively as part of product names: *(SAS) Enterprise Miner*, *(SAS) Text Miner*, and *SAS Companion*, which explains why they are not in the termbase. (SAS has chosen to exclude most product names from its termbase, unlike Symantec and HP.) The keywords *bar* and *scatter* are only present in the termbase in MWTs. The remaining keywords are present as unigrams and as nodes in MWTs.

7.5.2.3 Symantec

Keyword	Frequency of the keyword	Termbase terms with keyword	Frequency of the termbase terms	Termbase frequency as % of keyword frequency
worm	7,090	9	905	**12.76**
cloud	7,860	19	815	**10.37**
boot	6,285	35	3,173	50.49
wizard	6,168	92	2,801	45.41
cluster	8,490	28	6,448	75.95
portal	5,133	33	1,725	33.61
patch	3,968	13	1,898	47.83
snapshot	2,999	29	878	29.28
spam	10,411	25	1,255	**12.05**
key	18,801	39	5,645	30.02
Total	**77,205**		**25,543**	**Average: 34.88**

Table 66: Mid- and low-ranking keywords for Symantec

With the exception of *cloud*, all of these keywords are present in the termbase as unigrams. When there is a relatively high number of multi-word termbase terms containing the keyword, but this set of terms occurs comparatively infrequently in the corpus, such as for *wizard*, one can expect this set of terms to contain infrequently-occurring members. A concordance confirms this assumption.

7.5.2.4 Summary

Mid- and low-ranking keywords are a source of domain-specific terms that have the same surface form as a general lexicon word. Out of 27 keywords in our data sets, only two are not present in the termbases (either as unigrams or in n-grams), and these are associated with product names (*miner* and *companion*), which is a special case. This finding suggests that terminologists do not have difficulty in identifying domain-specific homographs.

7.5.3 Keywords that are non-existent or rare in the reference corpus

By sorting the keywords according to frequency in the reference corpus, we can focus on those that do not occur or are extremely rare in the reference corpus, as shown in the following screen capture for Minitab.

N	Key word	Freq.	%	Texts	RC. Freq.	RC. %	Keyness
1	DOTPLOT	348		43	0		1,080.27
2	UNCHECK	371		70	0		1,151.66
3	WOPEN	374		10	0		1,160.98
4	RQL	390		19	0		1,210.65
5	AQL	395		19	0		1,226.17
6	SUBDIALOG	431	0.01	81	0		1,337.92
7	LANEY	432	0.01	30	0		1,341.03
8	MINITAB'S	546	0.01	86	0		1,694.92
9	LSL	590	0.01	34	0		1,831.51
10	POPUP	626	0.02	58	0		1,943.27
11	USL	666	0.02	32	0		2,067.45
12	DEFECTIVES	723	0.02	57	0		2,244.40
13	OPTIMIZER	751	0.02	32	0		2,331.32
14	TAGUCHI	1,112	0.03	48	0		3,452.05
15	XBAR	1,153	0.03	60	0		3,579.34
16	SUBCOMMANDS	1,288	0.03	81	0		3,998.46
17	WEIBULL	1,371	0.03	69	0		4,256.15
18	SUBCOMMAND	1,490	0.04	84	0		4,625.61
19	MTW	2,047	0.05	93	0		6,355.01
20	SUBC	2,173	0.05	9	0		6,746.24
21	ENDOFTEXT	26,126	0.66	100	0		81,234.5
22	ODBC	457	0.01	52	1		1,404.86
23	STDEV	460	0.01	58	1		1,414.16
24	SIXPACK	475	0.01	31	1		1,460.66
25	NONNORMAL	670	0.02	65	1		2,065.32
26	CTRL	931	0.02	79	1		2,874.93
27	COUNT	1,363	0.03	95	1		4,215.35
28	SAME	4,932	0.12	100	1		15,295.9

Figure 26: Minitab keywords that are non-existent or rare in the reference corpus

While these keywords may not achieve a high rank using the keyness measure, their rarity in the reference corpus renders them uniquely distinctive in the company corpus. In this section, we examine some of these keywords.

7.5.3.1 Minitab

The following keywords do not occur in the reference corpus:

- dotplot
- RQL
- AQL
- subdialog
- LSL
- popup
- USL
- defectives
- xbar
- subcommand
- MTW

The following keywords are extremely rare in the reference corpus:

Keyword	Frequency in reference corpus
ODBC	1
sixpack	1
appraiser	4
boxplot	2
macro	3
toolbar	5
worksheet	17
repeatability	11
dialog	11
simplex	13
orthogonal	13
factorial	13
scatterplot	20

Table 67: Minitab - Keywords that are rare in the reference corpus

Of these 24 keywords, 19 are present in the termbase. The keywords missing from the termbase are *subdialog, popup, dialog, simplex,* and *orthogonal*. We will examine

concordances of some of these terms to determine if they are productive as nodes in forming MWTs.

7.5.3.2 SAS

The following keywords do not occur in the reference corpus:

- OLAP
- MDDP
- datalines

The following keywords are extremely rare in the reference corpus:

Keyword	Frequency in reference corpus
DBMS	2
toolbar	5
descriptor	6
pointer	10
locale	10
bytes	10
dialog	11
pane	11
widget	11
metadata	12
cursor	12
authentication	12
workspace	13
console	13

Table 68: SAS - Keywords that are rare in the reference corpus

Of these 17 keywords, 11 are in the termbase. The keywords missing from the termbase are: *descriptor, bytes, dialog, pane, console,* and *datalines.* Two of these, *dialog* and *meta-data*, occur frequently enough in the company corpus to also rank in the top keywords previously studied. We will examine concordances of some of these keywords to determine if they are productive as nodes in forming MWTs.

7.5.3.3 Symantec

The following keywords do not occur in the reference corpus:

- antivirus
- spyware
- netbackup
- phishing
- malware
- antispam
- deduplication
- adware
- reseller
- logon
- installer
- javascript

The following keywords are extremely rare in the reference corpus:

Keyword	Frequency in reference corpus
virtualization	1
scalable	6
dialog	11
pane	11
downtime	11
authentication	12
console	13
messaging	14
firewall	16
spam	16
hackers	33
trojan	34

Table 69: Symantec - Keywords that are rare in the reference corpus

Of these 24 keywords, 19 are in the termbase. Two of them, *antivirus* and *firewall*, occur frequently enough in the company corpus that they are also among the top-ranked keywords previously studied. The missing keywords are: *javascript, virtualization, dialog,*

pane, and *messaging.* We will examine concordances of some of these keywords to determine if they are productive as nodes in forming MWTs.

7.5.3.4 Summary

Keywords that do not occur or are extremely rare in the reference corpus are highly domain-specific. Generally, these keywords are fairly well represented in the termbases; in our data, 75 percent of them have been recorded. This suggests that highly specialised keywords are not difficult for terminologists to identify.

7.6 Collocate relationship measures

When looking for a word's collocates, we are interested in word associations that evoke a certain significance. Significance typically refers to relative frequency, i.e. two words occur more often together than they do alone, or a word occurs more often with one specific word than with some other word, or simply, that the co-occurrence of the two words is unlikely to be due to chance. Sinclair (2004: 28) notes the following:

> The test of whether two words are significant collocates... requires 4 pieces of data; the length of the text in which the words appear, the number of times they both appear in the text, and the number of times they occur together. (sic[84])

WordSmith offers six different statistical formulae to measure the significance of two collocates, i.e. the strength of their association with each other[85].

1. Log Likelihood
2. Z-Score
3. Specific Mutual Information (SMI)
4. Dice Coefficient
5. MI3
6. T-Score

84 Although Sinclair states four, he only lists three here. We suppose that he is counting the second item as two pieces of data, being applied to two words.

85 These measures are referred to in the literature as association measures, relationship measures, or significance measures.

Each method also requires a word list from the studied corpus, which provides the frequency of each individual word as mentioned by Sinclair.

In this section we briefly describe the six formulae. We then show the resulting collocate ranking for each formula using the term *factorial* and the Minitab corpus. Based on this comparative ranking, we chose the Dice method for our subsequent collocation search. Calculation formulae are shown in Appendix F.

7.6.1 Log likelihood

Log-likelihood is commonly used to measure the strength of association of collocates. However, Hoffman et al (2008: 152) claim that it has a bias towards high-frequency collocates, and Baker notes it places more emphasis on grammatical words (2006: 102).

N	Word	Relation	L3	L2	L1	R1	R2
1	FACTORIAL	89,657.	187	198	37	37	198
2	DESIGN	31,728.	198	129	65	1,892	27
3	DESIGNS	15,669.	77	68	52	760	27
4	FULL	10,013.	28	1	646	2	20
5	PLOTS	9,306.4	61	32	67	582	34
6	ANALYZE	8,131.9	31	12	452	69	55
7	DOE	6,286.5	107	50	259	18	4
8	FRACTIONAL	4,983.5	3		288	2	24
9	STAT	4,752.4	62	265		3	44
10	CREATE	4,344.5	30	27	259	55	25
11	A	3,663.0	107	301	233	1	37
12	CHOOSE	2,930.5	125	165		2	155
13	ENDOFTEXT	2,903.9	109	99	79	9	58
14	LEVEL	2,772.4	8	66	266		6
15	GENERAL	2,564.7	13	207	10	2	17
16	STATDOEFACTORIALANALYZE	2,420.2			120		1
17	PLOT	2,366.4	56	42	89	10	29
18	MAIN	1,991.6	16	15	1	3	102
19	TO	1,675.3	62	45	4	11	271
20	RESPONSE	1,577.2	34	41		53	73
21	ALSO	1,530.8	131	2	9		1
22	EFFECTS	1,488.1	25	26	11		9
23	USE	1,465.8	36	68	93	2	42
24	SURFACE	1,454.7	10	21	13	7	47
25	THE	1,414.0	77	107	126	11	42
26	SEE	1,351.0	10	9	7		4
27	CONTOUR	1,323.6	26	32		23	42
28	MODEL	1,287.5	26	31	31	38	9
29	TWO	1,277.6	47	91	1	1	27
30	POINTS	1,174.4	44	16	13	20	8

Figure 27: Log-likelihood ranking of the term: factorial

We can observe that indeed grammatical words have been ranked highly, such as *a*, *to*, and *the*, as well as potentially insignificant collocates such as *full*, *main*, and *also*.

7.6.2 Z-score

Z-score is a well-balanced hybrid measure which nevertheless tends to have a low-frequency bias (Hoffman). For better results, it should be used with a minimum threshold.

N	Word	Relation	L3	L2	L1	R1	R2	R3
1	FACTORIAL	892.303	187	198	37	37	198	187
2	DESIGN	260.338	198	129	65	1,892	27	164
3	DESIGNS	242.887	77	68	52	760	27	41
4	FULL	215.605	28	1	646	2	20	36
5	FRACTIONAL	182.694	3		288	2	24	11
6	DOE	160.640	107	50	259	18	4	40
7	ANALYZE	155.180	31	12	452	69	55	17
8	STATDOEFACTORIALANALYZE	146.774			120		1	47
9	PLOTS	125.034	61	32	67	582	34	91
10	STATDOEFACTORIALCREATE	90.052			30			15
11	STAT	84.401	62	265		3	44	33
12	CREATE	80.047	30	27	259	55	25	32
13	GENERAL	64.300	13	207	10	2	17	15
14	BURMAN	56.491	5		2		3	40
15	PLACKETT	55.409		2		3	40	23
16	FACTORIALANALYZE	55.069			10			
17	STATDOEFACTORIAL	53.771	3	17	2			12
18	STATDOE	47.319	8	12	10			14
19	STATDOEFACTORIALFACTORIAL	46.653						21
20	GENERATORS	39.834	2	3	2		19	2
21	LEVEL	39.609	8	66	266		6	30
22	FRACTION	38.431	7	3	9	2	3	7
23	CUBE	37.966	5	9	6	2	25	14
24	ANALYZING	37.875	24	12	4		24	
25	MAIN	33.532	16	15	1	3	102	45
26	OVERLAID	33.151	24			24	16	3
27	CONTOUR	32.857	26	32		23	42	21

Figure 28: Z-score ranking of the term: factorial

Several highly-unusual collocates have been highly ranked, such as on lines 8, 10, 17 and 19, which confirms the low frequency bias. These are menu paths, where the > character has not been properly parsed (for instance: STAT > DOE > FACTORIAL > CREATE)

7.6.3 Specific Mutual Information

Also known as Pointwise Mutual Information (PMI), Specific Mutual Information (SMI) is one of the standard association measures in collocation extraction (Bouma 2009). It was first adopted for lexicography by Church and Hanks in 1990. SMI compares the probability of co-occurrence of x and y given their joint distribution and the probability of their co-occurrence given only their individual distributions. Stated more simply, it compares the probability of observing two words together with the probability of observing each word independently, based on the frequencies of the words (Biber et al 1998: 266).

Bouma observes that with SMI, infrequent word pairs can receive relatively high scores and therefore tend to be ranked high. This is explained by the extreme cases where two parts of a bigram only occur together, i.e. p (x; y) = p (x) = p (y). This observation was confirmed by our samples, where the concordance with SMI ranked unique but rare collocates highly (some again are menu paths where the > character has not been correctly parsed), as shown in the following figure.

N	Word	Relation	L3	L2	L1	R1	R2
1	STATDOEANALYZE	10.889			2		
2	DESIGNCHOOSE	10.889		1		2	
3	DESIGNSANALYZING	10.889			1	1	
4	DESIGNSFITTING	10.889				1	
5	TOOLBARSDOE	10.889	1				
6	FACTORIALANALYZE	10.737			10		
7	STATDOEFACTORIALCREATE	10.654			30		
8	FACTORIALCREATE	10.626			3		
9	LEVELFRACTIONAL	10.474			2		
10	FACTORIAL	10.401	187	198	37	37	198
11	STATDOEFACTORIALANALYZE	10.282			120		1
12	DESIGNSNAMING	9.889					
13	PSSIF	9.889				2	
14	UNCONFOUNDED	9.889					
15	DOEFACTORIALFACTORIAL	9.889					
16	REDUCINGREDUCING	9.889	2				
17	LAPSE	9.889					
18	FITTINGFACTORIAL	9.889					
19	DOEFACTORIALANALYZE	9.889			10		
20	FACTORSBY	9.889					
21	LEVELSFACTORIAL	9.889					2
22	DESIGNSSETTING	9.889				4	
23	FRACTIONAL	9.869	3		288	2	24
24	STATDOEFACTORIAL	9.767	3	17	2		
25	STATDOEFACTORIALDEFINE	9.648	2	9			
26	FACTORIALFACTORIAL	9.474					
27	STATDOEFACTORIALFACTORIAL	9.410					

Figure 29: SMI ranking of the term: factorial

7.6.4 Dice Coefficient

The Dice Coefficient is a measure of the similarity between two samples. It tends to favour collocates that occur with a frequency close to that of the node, and its results are similar to the Z-score (Hoffmann et al 2008: 157).

N	Word	Relation	L3	L2	L1	R1	R2
1	FACTORIAL	1.426	187	198	37	37	198
2	DESIGNS	0.390	77	68	52	760	27
3	DESIGN	0.375	198	129	65	1,892	27
4	FULL	0.305	28	1	646	2	20
5	ANALYZE	0.244	31	12	452	69	55
6	DOE	0.213	107	50	259	18	4
7	PLOTS	0.207	61	32	67	582	34
8	FRACTIONAL	0.160	3		288	2	24
9	STAT	0.144	62	265		3	44
10	CREATE	0.136	30	27	259	55	25
11	GENERAL	0.101	13	207	10	2	17
12	STATDOEFACTORIALANALYZE	0.081			120		1
13	LEVEL	0.073	8	66	266		6
14	MAIN	0.065	16	15	1	3	102
15	SURFACE	0.060	10	21	13	7	47
16	EFFECTS	0.059	25	26	11		9
17	CONTOUR	0.058	26	32		23	42
18	CHOOSE	0.055	125	165		2	155
19	PLOT	0.049	56	42	89	10	29
20	BURMAN	0.047	5		2		3
21	ALSO	0.047	131	2	9		1
22	TOPIC	0.047	3	3		1	3
23	PLACKETT	0.046		2		3	40
24	RESPONSE	0.043	34	41		53	73
25	SEE	0.043	10	9	7		4
26	ENDOFTEXT	0.042	109	99	79	9	58

Figure 30: Dice ranking of the term: factorial

These results look promising, and there are fewer unusual collocates resulting from the menu paths than Z-score (only one in the above sample, on line 12).

7.6.5 MI3

MI3 is a variation of the mutual information formula that is intended to reduce the latter's strong low-frequency bias. Rather, it tends towards a high frequency bias (Hoffmann et al 2008: 156), and places more emphasis on grammatical words (Baker 2006: 102).

N	Word	Relation	L3	L2	L1	R1	R2
1	FACTORIAL	34.469	187	198	37	37	198
2	DESIGN	31.085	198	129	65	1,892	27
3	DESIGNS	29.700	77	68	52	760	27
4	FULL	28.628	28	1	646	2	20
5	ANALYZE	27.620	31	12	452	69	55
6	PLOTS	27.499	61	32	67	582	34
7	DOE	27.187	107	50	259	18	4
8	FRACTIONAL	26.893	3		288	2	24
9	STAT	25.503	62	265		3	44
10	CREATE	25.235	30	27	259	55	25
11	STATDOEFACTORIALANALYZE	24.808			120		1
12	GENERAL	23.798	13	207	10	2	17
13	A	23.741	107	301	233	1	37
14	LEVEL	23.363	8	66	266		6
15	CHOOSE	23.312	125	165		2	155
16	ENDOFTEXT	23.175	109	99	79	9	58
17	PLOT	22.669	56	42	89	10	29
18	MAIN	22.420	16	15	1	3	102
19	THE	22.051	77	107	126	11	42
20	BURMAN	21.783	5		2		3
21	TO	21.774	62	45	4	11	271
22	PLACKETT	21.686		2		3	40
23	EFFECTS	21.655	25	26	11		9
24	SURFACE	21.625	10	21	13	7	47
25	RESPONSE	21.509	34	41		53	73
26	ALSO	21.476	131	2	9		1

Figure 31: MI3 ranking of the term: factorial

Indeed, grammatical words have been ranked high, such as *a*, *the*, and *to*, as well as the potentially insignificant words mentioned earlier (*main*, *full*).

7.6.6 T-Score

T-Score is a hybrid measure with results similar to log-likelihood but with a stronger high-frequency bias (Hoffman 2008: 155). It is more suitable for researching how pairs of words are used differently, rather than as a measure of the strength of association between two words (Biber et. al 1998: 267). Biber notes that T-scores are not appropriate for compiling a list of the most important collocates for a single node word (p. 268).

N	Word	Relation	L3	L2	L1	R1	R2	R3
1	FACTORIAL	77.254	187	198	37	37	198	187
2	DESIGN	55.337	198	129	65	1,892	27	164
3	DESIGNS	37.478	77	68	52	760	27	41
4	PLOTS	32.528	61	32	67	582	34	91
5	A	30.164	107	301	233	1	37	93
6	FULL	29.273	28	1	646	2	20	36
7	ANALYZE	28.249	31	12	452	69	55	17
8	ENDOFTEXT	24.188	109	99	79	9	58	15
9	THE	23.971	77	107	126	11	42	170
10	STAT	23.853	62	265		3	44	33
11	DOE	23.762	107	50	259	18	4	40
12	CHOOSE	22.969	125	165		2	155	25
13	CREATE	22.879	30	27	259	55	25	32
14	TO	22.849	62	45	4	11	271	29
15	PLOT	21.076	56	42	89	10	29	61
16	LEVEL	20.858	8	66	266		6	30
17	FRACTIONAL	19.085	3		288	2	24	11
18	AND	18.658	25	29	10	44	91	40
19	USE	18.169	36	68	93	2	42	31
20	MAIN	17.739	16	15	1	3	102	45
21	GENERAL	17.436	13	207	10	2	17	15
22	RESPONSE	17.367	34	41		53	73	25
23	MODEL	16.735	26	31	31	38	9	30
24	ALSO	16.688	131	2	9		1	3
25	SEE	15.944	10	9	7		4	10
26	TWO	15.697	47	91	1	1	27	20

Figure 32: T-score ranking of the term: factorial

Here, we can notice a number of grammatical and common words ranking highly, such as *a*, *the*, *to*, and *and*.

7.6.7 Comparison and selection

The following table summarises our assessment of relationship measures.

Measure	Pros	Cons	Comment	Observation from trials
SMI	Standard measure	Ranks rare but unique collocates high		Not suitable because of extremely rare collocates involving non-terms (menu paths)
DICE		Favours collocates close to the frequency of the node	Similar to Z-score	Good results
MI3		High frequency bias More grammatical words	Used by Daille	Does contain some grammatical words
Z-Score	Hybrid, well balanced	Low frequency bias	Use with a minimum frequency threshold	A few extremely rare collocates involving non-terms (menu paths)
T-Score	Hybrid	More suitable for comparing pairs of words	Similar to LL but with stronger high frequency bias	Very close to pre-calculation (raw frequency) ranking
Log-Likelihood	Standard measure	High frequency bias More grammatical words		Does contain some grammatical words

Table 70: Comparison of collocate relationship measures

Measures with a high frequency bias favour grammatical words, and measures with a low frequency bias favour accidental anomalies such as file paths. The Dice and Z-Score measures emerge as most suitable for our purposes, both effectively filtering out grammatical words. But the Dice measure eliminates more of the rare anomalies than Z-Score. This improved performance seems to reflect Hoffman's observation that it favours collocates that occur with a frequency close to that of the node. This behaviour may be ideal for our purposes; since keywords are domain-specific unigrams, and we are interested in discovering MWTs of similar status with respect to both domain-specificity and frequency, it would seem that MWTs whose frequency approaches that of the keyword would be most relevant.

Relying on visual observations of the ranking of one keyword in one corpus, as well as advice from scholars, our evaluation of the six relationship measures is more anecdotal than empirical. This is sufficient given our aim, which is to explore, later, whether or not a

relationship measure can lead to the discovery of additional terms beyond those identified with raw frequencies alone. Possibly all of the six relationship measures can achieve this to some degree, but our analysis suggests that Dice would perform the best in this regard.

Based on the previous analysis, we have chosen the Dice measure for identifying collocates of keywords in the subsequent sections.

7.7 Concordances and collocations

A *collocation* is the occurrence of two or more words within a space of each other in a text (Sinclair 1991: 170). The lexical units in a collocation are called collocates. When a collocation expresses a delimitable concept, it could be a MWT. For instance, as we will show later, the words *exponential*, and *smoothing* occur frequently together forming the term *exponential smoothing*, and numeric qualifiers such as *single* and *double* occur frequently with this term, to form *double exponential smoothing*, and so forth. However, when a collocation expresses a multi-concept phrase, such as *boot* and *computer* as in "boot the computer," according to terminology theory it is not one term but several, since in this case the verb expresses one concept and the noun another. Typically, sequences comprising a verb and a noun are phrases, and often sequences involving prepositions and other function words are as well. Nevertheless, we have observed some such phrase constructions in the four termbases in this study, suggesting that there is sometimes a need to record certain phrasal collocations in termbases designed to support production-oriented requirements.

In this section we will examine collocates of some of the more prominent keywords identified in previous sections. We achieve this by concordances of the keyword in WordSmith. A *concordance* is a collection of occurrences of a word, each in its own textual environment, where the keyword plays the role of the *node*, and the words we find in its environment are the *collocates*, to adopt the terminology used by Sinclair (1991: 32, 115, 175). We now look at concordances of some of the keywords presented in Section 7.5 that are under-represented in the termbases. We also apply the Dice relationship measure to identify collocates that, with the keyword, potentially form a MWT.

7.7.1 Top-ranking keywords

7.7.1.1 Minitab

Keyword	Keyword frequency	Termbase terms with keyword	Frequency of termbase terms	Termbase freq. as % of keyword frequency	Termbase terms in range A or absent	% of termbase terms in Range A or absent
data	38,525	53	5,373	**13.95**	15	28.30

Five of the 53 termbase terms containing *data* do not occur in the corpus, and ten occur in the low frequency range A (10 times or less, see Section 6.2.1.3). Using the Patterns and Collocates functions in WordSmith, we identified the following terms that are missing from the termbase and that are potentially significant because of their frequency.

Term	Frequency
data set	2,013
sample data	1,018
data description	672
response data	477
data collection	1,123
example data	232
Total	**5,535**

Table 71: MWTs from the Minitab corpus containing the node term: data

The total occurrences of these six terms in the corpus exceeds the total occurrences of all 53 terms containing *data* that are currently in the termbase. The Dice relationship measure produces interesting results shown in the following figure.

N	Word	Relation	L4	L3	L2	L1	R1	R2	R3	R4
1	DATA	3.406	764	663	659	98	98	659	663	764
2	DESCRIPTION	0.274	20	19	32	206	672	8	855	37
3	DISTRIBUTION	0.231	227	414	79	18	40	46	107	380
4	YOUR	0.225	93	96	253	2,263	9	91	97	78
5	SAMPLE	0.221	94	55	75	1,014	21	63	74	115
6	SET	0.218	47	26	58	8	1,718	35	18	63
7	MINITAB	0.207	192	192	50	49	139	166	146	128
8	COLUMN	0.192	151	260	138	44	104	77	159	214
9	COLUMNS	0.185	168	209	234	11	72	148	126	130
10	PLOT	0.181	219	154	113	99	6	75	125	128
11	COLLECTION	0.169	9	69	14	12	1,123	7	1	
12	WORKSHEET	0.164	120	66	62	80	3	120	291	203
13	DISPLAY	0.159	84	55	98	65	529	31	71	23
14	CHOOSE	0.154	102	145	121	133	163	92	46	124
15	CHART	0.153	131	155	68	155		47	52	160
16	ENTER	0.151	43	85	202	148	107	98	135	87
17	PROCESS	0.145	94	114	159	260	23	62	118	137
18	VARIABLE	0.137	31	44	16	37	1	646	68	57
19	INTERPRETATION	0.136	537	21	2	1		3	2	1
20	OVERVIEW	0.133	226	38	228	53	40	97	36	59
21	WINDOW	0.131	47	39	25	14	606	4	12	78
22	VALUES	0.130	54	177	27	12	384	64	76	112
23	ANALYSIS	0.129	166	189	99	63	280	78	124	106
24	USE	0.126	196	133	182	147	120	108	142	83
25	NORMAL	0.125	78	16	10	111	37	98	259	199
26	MODEL	0.114	166	361	136	29	12	57	33	49
27	EXAMPLE	0.114	82	101	125	308	56	54	83	45
28	REGRESSION	0.109	49	292	20	51	18	60	61	63
29	POINTS	0.107	13	28	10		597	36	28	43
30	GRAPH	0.104	50	37	29	34	3	75	136	97

Figure 33: Dice-ranked collocates of the term: data

The Dice ranking confirms the salience of most of the MWTs we previously identified through raw frequencies. However, it also enables us to identify additional terms that are missing from the termbase, such as *data folder* (570), and *random data* (227). The former occurs more often than the most frequent termbase term (526), and the latter occurs more often than 83 percent of the termbase terms. Investigating collocates in the L2 or R2 position helps to identify frequently-occurring trigrams. The keyword *variable* for instance, which occurs 646 times in R2 position, often occurs as *data set variable*.

Keyword	Keyword frequency	Termbase terms with keyword	Frequency of termbase terms	Termbase freq. as % of keyword frequency	% of termbase terms in Range A or absent
model	15,049	30	2,813	**19**	33

Three of the 30 termbase terms containing *model* do not occur in the corpus (one has a potentially unnecessary pre-modifier: *special quartic model*), and seven occur in the low frequency range A (ten times or less). The 30 termbase terms occur 2,813 times in the corpus. The following MWTs containing *model* are missing from the termbase.

Term	Frequency
regression model	810
quadratic model	154
response surface model	65
Total	**1,029**

Table 72: MWTs from the Minitab corpus containing the node term: model

By adding three terms, we have increased the correspondence of the termbase terms to the corpus by 36 percent.

We are not able to close the frequency gap for *model* further by discovering additional MWTs because the term *model* occurs frequently on its own; over 7,000 occurrences of *model* are preceded by an article, pronoun, simple adjective or other function word, such as *the model*, *one model*, *your model*, *a model*, *any model*, *new model*, and *best model.*

The Dice relationship measure enables us to confirm the salience of some of the terms identified earlier through raw statistics: *regression model* ranks second and *quadratic model* ranks twenty-fourth. However, it also serves to identify several additional terms, such as *residuals model* (141), and *stored model* (138).

Keyword	Keyword frequency	Termbase terms with keyword	Frequency of termbase terms	Termbase freq. as % of keyword frequency	% of termbase terms in Range A or absent
column	10,603	8	144	**1.36**	25.00

Two of the eight termbase terms containing *column* occur in the low frequency range A (10 times or less). The termbase terms occur only 144 times in the corpus. The following frequently-occurring MWTs containing *column* are missing from the termbase.

Term	Frequency
storage column	134
numeric column	137
response column	119
worksheet column	99
variable column	43
label column	20
grouping column	78
factor column	70
data column	121
censoring column	103
column data	52
column delimiter	19
column variable	32
Total	**954**

Table 73: MWTs from the Minitab corpus containing the node term: column

With these 13 new terms, the frequency of the termbase terms containing *column* in the corpus would increase from 144 to 954. The Dice relationship ranking of *column* once again confirms the salience of the terms previously identified.

Keyword	Keyword frequency	Termbase terms with keyword	Frequency of termbase terms	Termbase freq. as % of keyword frequency	% of termbase terms in Range A or absent
process	12,553	17	1,686	**13.43**	35.29

One of the 17 termbase terms that contain *process* does not occur at all in the corpus, and five others occur in the low frequency range A (ten times or less). The total number of occurrences of the termbase terms containing *process* is 1,686. The following frequently-occurring MWTs containing *process* are missing from the termbase.

Term	Frequency
mixture process variable	51
process spread	210
process mean	653
process standard deviation	104
manufacturing process	111
process capability	469
process tolerance	94
Total	**1,692**

Table 74: MWTs from the Minitab corpus containing the node term: process

By adding these seven terms, we have doubled the correspondence of the termbase terms to the corpus. The Dice relationship ranking also identifies these terms, plus additional ones: *process variation* (644), *process stability* (74) and *process parameter* (100).

Keyword	Keyword frequency	Termbase terms with keyword	Frequency of termbase terms	Termbase freq. as % of keyword frequency	% of termbase terms in Range A or absent
plot	15,339	92	12,232	79.74	36.96

The keyword *plot* is relatively well represented in the termbase, as reflected by the figure of nearly 80 percent. Of the 92 termbase terms containing *plot*, 21 do not occur, and 13 occur in the low frequency range A (ten times or less). The percentage of infrequent termbase terms can be reduced by checking concordances and making adjustments, such as by adding *capability plot*, which occurs 69 times, to complement *binomial capability plot*, which is a new term not yet reflected in the corpus. The following terms are missing from the termbase.

Term	Frequency
surface plot	495 [86]
split-plot (adj)	155 [87]
plot point	162
Total	**1,075**

Table 75: MWTs from the Minitab corpus containing the node term: plot

86 Occurrences without *3D* as pre-modifier
87 Occurrences without *design*

As stated earlier, we did not expect to find many productive MWTs containing *plot* that are missing from the termbase, due to the relatively high frequency rate of the termbase terms compared to all terms containing *plot* (80 percent). In fact, virtually all the types of plot found in the corpus are also documented in the termbase.

7.7.1.2 SAS

Keyword	Keyword frequency	Termbase terms with keyword	Frequency of termbase terms	Termbase freq. as % of keyword frequency	% of termbase terms in Range A or absent
statement	165,489	17	2,848	**2**	35

Six of the 17 termbase terms containing *statement* occur in the low frequency range A (58 occurrences or less). The total occurrences of the termbase terms is 2,848. The following frequently occurring terms containing *statement* are missing from the termbase.

Term	Frequency
model statement	5,146
output statement	3,381
class statement	3,210
weight statement	3,161
test statement	1,822
input statement	1,755
plot statement	1,413
statement details	5,402
Total	**25,290**

Table 76: MWTs from the SAS corpus containing the node term: statement

Many are written in upper case in the corpus, such as *MODEL statement*. By adding just eight terms, the correspondence between the termbase and the corpus has increased nearly nine fold. Counted individually, each of these terms occurs more often than the complete set of existing termbase terms containing *statement*. The collocations indicate that one of the termbase terms, *statement label*, is in the wrong order; it should be *label statement*.

The Dice ranking is similar. Six of the above eight terms rank in the top 25 collocates, the exceptions are *plot statement* (33) and *test statement* (55), which still rank fairly high. However, additional terms are revealed: *slice statement* (1,066), *strata statement* (1,052), and *forecast statemen*t (740), among others. Although of lower frequency, several others appear interesting, such as *statement spline* (113) and *statement lag* (113).

Keyword	Keyword frequency	Termbase terms with keyword	Frequency of termbase terms	Termbase freq. as % of keyword frequency	% of termbase terms in Range A or absent
page	99,728	23	789	**1**	**91**

Two of the 23 termbase terms containing *page* do not occur in the corpus, and the remaining occur in the low frequency range A except *page size* (413) and *page view* (72). In this case, we are curious to discover whether the keyword *page* refers to its primary meaning (i.e. pages in a document), or to another concept. The most frequent collocates indicate the former, such as *next page*, *previous page*, and *last page*. Due to their general meaning, these do not need to be in a termbase, and they account for a large portion of the gap between the termbase and the corpus. Other frequent terms, such as *welcome page* and *home page*, may need to be included in a termbase because they refer to types of Web pages, they appear on user interfaces, and therefore need to be translated consistently. However, there are some domain-specific uses, such as the following, none of which are in the termbase.

Term	Frequency
code page	103
page layout	130
buffer page	77
page template	249
page break	99
page overlay	106
Total	**764**

Table 77: MWTs from the SAS corpus containing the node term: page

By adding only six terms, all of which are domain-specific, we have nearly doubled the correspondence between the termbase and the corpus. In this case, the Dice ranking did not offer any additional suggestions.

Keyword	Keyword frequency	Termbase terms with keyword	Frequency of termbase terms	Termbase freq. as % of keyword frequency	% of termbase terms in Range A or absent
procedure	63,804	13	986	**2**	**62**

Two of the 13 termbase terms containing *procedure* do not occur in the corpus, and six occur in the low frequency range A (up to 58 occurrences). Two evoke redundancy; the termbase contains both *catalogued procedure* (59) and *SAS catalogued procedur*e (32). Only two occur frequently, *SAS procedure* (564) and *procedure output file* (136), although it could be argued that both these terms are ill-formed; the former would be more repurposable as simply *procedure*, and the latter should be *procedure output* since this bigram occurs 1,698 times alone and in combination with other headwords. We found the following frequent terms in the corpus that are missing in the termbase.

Term	Frequency
syntax procedure	2,072
model procedure	926
capability procedure	625
procedure output	1,698
procedure code	185
Total	**5,506**

Table 78: MWTs from the SAS corpus containing the node term: procedure

By adding five terms, we have increased the termbase/corpus correspondence more than five-fold. All the newly proposed collocates rank in the top 30 according to the Dice coefficient, except *code* which appears in position 150. However, Dice identifies additional terms: *sort procedure* (624), *access procedure* (803), *print procedure* (434), *univariate procedure* (524), *mixed procedure* (536) and *logistic procedure* (397), among others, many occuring more frequently than all the terms currently in the termbase. If we removed the

two nonextant termbase terms, added the five shown in the previous table, and replaced the six existing low-frequency termbase terms with the six identified through Dice, the number of occurrences would increase from the current 986 to 9,671, a near ten-fold improvement.

Keyword	Keyword frequency	Termbase terms with keyword	Frequency of termbase terms	Termbase freq. as % of keyword frequency	% of termbase terms in Range A or absent
syntax	38,808	5	197	**1**	**75**

All the termbase terms containing *syntax* occur in the low frequency range A except one: *syntax error* (116). The following terms, frequent in the corpus, are not in the termbase.

Term	Frequency
syntax description	2,426
syntax procedure	2,072
syntax arguments	596
Total	**5,094**

Table 79: MWTs from the SAS corpus containing the node term: syntax

By adding only these three terms, we have increased the correspondence between the termbase and the corpus 25 fold. All three of these collocates rank in the top 15 by the Dice measure. But Dice gives us more: *syntax details* (4,000), *example syntax* (1,463), *syntax RC* (224) and *command syntax* (141) are worth mentioning. Further investigation reveals that *RC* stands for *return code*. All of these seven newly-identified terms are much more frequent than the terms currently in the termbase.

7.7.1.3 Symantec

Keyword	Keyword frequency	Termbase terms with keyword	Frequency of termbase terms	Termbase freq. as % of keyword frequency	% of termbase terms in Range A or absent
computer	48,657	14	1,573	**3**	64

Two of the 14 termbase terms that contain *computer* do not occur in the corpus, and seven occur in the low frequency range A (up to 52 occurrences). The only frequent terms are *client computer* (639), *host computer* (350), and *computer system* (313). The following frequently-occurring terms are not in the termbase:

Term	Frequency
compromised computer	1,063
remote computer	1,935
commercial computer software	322
infected computer	709
bot-infected computer	227
desktop computer	402
source computer	251
destination computer	341
target computer	319
computer security	341
computer disaster	278
Total	**6,188**

Table 80: MWTs from the Symantec corpus containing the node term: computer

Terms such as *compromised computer*, *infected computer*, *bot-infected computer*, *computer security* and *computer disaster* are clearly specific to Symantec's business sector. By adding these 11 terms, we have increased the correspondence between the termbase and the corpus four-fold. Using the Dice relationship measure to calculate salient collocates produces the very interesting results shown in the following figure.

N	Word	Relation	L4	L3	L2	L1	R1	R2	R3	R4
1	COMPUTER	21.362	444	388	159	17	17	159	388	444
2	RESTART	3.066	17	69	3,324	15	46	133	46	16
3	YOUR	2.066	276	202	669	15,034	165	496	411	370
4	NORTON	1.761	395	44	10	10	74	271	345	319
5	REMOTE	1.429	112	49	145	1,209	17	35	130	120
6	INSTALLED	1.419	363	1,055	24	2	39	15	89	66
7	WINDOWS	1.331	128	214	376	221	32	247	217	188
8	BACKUP	1.106	203	110	61	12	82	101	311	153
9	SOFTWARE	1.013	227	413	56	32	440	144	107	157
10	FILES	0.933	255	325	99	7	35	124	118	113
11	SERVER	0.870	177	117	63	199	18	29	87	108
12	COMPROMISE	0.806	1	2		986	1	4	9	16
13	INFECTED	0.776	19	11	90	338	15	298	310	103
14	INSTALL	0.735	101	62	6	2	35	86	178	111
15	THREATS	0.715	124	177	20	8	62	28	130	246
16	RUNNING	0.711	115	167	8	2	426	179	181	176
17	CLICK	0.661	48	259	221	36	158	250	210	85
18	GHOST	0.660	125	59	168	2	30	58	109	87
19	INTERNET	0.655	112	65	33		8	56	331	268
20	VIRUSES	0.650	56	88	9	1	188	278	97	64
21	CLIENT	0.633	50	34	19	639		4	15	44
22	BOOT	0.602	47	39	370	9	19	37	30	69
23	PROTECTION	0.597	110	329	32	18	109	39	104	137
24	SCAN	0.594	45	102	340	8	42	25	71	42
25	PROTECTED	0.538	6	9	7	30	414	114	96	51
26	DISK	0.535	101	97	14	2	14	19	23	93
27	VIRUS	0.532	114	96	47	2	166	76	82	88
28	NETWORK	0.524	76	56	44	53	54	60	267	148
29	SECURITY	0.521	265	196	35	15	336	63	144	126
30	RECOVERY	0.508	96	29	13	12	5	37	111	119

Figure 34: Dice-ranked collocates of the term: computer

Here we can easily identify not only frequent collocates forming MWTs, such as *remote computer* and *compromised computer*, but also potentially interesting verbs, such as *restart*, *boot*, and *scan*, which occur hundreds or thousands of times.

The most frequent termbase term (*client computer*, 639) ranks in 21st position in the Dice calculation. In the L1 position, *compromised computer* and *remote computer* are much stronger collocations, yet neither are in the termbase. If we take the top 14 collocates according to Dice, focusing for the moment on bigrams, we obtain the following terms:

	Term	Frequency
1	remote computer	1,209
2	computer software	440
3	server computer	199
4	compromised computer	986
5	infected computer	338
6	client computer	639
7	computer virus	166
8	computer security	336
9	desktop computer	238
10	host computer	350
11	commercial computer	355
12	destination computer	231
13	computer system	317
14	computer users	240
	Total	**6,044**

Table 81: MWTs from the Symantec corpus containing computer, ranked by Dice

The frequency of these terms is nearly four times that of the 14 existing termbase terms (1,573). In addition, interesting collocates in the Dice list, such as *ghost*, can be searched to find other domain-specific terms, in this case *ghost server computer*, which occurs 120 times. Checking the concordances directly as well as via the Clusters function allows us to determine that occasionally bigrams join together to form a trigram, for instance, *computer software* and *commercial computer* actually form *commercial computer software* (322).

Keyword	Keyword frequency	Termbase terms with keyword	Frequency of termbase terms	Termbase freq. as % of keyword frequency	% of termbase terms in Range A or absent
subscription	18,455	7	562	**3**	43

Two of the seven termbase terms that contain *subscription* do not occur in the corpus, and two occur in the low frequency range A (up to 52 times). The only frequent terms are *Norton subscription* (140 occurrences), *paid subscription* (83), and *Symantec subscription*

(241). The following frequently-occurring terms are not in the termbase:

Term	Frequency
subscription renewal	1,173
service period subscription	343
subscription term	880
subscription service	923
subscription key	819
subscription period	759
academic subscription	175
trial subscription	114
subscription troubleshooter	47
subscription clock	48
Total	**5,172**

Table 82: MWTs from the Symantec corpus containing the node term: subscription

Note that *subscription service* is in the termbase, but only in the four-gram term *virus definition subscription service*, which does not occur in the corpus. This is clearly a case of a boundary-setting problem where two terms are recorded as one, i.e. *virus definition* (which occurs 1,233 times) and *subscription service*. Perhaps the notion of a *virus definition subscription service* exists in the company, or existed at one time. Whatever the case, it is not present in the corpus under study. In cases like this, the terminologist may have to conduct further research and consult subject-matter experts. Regardless of the decision with respect to this four-gram, the two bi-grams are frequent on their own and in combination with other collocates, and so they should be included in the termbase.

The Dice ranking gives us the following additional terms: *product subscription* (804), *subscription server* (159), and *subscription price* (139). As we did for many of the terms in this study, concordances of the terms *subscription troubleshooter* and *subscription clock* were examined directly to confirm that they are valid terms, since they appeared slightly unique and occurred less frequently than many of the other terms.

```
This was due to a product error whereby the subscription
clock was not deducting time from user's subscriptions
if those systems entered "hibernate" or "stand by" mode.
```

```
The issue at hand is that products subscription clock is
displaying an inaccurate amount of time remaining.

Please make use of our subscription troubleshooter to
order your Virus Update Subscription.

The Symantec Subscription Troubleshooter is an automated
tool designed to help you find the documents that you
need.
```

Keyword	Keyword frequency	Termbase terms with keyword	Frequency of termbase terms	Termbase freq. as % of keyword frequency	% of termbase terms in Range A or absent
installation	22,053	22	1,108	**5**	59

Seven of the 22 termbase terms that contain *installation* do not occur in the corpus, and six occur in the low frequency range A (up to 52 times). The only frequent terms are *Installation Wizard* (299), *installation package* (164), and four others that occur between 100 to 140 times. The following frequently-occurring terms are not in the termbase:

Term	Frequency
initial installation	1,988
product installation	722
software installation	294
installation CD	621
installation file	740
installation program	294
remote installation	222
server installation	171
client installation	178
installation process	379
Total	**5,609**

Table 83: MWTs from the Symantec corpus containing the node term: installation

The Dice ranking identifies additional terms: *Backup Exec installation* (343), *ghost installation* (64), and *expert installation service* (138). Examining the concordances directly

reveals frequent domain-specific pre-modifiers: *problem-free*, *failed*, *successful*, *fresh*, *scripted*, and *silent*, among others. It is possible to find groups of terms that may be synonyms or near synonyms and therefore possibly subject to prescription within the company, such as *problem-fre*e and *successful* among the pre-modifiers just mentioned, as well as *bad software installation*, *failed software installation*, and *improper software installation*. Sometimes the contextual environment points out synonyms, as in the following excerpt, "silent mode installation, also known as command line installation." It is also easy to spot less frequent but noteworthy terms such as *installation footprint* (32), *installation path* (108) and *installation script* (90). Terms that are frequently found as titles of topics, such as *installation instructions* (138) and *installation guide* (143), might need to be documented in the termbase to ensure consistent translations. Such contextual environments can be discovered in the concordances. Finally, the formulation *initial installation and activation*, slightly unusual in that it contains the conjunction *and*, occurred a surprising 754 times.

Keyword	**Keyword frequency**	**Termbase terms with keyword**	**Frequency of termbase terms**	**Termbase freq. as % of keyword frequency**	**% of termbase terms in Range A or absent**
download	20,369	13	1,158	**6**	36

One of the 13 termbase terms that contain *download* does not occur at all in the corpus, and four occur in the low frequency range A (up to 52 times). The only frequent terms are *Download Insight* (420), *Download Manager* (168), and *Norton Download Insight* (166), the latter seeming somewhat redundant given the first. Indeed, five of the terms start with *Norton*, and four of these already exist in the termbase without it, such as *Download Intelligence* (38) and *Norton Download Intelligence* (occurring only once). It appears that ten of the terms are proper names, as they are capitalised. The non-capitalised terms are *download service* (95), *drive-by download* (62) and *internet download* (27). The following frequently-occurring terms are not in the termbase:

Term	Frequency
download process	174
download button	128
download link	72
product download	195
agent download	114
framework download	97
software download	245
dangerous download	229
Total	**1,254**

Table 84: MWTs from the Symantec corpus containing the node term: download

The unigram *download* is frequently used as a noun ("the download") and as a verb ("to download," "please download," etc.), which accounts for its high frequency as a keyword. Both a noun and a verb entry for the unigram are available in the termbase.

The Dice measure identifies the following additional terms: *download file* (378), *download folder* (50), *download dialog* (53) and *file download* (234). The concordances of *download dialog* and *file download* enables us to observe that the longer term *file download dialog box* occurs 33 times, which is fairly frequent for a term of this length. Similarly, a concordance of *download service* enables us to observe that *Extended Download Service* occurs 82 times. Checking the concordances directly allows us to identify some additional terms that, although less frequent, are of potential interest. For instance, the difference between *trialware download* (61) and *trial download* (27) could be investigated and clarified.

In total these additional terms occur 2,175 times, which is nearly a 90 percent increase over the existing termbase terms. By including in this set several of the more frequent existing termbase terms, we could triple the frequency to nearly 3,000.

7.7.1.4 Summary

When the occurrence of termbase terms containing a top-ranking keyword is low compared to the occurrence of the keyword alone, some very important MWTs are missing from the

termbase. These MWTs can be easily identified using a concordancing software. Application of the Dice relationship measure will identify additional important MWTs that would otherwise be overlooked by using the raw frequencies alone. We have shown that this methodology can significantly increase the correspondence of the termbase to the corpus.

7.7.2 Mid- and low-ranking keywords

As discussed earlier, among the mid- and low-ranking keywords, one finds words that exist both in the company corpus and the reference corpus; often the term in the company corpus has assumed a domain-specific meaning. We are interested in determining whether these types of polysemic keywords are sufficiently represented in commercial termbases and whether they are productive in forming MWTs of significant frequency. For this purpose we will examine concordances of some keywords. Our selections focus on three types:

1. keywords forming few termbase terms
2. keywords forming termbase terms that occur very infrequently compared to the frequency of the keyword
3. keywords that are present in a significant number of termbase terms, yet their total frequency remains relatively low compared to the frequency of the keyword

7.7.2.1 Minitab

The keyword *smoothing* fits into type three. There are nine termbase terms containing this keyword, but their total frequency is only 52 percent that of the keyword itself.

- 0 concordance lines found for *Brown's double exponential smoothing*
- 0 concordance lines found for *Holt-Winters double exponential smoothing*
- 2 concordance lines found for *Holt-Winters seasonal exponential smoothing*
- 11 concordance lines found for *degree of smoothing*
- 120 concordance lines found for *double exponential smoothing*
- 2 concordance lines found for *resistant smoothing*
- 2 concordance lines found for *single exponential smoothing method*
- 140 concordance lines found for *single exponential smoothing*
- 22 concordance lines found for *smoothing constant*

Nearly half the 576 occurrences of *smoothing* are unaccounted for in the termbase. Furthermore, five of the nine termbase terms occur in the low frequency range A (up to 10 occurrences), suggesting that this set of terms is not optimised. We can also see that the terms with proper names as pre-modifiers are lacking in the corpus or are extremely rare. Two terms that do not occur are redundant, given that the term *double exponential smoothing* is already in the termbase. Six of the nine terms contain *exponential smoothing*. Indeed, this term occurs 282 times in the corpus as a bigram and combined with various pre-modifiers, yet it is not in the termbase. Including just this one term in the termbase would allow us to consider eliminating six less productive terms (the pre-modifiers *single* and *double* appear to be non-essential) and at the same time increase the correspondence of the termbase to the corpus. The Dice measure identifies one additional term, *smoothing constant* (72).

The keyword *bar* fits into type one. There are only three terms in the termbase:

- 836 concordance lines found for *bar chart*
- 67 concordance lines found for *interval bar*
- 23 concordance lines found for *status bar*

Compared to *smoothing*, which was present in nine termbase terms, *bar* forms much fewer termbase terms and yet those terms occur more frequently (56 as opposed to 52 percent of the keyword occurrences). Here we see that all three terms occur above the low frequency range A. Nevertheless, *bar* also occurs with other frequent collocates that are missing from the termbase, such as *title bar* (22), *stacked bar chart* (59), *clustered bar chart* (57), *interval bar* (67), and *menu bar* (19). Unlike the trigrams *single exponential smoothing* and *double exponential smoothing* in the previous example, which one could say are redundant given the term *exponential smoothing* and the simplicity of the pre-modifiers, the words *stacked* and *clustered* are unusual and would pose translation difficulties. The keyword *bar* occurs rarely in the corpus as a unigram, but appears in a number of distinct bigrams, such as *soap bar*, *fruit bar*, *hot bar* (*pressure* and *temperature*), *steel bar*, and even the expression *bar none!* Therefore, if *bar* is included in the termbase as a unigram, a separate entry is required for each of its different meanings (*bar* as in *fruit bar*, *bar* as in *soap bar*, *bar* as in *steel bar*, and so forth), since each may require a different translation. This example demonstrates the value of concept-orientation. Interestingly, the Dice measure helps to separate

the truly domain-specific *bar* terms from those that are more general, as shown in the following figure, where the collocates *fruit* and *soap* rank lower than *interval* and *stacked*.

N	Word	Relation	Texts	L4	L3	L2	L1	R1	R2	R3
1	BAR	1.923	88	53	32	34	5	5	34	32
2	CHART	1.830	60	36	42	11	17	837	2	51
3	CHARTS	0.330	26	9	3	13	5	134		3
4	GRAPH	0.241	29	9	14	5	67	3	8	15
5	DISPLAY	0.228	24	6	35	81	15	1	25	18
6	EDIT	0.224	9	18	8	23	11		33	6
7	PLOT	0.206	22	20	13	24	42		6	5
8	INTERVAL	0.190	19	4	2	18	67	7	5	11
9	STACKED	0.164	10		1	1	65			1
10	OPTIONS	0.147	18	6	9	9	15	22	21	2
11	BARS	0.134	21	6	10	11	6		3	23
12	COUNTS	0.131	16	4		13	2		16	46
13	CLUSTERED	0.121	10	1	4		58			
14	HISTOGRAM	0.099	21	6	4	6	12	1	3	13
15	CHOOSE	0.092	19	8	2	37	2	1	30	3
16	MINITAB	0.092	14	1	6	1			9	10
17	PARETO	0.082	13	7	10	4			4	5
18	ATTRIBUTES	0.068	8	2	3	4		15	2	
19	COMMANDS	0.068	6	3	1	15	1			16
20	SKEWNESS	0.061	24			32				
21	INTERPRETING	0.061	5	30						4
22	FRUIT	0.060	3				40			
23	BOXPLOT	0.058	10	11	4	1	1		4	3
24	SOAP	0.058	3				16	12		2
25	PIE	0.055	16		1				11	10
26	TAILS	0.055	24		32					
27	VALUES	0.054	21	8	16	21	7	1	15	7
28	REPRESENTS	0.054	7		2			17	1	3
29	SIMPLE	0.050	7	1			44			1
30	MENU	0.049	9	3		1	19		1	1

Figure 35: Dice ranked collocates of the term: bar

Note also that *fruit bar* and *soap bar* occur in only three files (as indicated in the Texts column), out of a total of 100 files in the corpus, meaning that these terms are confined to a

small section of the corpus, and are probably restricted to a specific context. Indeed, further examination reveals that they are used in examples for statistical calculations:

```
Fruit Bar Appearance

Problem
Food engineers need to find the combination of
components that maximizes the appearance for the new
fruit bar. The filling contains strawberry, raspberry,
and blueberry purees. In addition, two process variables
—the amount of vanilla and malt in the crust—may affect
appearance. Use an extreme vertices design with two
process variables to investigate appearance.

Data collection
A total of 104 trained sensory panelists evaluate 4
fruit bars, each having a different combination of
ingredients. Panelists evaluate the fruit bars in random
order (run order).


Soap Bar Appearance

Problem
Overweight soap bars cause smudged logos when the
machine squeezes them into cartons. To evaluate the
molding process over time, engineers use control charts
to plot soap bar weights.

A machine stamps extruded logs of soap into five bars
using a single 5 cavity die. The cavities are labeled.
Each cycle produces five bars, one from each mold.
```

7.7.2.2 SAS

The keyword *filter* fits into type two. The eight termbase terms account for only one percent of the occurrences of *filter* in the corpus; they all occur in the low frequency range A (up to 58 occurrences).

- 5 concordance lines found for *MIME type filter*
- 3 concordance lines found for *entry filter*
- 13 concordance lines found for *name/value filter*
- 1 concordance lines found for *package entry MIME type filter*
- 2 concordance lines found for *package entry type filter*
- 6 concordance lines found for *package filter*
- 30 concordance lines found for *prompted filter*
- 27 concordance lines found for *reply filter category*

A concordance and Dice measure of *filter* reveals many more potentially significant terms: bigrams such as *filter window* (111), *filter node* (255), *filter pane* (192), *filter criteria* (213), *filter section* (80), *filter expression* (129), *filter options* (105), and *filter viewer* (66) as well as trigrams such as *text filter node* (171), *Filter dialog box* (147), and *Interactive Filter Viewer* (57). Altogether, these terms alone occur 1,526 times, compared to only 67 occurrences for the existing termbase terms.

Keywords that occur infrequently in the termbase (type one), such as *wizard* and *locale*, both of which form only two terms in the termbase, may be productive in the formation of additional compounds that are missing from the termbase. A concordance shows this to be the case. While only *SAS Deployment Wizard* (855) and *migration wizard* (23) are in the termbase (accounting for only 23 percent of the concordances of the keyword), we found *Adapter Setup Wizard* (565), *Cube Designer Wizard* (234), *New Library Wizard* (118), *Backup Wizard* (51) and many others. And while the termbase only contains *locale* and *locale list*, we found *locale system* (292), *data locale* (105) and *default locale* (72) among others. We also found *Locale Setup Manager*, and while it only occurs 18 times, it has a corresponding acronym, *LSM*, which occurs 31 times. Acronyms, even if infrequent, are important to record in termbases to provide users with knowledge about their meaning which is reflected in their full form. Finally, the adjectival form *locale-specific* is used 48 times with different nouns including *character*, *date format*, and *punctuation*.

A keyword that forms quite a few terms in the termbase, but yet these terms account for proportionally few occurrences in the corpus (type three), such as *block* (20 terms but only 15 percent of the corpus occurrences), is a likely candidate as the node of termbase terms that are very infrequent. Indeed, the corpus evidence supports this:

- 3 concordance lines found for *Application Control Block*
- 2 concordance lines found for *I/O Program Communication Block*
- 0 concordance lines found for *IP block*
- 9 concordance lines found for *Program Communication Block*
- 11 concordance lines found for *Program Specification Block*
- 135 concordance lines found for *SUBMIT block*

- 13 concordance lines found for *block cipher*
- 30 concordance lines found for *block map*
- 42 concordance lines found for *block menu*
- 41 concordance lines found for *block plot*
- 296 concordance lines found for *block size*
- 267 concordance lines found for *block variable*
- 25 concordance lines found for *cell block*
- 2 concordance lines found for *child block*
- 16 concordance lines found for *data control block*
- 5 concordance lines found for *define block*
- 95 concordance lines found for *layout block*
- 76 concordance lines found for *method block*
- 2 concordance lines found for *parent block*
- 115 concordance lines found for *statement block*

Seventy percent of the 20 termbase terms occur in the low frequency range A. Other more frequent terms are missing from the termbase; 17 bigram and trigram terms occur in the higher frequency range B or above (over 58 occurrences), for instance, *holder block* (115), *queue block* (88), and *stats collector block* (78). If we were to replace the low frequency termbase terms with these 17, the value of this set of terms, in so far as repurposing potential is concerned, would be much higher.

7.7.2.3 Symantec

The keyword *spam* is frequent in the termbase (forming 25 terms) yet these terms account for a low proportion (12 percent) of the total frequency of the keyword. It therefore fits into type three.

- 0 concordance lines found for *Anti-Spam scanner*
- 1 concordance lines found for *Brightmail Spam Folder Agent for Exchange*
- 17 concordance lines found for *Email Anti-Spam Service*
- 12 concordance lines found for *MessageLabs Email Anti-Spam Service*
- 0 concordance lines found for *predictive spam detection*
- 0 concordance lines found for *Responsive Spam Detection*
- 3 concordance lines found for *SMS Anti-Spam*

- 7 concordance lines found for *spam definition*
- 253 concordance lines found for *spam filter*
- 302 concordance lines found for *spam filtering*
- 84 concordance lines found for *Spam folder*
- 87 concordance lines found for *Spam Manager*
- 16 concordance lines found for *spam message*
- 1 concordance lines found for *Spam Monitor*
- 0 concordance lines found for *spam originator*
- 327 concordance lines found for *spam prevention*
- 70 concordance lines found for *spam quarantine*
- 22 concordance lines found for *spam rule*
- 18 concordance lines found for *spam tagging*
- 1 concordance lines found for *spam trap*
- 20 concordance lines found for *Symantec Brightmail Anti-Spam*
- 10 concordance lines found for *Symantec Spam Folder Agent*
- 4 concordance lines found for *Symantec Spam Folder Agent for Exchange*
- 0 concordance lines found for *Symantec Spam Folder Agent for Microsoft Exchange*
- 0 concordance lines found for *Symantec Spam Plug-in for Outlook*

Over 75 percent of the 25 termbase terms occur in the low-frequency range A (up to 52 occurrences). The only frequently-occurring terms are *spam prevention*, *spam filtering*, and *spam filter*. However, the following terms are missing from the termbase: *spam email* (360), *spam threat* (146), *spam messages* (211), *spammer* (336) and more. In fact, taking only a few minutes to examine the concordances and Dice measures, we were able to create a different list of 25 terms containing *spam* whose total frequency is 2,723 occurrences, which is an increase of over 100 percent compared to the original list (1,255).

The concordances also identify differences in word formation, for instance, *antispam* vs. *anti-spam*, *multi-layer spam prevention* vs. *multi-layered spam prevention*, and *spam e-mail* vs *spam email.* While such minor inconsistencies have negligeable impact on SL text, they do reduce the exact match rate of TM in CAT tools, which increases translation costs (Warburton 2001b: 680). Companies therefore are well-advised to address them.

Another type three keyword is *cloud*. It is frequent in the termbase (forming 19 terms) yet these terms account for a low proportion (10 percent) of the total frequency of the keyword.

- 0 concordance lines found for *cloud definition*
- 7 concordance lines found for *Cloud Scan*
- 0 concordance lines found for *Cloud Scanning*
- 0 concordance lines found for *Cloud Services for Backup Exec*
- 139 concordance lines found for *cloud storage*
- 11 concordance lines found for *cloud technology*
- 0 concordance lines found for *cloud-based scanning*
- 0 concordance lines found for *Discovery Enterprise Vault.cloud*
- 0 concordance lines found for *Enterprise Vault.Cloud administration console*
- 0 concordance lines found for *Enterprise Vault.cloud Console*
- 6 concordance lines found for *NetBackup Cloud Storage*
- 0 concordance lines found for *NetBackup Cloud Storage for Nirvanix OpenStorage*
- 0 concordance lines found for *NetBackup Cloud Storage Server*
- 0 concordance lines found for *Norton Utilities Cloud*
- 630 concordance lines found for *Symantec.cloud*
- 5 concordance lines found for *Symantec.cloud agent*
- 17 concordance lines found for *Symantec.cloud Management Console*
- 0 concordance lines found for *Symantec.cloud Protection Agent*
- 0 concordance lines found for *tag cloud*

Ninety percent of the 19 termbase terms occur in the low frequency range A, 12 not at all. Only two occur frequently: *Symantec.cloud* and *cloud storage*. Many of the terms appear to be product names[88]. Note the unusual use of punctuation in several, which contain a period. None of the termbase terms containing *Vault.cloud* occur in the corpus, but *Vault.cloud* does occur 262 times as *Enterprise Vault.cloud*. This is a clear example of a boundary-setting problem. Including long MWTs containing an already documented node, such as *NetBackup Cloud Storage for Nirvanix OpenStorage* and *NetBackup Cloud Storage Server* (nodes: *cloud storage* and *NetBackup Cloud Storage*) has resulted in redundancy.

88 See for instance: http://www.symantec.com/en/ca/products-solutions/families/?fid=symantec-cloud

As noted earlier, the unigram *cloud* is not present in the termbase. If it were, it would cover the frequent use of *cloud* as a unigram in the corpus (as in "in the cloud") as well as reduce the need to include low frequency terms such as *cloud technology* not to mention terms that do not occur, such as *cloud definition*. A concordance and Dice measure on the node *cloud* reveal some significant terms that are missing from the termbase, such as *Backup Exec Cloud* (632), *endpoint protection cloud* (307), *cloud service* (353), *cloud solution* (153), *email continuity cloud* (102), *cloud computing* (428), *cloud securit*y (100) and *virtualization cloud* (68). In addition, one can easily find term pairs, such as *private cloud* (382) and *public cloud* (61). Finally, frequent adjectival formulations such as *cloud-based* (835) and *cloud-managed* (168) can be documented in the termbase to cover an unlimited number of noun phrases such as *cloud-based scan* and *cloud-managed service*.

If we add the 12 aforementioned terms to *cloud storage* and *Symantec.cloud* from the original list, the total frequency count rises to 4,358, an increase of over 400 percent compared to the original set of terms (815 occurrences).

The keyword *worm* combines features of types one and two; compared to the other mid- and low-ranking keywords, there are relatively few terms in the termbase (9), and they still occur infrequently (13 percent of the keyword occurrences). Thus, the problem shifts from too many infrequent terms being documented, to not enough terms documented. Two of the nine terms do not occur in the corpus.

- 319 concordance lines found for *Internet Worm Protection*
- 358 concordance lines found for *mass-mailing worm*
- 28 concordance lines found for *Norton Internet Worm Protection*
- 0 concordance lines found for *Symantec defined worm list*
- 13 concordance lines found for *worm attack*
- 88 concordance lines found for *worm blocking*
- 0 concordance lines found for *worm signature*
- 11 concordance lines found for *WORM storage*

A concordance and Dice measurement reveal missing terms: *email worm* (and *e-mail worm*), and *worm variant*, as well as named worms such as *Conficker worm*, *CodeRed*

worm and *Blaster worm*. With function words such as *the*, *a*, and *this* the most frequent left collocates, *worm* is primarily a unigram. The term *WORM media* occurs 74 times, but it is an acronym (write once read many). Two variants of the termbase term *mass-mailing worm* were identified: *mass-mailer worm* and *mass mailing worm* (no hyphen).

7.7.2.4 Summary

Domain-specific terms that share their surface form with a general lexicon word can be found in mid- and low-ranking keywords. They are highly productive in forming MWTs and therefore in most cases should be present as unigrams in the termbase. Terminologists should take care to determine the most productive boundaries of MWTs to avoid redundancy in the termbase, as in the case of the potentially unnecessary trigrams and four-grams based on *exponential smoothing* (see Section 7.7.2.1).

When the frequency of termbase terms based on these keywords is low compared to the frequency of the keyword itself, some important terms are missing from the termbase. When the number of termbase terms containing the keyword is relatively high, this is a clue that there are redundant terms in the termbase. Redundancies can result from term boundary-setting problems, such as the case of *Enterprise Vault.cloud* which was only entered in the termbase in the form of longer MWTs, none of which are present in the corpus. Likewise, adding a proper noun to a core term can result in a term that is less prevalent in the corpus, as in the case of *Brown's double exponential smoothing* which does not occur in the corpus whereas *double exponential smoothing* occurs 120 times. However, one cannot generalise and assume that proper nouns should be stripped from all MWTs, as there are cases where they form a valid term, such as the various named worms (*Conflicker worm*, etc.) and wizards (*SAS Deployment Wizard*, etc.). Using corpus evidence, the terminologist can decide which terms need to be recorded in the termbase, the term with the proper noun retained, the term without the proper noun, or both.

As we saw with the keywords *bar* and *smoothing*, the Dice relationship measure helps to distinguish between domain-specific and non-domain-specific uses of the keyword.

7.7.3 Keywords that are non-existent or rare in the reference corpus

As shown in Section 7.5.3, keywords that are rare or non-existent in the reference corpus are highly domain-specific terms. Thus, even if they occur relatively infrequently compared to high-ranking keywords, due to their domain specificity they and the MWTs they form may be important for a company termbase. In this section, we will examine concordances of these types of keywords, focusing on those that are also missing from the termbase as unigrams. Our aim is to determine if these keywords are frequent enough in the corpus, and productive enough in forming compounds, to justify their inclusion in termbases.

7.7.3.1 Minitab

Keywords that are absent or rare in the reference corpus and are also missing from the termbase are shown in the following table along with corpus frequencies.

Keyword	Frequency
subdialog	431
dialog	7,485
simplex	451
orthogonal	485

Table 85: Minitab keywords that are absent or rare in the reference corpus

Although the keywords *dialog* and *subdialog* are not in the termbase, the terms *dialog box* and *subdialog box* are. A concordance search of *dialog* and *subdialog* shows that about 90 percent occur in combination with *box,* and this form is indeed the standard that Minitab has adopted. However, there are other occurrences of *dialog* and *subdialog*, both as a unigram and in combination with other collocates such as *dialog instruction*, in sufficient numbers to justify the inclusion of these keywords in the termbase as unigrams. There are only 18 occurrences of the hyphenated termbase form *sub-dialog* in the corpus, demonstrating that this form is an inconsistency.

The keyword *orthogonal*, being an adjective, occurs in the termbase but only in three MWTs: *Taguchi orthogonal array design* (6), *orthogonal design* (31), and *orthogonal regression* (187), accounting for less than half the occurrences of the keyword. However, the keyword also occurs in the corpus as a unigram (as in "the model is orthogonal"), and with other head nouns: *orthogonal array*, *orthogonal blocking*, *orthogonal blocks*, *orthogonal matrix*, and more. These collocates justify the inclusion of *orthogonal* as a unigram in the termbase. We even found the derived noun *orthogonality* a surprising 90 times.

The keyword *simplex* occurs in four MWTs in the termbase; one is not found in the corpus (*L-simplex design*), but this is not unexpected because this term is a new standard. The remaining three, all trigrams, occur frequently (280 occurrences or 62 percent of the keyword occurrences). Each also exists in the corpus in an abbreviated bigram form which is absent from the termbase, for instance, *simplex centroid design* also occurs as *simplex centroid*, and *simplex lattice design* also occurs as *simplex lattice* (i.e. without *design*, 21 and 39 occurrences respectively). It is also used occasionally as a noun itself ("a simplex"). For these reasons, documenting the keyword as a unigram in the termbase is justified.

7.7.3.2 SAS

Keywords that are absent or rare in the reference corpus and are also missing from the termbase are shown in the following table along with corpus frequencies.

Keyword	Frequency
descriptor	4,226
bytes	2,352
dialog	15,604
pane	2,790
console	3,224
datalines	2,760

Table 86: SAS keywords that are absent or rare in the reference corpus

The keyword *descriptor* is present in the termbase in eight MWTs (seven bigrams and one trigram), but not as a unigram. These MWTs account for 3,377 occurrences, or about 80

percent of the total, even though four of them occur in the low frequency range A (up to 58 occurrences). This is due to the fact that two of the bigrams are very frequent: *view descriptor* (1,588) and *access descriptor* (1,331). This keyword thus appears to be already sufficiently documented in the termbase without existing as a unigram, although, eliminating several of the lowest frequency MWTs and adding the keyword as a unigram would improve the overall repurposability of this set of terms. It should be noted that two similar unigram terms, *subdecriptor* (22) and *superdescriptor* (27), are present in the termbase although they are much less frequent by comparison, and so it would seem odd that their more frequent base form would be overlooked.

The keyword *bytes* occurs exclusively as a unigram in the corpus and therefore should be documented in the termbase.

The case of *dialog* is similar with that of Minitab. Only the term *dialog box* appears in the termbase. Ninety percent of its corpus occurrences are also *dialog box*. Nevertheless, this leaves over 1,500 occurrences as a unigram, or with different collocates, such as *dialog window* (100). This justifies its inclusion as a unigram in the termbase.

The keyword *pane* does not exist in the termbase either as a unigram or in any MWT. There is an abundance of bigram terms in the corpus: *properties pane* (145), *gallery pane* (121), *filter pane* (192), *navigation pane* (99), *workspaces pane* (72), *contents pane* (133), and more. Clearly, both *pane* and the frequent bigrams should be documented in the termbase.

The three MWTs containing *console* in the termbase account for 80 percent of the occurrences of *console*, with the term *SAS Management Console* occurring most often (2,550). There are, however, other important collocates including *administration console* (115) and its variant *administrative console* (46). There are also many occurrences as a unigram ("the console"). This justifies the inclusion of *console* in the termbase.

The keyword *datalines* is not in the termbase either as a unigram or in any MWT. This term appears most often in the corpus in plural (only six occurrences of the singular form were

found, and they appear to be part of code strings). This term occurs exclusively as a unigram in the corpus and therefore should be included as a unigram in the termbase.

7.7.3.3 Symantec

Keywords that are absent or rare in the reference corpus and are also missing from the termbase are shown in the following table along with corpus frequencies.

Keyword	Frequency
javascript	2,555
virtualization	4,040
dialog	6,047
pane	5,795
messaging	5,596

Table 87: Symantec keywords that are absent or rare in the reference corpus

An examination of its collocations confirms that the word *javascript* is a code string and therefore not a keyword. The keyword *virtualization* appears in 10 MWTs in the termbase but not as a unigram. These terms account for only 15 percent of the occurrences of the keyword. This suggests that some important MWTs are missing and also that the unigram itself might be frequent. Furthermore, eight of the termbase terms are in the low frequency range A (up to 52 occurrences), with two not occuring at all, and another four occurring less than ten times. So clearly, the terms in the termbase are not optimised.

Terms that are missing from the termbase include: *server virtualization* (294), *storage virtualization* (204), *endpoint virtualization* (244), and *virtualization technology* (176) among others. This keyword also occurs frequently as a unigram. It should be present in the termbase as a unigram as well as some of the more frequent bigrams.

The keyword *dialog* is not in the termbase, but there are three MWTs: *dialog box* (4,325), *Dialog Workflow* (1) and *Virus Definitions Dialog Box* (0 occurrences). *Dialog box* accounts for 70 percent of all keyword occurrences, a smaller proportion than for the other

two companies, and therefore the inclusion of *dialog* as a unigram in the termbase is even more justified to account for the remaining 30 percent of the occurrences. Other terms include *send dialog* (257), *properties dialog* (184), *status dialog* (68), and *dialog window* (44), as well as a large number of less frequent bigrams and trigrams.

As for the keyword *pane*, the termbase contains only one term, *list pane*, which occurs only 11 times in the corpus. This keyword occurs frequently as a unigram ("the pane") as well as forming many frequent MWTs: *task pane* (1,231), *properties pane* (1,210), *results pane* (253), *selection pane* (163), *preview pane* (113), *backup selections pane* (76), *active alerts pane* (60), *job history pane* (67), and *current jobs pane* (69), among others.

The termbase has seven terms containing *messaging,* but not the unigram. These terms account for 31 percent of the occurrences, almost all of which are attributed to two termbase terms: *instant messaging* (1,572) and *Symantec Messaging Gateway* (157). The remaining five terms are very rare: two not occurring at all, and the remaining three occurring once, twice, and seven times. Clearly, the set of termbase terms based on this keyword is not optimised.

Terms that are missing in the termbase include *messaging server* (168), *messaging security* (734), *messaging system* (68), *text messaging* (79), and *messaging management* (258) as well as contrasting collocate pairs such as *soft messaging* (112) versus *strong messaging* (97), and *inbound messaging* versus *outbound messaging*. Collocates such as *mission-critical messaging* and *business-critical messaging* may have some marketing importance. Some frequent terms occur both as general concepts and in product names, such as *messaging gateway* (132) and *Symantec Messaging Gateway* (157), *messaging management* (170) and *Symantec Messaging Management* (55), the translations of which would likely differ. Finally, we discovered unusual variations in formulation: *softer alternative messaging* (68), *alternative soft messaging* (42), and *softer messaging alternative* (21), the differences in meaning of which are unclear and therefore warrant further investigation to avoid translation problems.

7.7.3.4 Summary

Due to their pronounced domain-specificity, keywords that are absent or rare in the reference corpus are extremely important to document in a company termbase. In our sample data, they include acronyms, designations of user interface objects (dialog, pane, console, popup), elements of product names (such as Messaging), and other specialised terms, and they tend to occur frequently. Their inclusion in the termbase as unigrams is justified.

Overall, the set of MWTs based on these keywords that are documented in the termbase is under-optimised; there are too many rare terms and not enough frequent ones. Our research shows that basing term selections on concordances can vastly improve this situation.

CHAPTER 8 CONCLUSIONS AND IMPLICATIONS

We begin this chapter by answering the research questions and then we draw conclusions based on the overall research objectives outlined in chapter 4.

8.1 The gap between termbases and corpora

In the companies we studied, the combined percentage of termbase terms that do not occur or occur infrequently in the corpus ranges from 35 percent to 73 percent[89]. This is a significant gap indeed. We have previously acknowledged that some of this gap is likely due to the corpora being incomplete. We also pointed out that when the termbase is used for controlled authoring, as in the case of Minitab, some terms are added to the termbase before they are used by content producers and will therefore not be observed in the corpus for a certain period. Nevertheless, we discovered that a significant portion of the gap is caused by linguistic and methodological weaknesses which, if addressed, would improve the value of the termbase. In the remainder of this section, we discuss this portion of the gap.

There are two types of terms that contribute to the gap: (a) terms in the termbase that are absent or infrequent in the corpus, and (b) important terms in the corpus that are missing in the termbase. Viewing these terms from the perspective of deficiencies in the termbase, we call the former category *under-optimised* and the latter *un-documented*.

Generally speaking, the longer the term, the less frequently it occurs. The most productive terms are two and three tokens in length. While this observation is not new, our research provides empirical evidence that can now serve to justify a guideline: terms longer than three tokens should be validated using corpus evidence before being included in a termbase.

Upper-case terms tend to be less frequent than their lower-case counterparts, and they account for a disproportionate number of the under-optimised termbase terms. Accidental

89 HP has not been included here, due to the logistical problem previously described in the termbase/corpus correspondence. The exclusion of HP was justified in Section 6.8

differences in case between the termbase and the corpus can potentially account for between 10 and 20 percent of the gap between the two.

When added to a core term, a non-essential or incidental word produces a MWT that is often much less frequent in the corpus than the core term without that word. The same was observed for words that render the core term overly-specific. As a general best practise, MWTs containing such words should be checked against the corpus in order to make sound decisions about setting term boundaries. Likewise, the addition of a proper noun to a core term, such as names of companies, can decrease the term's frequency significantly, since these parts of a MWT are often dropped in running text for purposes of stylistic economy. And once again, we suspect that the predominance of product names in Symantec's term-base partially accounts for its high rate of infrequent terms; product names change regularly and obsolete names often remain in termbases. When included in termbases, product names and other proper nouns should at least be characterised with the proper metadata so that they can be subject to management procedures and end uses that might differ from more conventional types of terminologies.

There are more variants in the corpora than in the termbases. Variants occurred more frequently in the corpora than conventional theories account for. Perhaps because terminologists are guided by those theories, variants are under-represented in the termbases. Furthermore, failure to record variants properly as terms in the termbase has the same effect on repurposability as not recording them at all. Our research indicates that commercial terminographers should pay special attention to variants and should adopt strict concept-oriented approaches when recording them in termbases.

Less than ten percent of the termbase terms in our study are non-nouns. This appears to be lower than the proportion of non-nominal terms in the corresponding corpora. Domain-specific verbs occur fairly frequently in the IT field due to the prevalence of action-oriented concepts on software user interfaces. Domain-specific adjectives that form many MWTs may be justified in entries as unigrams due to their term formation potential.

Finally, many frequently-occurring MWTs that contain domain-specific unigram terms -- the keywords in our research -- are also missing from the termbase.

Some of the gap between the termbases and the corpora can be attributed to term collection methods. For example, many of the nonextant or infrequent HP termbase terms are strings that were imported from TM segments. A TM segment need only occur once to be included in the import procedure, and the text from which this segment originated may at any time be modified or deleted from the current active company materials, rendering the imported "term" nonextant. Symantec routinely uses an automatic term harvesting function (a form of term extraction). This function, while it may take into account term frequency, is carried out on a project basis, therefore, on a sub-corpus. Furthermore, unless the extracted term candidates are carefully cleaned manually using the full corpus as evidence, a number of poorly-delineated terms will end up in the termbase.

Aside from the above two cases involving TM segments and automatic term extraction, the primary term collection method used in all companies is one whereby the terminologist adds terms on an ad-hoc basis or imports terminology files obtained from other sources. These files are typically glossaries prepared by technical writers or bilingual lists of terms provided by translators. The files have no *direct* relation to the current company corpus as a whole. They may have an *indirect* relation to the corpus by virtue of the terms having been selected by the technical writer or translator because they appeared in some company text. However, under these circumstances, the nature of the terms thus identified may not be optimal for serving the terminological needs of the company as a whole, for two main reasons. First, writers and translators are not terminologists and so they may lack knowledge about best practises for selecting terms. Second, the term is often selected from an isolated text -- a single document or file without the larger contextual framework of other documents or files in the company. The writer or translator works within the limited confines of one project, one product, one technology or one department. The terminologist as the gatekeeper attempts to filter out redundant terms, modify term boundaries, and identify important terms that individual translators or writers may have missed. However, with the exception of Minitab (330 employees), the number of employees in the companies in our

study range from 13,000 to 330,000. On this scale, one dedicated terminologist cannot compensate for all shortcomings. We can perhaps now better appreciate why TM segments would be imported into a termbase of a very large company. The sheer volume of content and terminology would be overwhelming.

The main motivation for managing terminology in a commercial setting is to reduce costs for content authoring and translation. To this we add that enterprises are interested in developing multi-purpose terminological resources that can be leveraged in various extended NLP-based applications, as discussed in Section 1.7.3. Bourigault and Slodzian point out that these applications are primarily "textual" which means that terminological resources intended to serve them must necessarily be corpus-based in order to reflect those texts. We like the simple way that they express this bi-directional dependency on corpora :

> La terminologie doit venir des textes pour mieux y retourner. (1999: 30)
> (Translation) Terminology must come from texts so that it can return to texts.

In other words, terminological resources must reflect terms in active use in order to enable productive *reuse*. In an article specifically dealing with workplace applications, Condamines also stated this very directly: "Terminology has to be drawn from texts written in the workplace" (2010: 45). This perspective contrasts with other environments such as public institutions where the motivation centres around language preservation or conceptual standardisation. It shifts the focus of *what terminology is* from semantic criteria to authentic discourse, purpose-driven communication needs and particularly to degrees of repurposability. The more a term is used, the more it will be required in various applications and situations, and thus, the more it should be recorded and managed in a structured digital format so that the information necessary for these uses can be leveraged in various production-oriented NLP technologies. Linguistic properties of various sorts, while important, are secondary to this pragmatic criterion. This may be difficult for some to accept, but it is reality for terminologists working for companies.

In this context, the frequency of occurrence of termbase terms in the company corpus is a major indicator of their value. From a business perspective, there is little justification for incurring the costs of managing (and therefore translating) a term that is rarely used, with

some exceptions such as when the concept has critical legal or safety ramifications. We believe that the scope of non-extant and infrequent terms discovered in this study renders a termbase under-optimised for meeting the objectives of terminography in a company, particularly given the future likelihood that its terminological resources will be required for various NLP applications, as they already are in some cases.

8.2 Economic impacts

In the preceding sections we used concordances and keywords to find productive terms that are missing from termbases. In doing so, we also identified under-optimised terms in termbases, that is, terms that bring insufficient value to the content production process to justify their inclusion in the termbase. A few comments about the significance of the problem of underoptimisation for commercial terminography is warranted.

Building and maintaining a termbase incurs costs. Teubert describes terminology as "a commercial commodity, an important economic factor" (2005: 97). He notes that small companies pay "large sums" of money for multilingual terminologies, and that large companies spend even more money on developing their own terminological resources.

Including under-optimised terms in a termbase does not come without associated costs. Aside from the IT costs of data storage and maintenance, there is the cost of the terminologist creating the entry and then the cost of finding and entering TL equivalents. According to a study commissioned for the Government of Canada (Champagne 2004: 8, 25), initially creating an entry takes about 15 minutes, and searching for and deciding on a translation takes about 20 minutes. Thus we can consider that it takes about 35 minutes to create a bilingual entry, and each additional language adds a further 20 minutes, or, that a trilingual entry takes about an hour. In another study carried out by Tekom, creating a monolingual entry takes on average 30 minutes, and adding translations takes 10 to 25 additional minutes (Schmitz and Straub 2010: 59).

Champagne also stresses that the number of times that an entry is accessed is an important factor in measuring its value to the organisation (p. 9, 30, 31). If a term in the termbase is not an element of the organisation's corpus, then it is unlikely that end-users will need to enquire about it. Likewise, if the term occurs very rarely in the corpus, then it will probably be rarely queried in the termbase as well. Below a certain threshold of use, it becomes economically unjustified to include a term in the termbase when users could probably find the information they need elsewhere, such as by conducting an internet or intranet search. This type of un-focused search is not efficient if it is repeated many times, but is reasonably justified if it is repeated infrequently. On the other hand, a termbase is cost-effective by reducing the time it takes for employees to find information that they require on a frequent basis. This is why frequency of occurrence is an important term selection criterion for termbases that are developed for production-driven requirements.

Of course, frequency of occurrence is not the only valid term selection criterion; certain other criteria justify the inclusion of infrequent terms, such as domain specificity, translation difficulty, or legal or marketing importance. When there is a need to replace a term currently used in the company by another term, such as to align with industry standards, the new term must also be added to the termbase even though it may not even exist yet in the corpus. Termbases that are used for controlled authoring need to contain new standardised terms. However, infrequent terms that have no such special status and are present accidentally, such as when the selected term boundaries are not optimal, are unjustified and add undue costs. The termbase will be much more effective if these potentially redundant terms could be replaced by more productive ones.

The tables in Section 7.5 indicate what percent of the keyword-based termbase terms occur in the low frequency range A or are absent entirely. These tables show that every set of MWTs in the termbase that contain a given keyword comprises some infrequent terms, many of which are extremely infrequent. While the presence of some infrequent terms in a termbase is to be expected and is even justified, the percentage of termbase trms that are infrequent in our four case studies is too high. Averaged across the set of keywords, these percentages are 27 for Minitab, 52 for SAS and 66 for Symantec.

If we assume that each termbase term is translated into two languages on average, it therefore takes about an hour to produce a full terminological entry. Using 60 USD per hour as a fully-burdened salary rate, we can estimate the costs of termbase entries containing terms that are nonextant or extremely infrequent in the corpus as follows[90]).

	Minitab	**SAS**	**Symantec**
Nonextant terms	203	530	2,240
Extremely infrequent terms (Range A-s[91])	127	1,086	1,358
Total	330	1,616	3,598
Cost (USD)	19,800	96,960	215,880

Table 88: Estimated cost of under-optimised entries

As we have pointed out earlier, for Minitab, and also for SAS to a lesser degree, some of these under-optimised entries are required for controlled authoring purposes. However, their large number cannot be accounted for by controlled authoring needs alone. Furthermore, we have provided sufficient examples of other causes of underoptimisation, such as boundary setting problems and unjustified use of proper names, to support our claim that under-optimised terms in termbases are indeed a major problem that impacts the return-on-investment of termbases.

Even if one could argue that the estimated entry costs are inflated (although we have used industry benchmarks in the calculation), or that a portion of under-optimised terms can be explained by deficiencies in corpus compilation, we must acknowledge that there are significant costs associated with under-optimised termbase terms and that companies therefore should be motivated to reduce their occurrence. Note also, that this cost estimate includes only nonextant and *extremely* infrequent terms and also, only the costs of entry creation. The total cost of under-optimised terms would be higher if *all* levels of low productivity with respect to term repurposability and other associated overhead costs were included.

90 In Schmitz and Straub, the higher figure of 60 Euros is used. See also: http://www.bls.gov/oes/current/oes273042.htm and http://smallbusiness.chron.com/calculate-fully-burdened-labor-costs-33072.html

91 See Section 6.4.2.

In this section, so far, we have only discussed under-optimised termbase terms. There is also undisputedly an economic impact of not documenting key terms in the termbase. It would be more difficult to quantify, as it involves not measuring the inadequacies of what we have, but the potential benefits of what we don't have. Given the evidence that we have shown where termbase/corpora correspondence can be increased many-fold with the simple addition of a few highly-productive terms, we suggest that the economic impacts of failing to document key terms are even greater than those of documenting under-performing terms.

8.3 A purpose-driven notion of terminography

We have pointed out that generally speaking, frequency of occurrence should be considered when selecting terms for a termbase. However, the types of linguistic expressions and associated metadata that need to be managed in a commercial setting also depend on the specific communicative needs and technological landscape of each company. Minitab, for instance, requires special surface form entries that are a concatenation of two or more terms to ensure that writers provide the expanded form of abbreviations and synonyms of unfamiliar terms in their texts. It also requires general lexicon expressions to serve the requirements of automated controlled authoring. Further, because its termbase is used for controlled authoring, to impose a new usage requirement, the Minitab terminologist needs to add some terms to the termbase *before* they are used. HP requires some TM segments in its termbase to raise the consistency of translations at the sub-segment level produced by translators using CAT tools. SAS requires source-language synonyms to guide authors in selecting terms consistently, as well as definitions for publishing glossaries. Symantec's termbase is heavily-laden with proper nouns, possibly due to a history of mergers and acquisitions and the high level of integration of its products with external suppliers. All companies are concerned about trademark protection, which explains the occurrence of some proper names in all the termbases. Finally, to achieve consistency between the UI and related documentation, all companies have taken specific measures some of which result in phrasal expressions in the termbase.

Thus we have seen many examples where the company terminologist accepted a term into the termbase based on application-oriented needs: Minitab's *Surface* form and general lexicon units, Symantec's proper nouns and harvested terms, SAS' deprecated synonyms, and, although we maintain that this is not recommended, HP's TM segments.

Adopting a corpus-based approach to term identification would eliminate most of the linguistic causes of the gap described in the previous section. It would allow the terminologist to see the most productive terms in context, thereby enabling him or her to make the appropriate adjustments to term boundaries or surface form, such as removing non-essential words in a MWT and choosing the appropriate case.

Commercial terminologists need to include more variants in termbases and encode them properly in the concept-oriented structure. Non-nouns should not be neglected, and terms that have a homographic surface form can be particularly important.

Multi-word terms are very important in terminology. Keywords are highly productive nodes for discovering MWTs that are either already lexicalised or occur frequently enough to warrant their inclusion in a termbase.

There is little evidence that semantic criteria is the primary motivation for the selection of many of the terms in our four company termbases. The subject field data category is not used at all in three of the termbases, and in the fourth is not yet meaningful due to the nature of the values currently available. Definitions are not included at all in two of the termbases, and in a third they are present in less than half the entries. Only the SAS termbase contains definitions for all the terms, and this is because it is monolingual and is used primarily for glossary generation. These patterns confirm Van Campenhoudt's claim that in authentic terminology resources, definitions are “systematically absent” and semantic classifications are “rare” (2006: 5). The omission of such key semantic information is in total contradiction with the GTT. Although terminologists generally acknowledge that the lexical units they add to the termbase should in theory be domain-specific, in practise other criteria often take precedence such as frequency, visibility, and the risk of inconsistency.

While a semantic basis for term selection seems to be nearly absent, usage information is important. When a termbase is used for controlled authoring purposes, and even for controlled translation purposes, terms that should not be used, or should be used only in restricted contexts, need to be recorded in the termbase with appropriate metadata to indicate their restricted use. If content producers are following these usage guidelines, then these terms should occur infrequently in the corpus.

As we saw with Minitab, when used for controlled authoring, a termbase also needs to include some expressions from the general lexicon. These expressions are not domain-specific; therefore, they would not qualify as terms according to conventional theory. It is interesting to note that all four companies in our study are implementing some level of automated controlled authoring, yet only Minitab has managed to use its termbase directly for this purpose. Controlled authoring is a recently-introduced application in business environments and most terminologists in this situation are struggling to adapt terminographical methods to what are essentially lexical resources.

8.4 Implications for theory and practise

In Section 3.3, we suggested that among the theories of terminology, the GTT is most distant from the needs of commercial applications of terminological resources. We believe we have shown this clearly to be the case with empirical evidence demonstrating a departure from purely semantic criteria, a model for term selection that is purpose-driven, and a valuation of repurposability based on corpus frequency. The various theoretical perspectives on the notion of *term* that evolved post GTT give greater importance of varying degrees to the communicative intent of interlocuteurs, application of the terminological resource, and the role of corpora for providing empirical linguistic evidence. We find that these perspectives resonate for commercial terminography. Condamines has already claimed that textual terminology "constitutes an important part of linguistics of the workplace" (2010: 46). There is definitely a place for terminology management in the private sector among the modern theories of terminology.

On the other hand, neither the production-oriented factors that are the primary motivation for commercial terminography, nor the opportunities and constraints afforded by the linguistic technologies that are increasingy available in commercial settings, have been considered in existing theoretical and methodological frameworks. Technnologies such as CAT and controlled authoring tools provide the means to achieve unprecedented levels of productivity and quality, yet their limitations can undermine the stability, quality and repurposability of terminological resources. We must prevent terminological resources from being driven in the wrong directions by technology, as we have witnessed with certain approaches adopted to handle sub-segment-level text strings and with various violations of fundamental principles for the design and use of termbases.

As we have shown in the previous section, commercial terminography is essentially pragmatic with respect to term selection. This pragmatism also extends to methodological approaches, were flexibility is necessary to address practical needs:

> Corporate standards (in terminographic approaches) need to consider, in addition to terminology theory and standards, many practical issues: the dependencies on existing tools, the conservation of existing terminology assets, the preservation of locally-realized economies, and production output requirements." (Warburton 2001b: 691).

The principle of concept-orientation, for instance, has not been adopted in the strictly localisation-oriented termbases (Symantec, HP). We maintain, however, that as the needs for terminological resources extend beyond localisation, the value of the concept-oriented approach will become evident. IBM (Warburton 2001b: 691) and Microsoft (Karsch et al 2008) have already recognised this by developing strictly concept-oriented termbases. The onomasiological and thematic methodologies adhered to by the GTT are rarely if ever used in commercial settings. The semasiological, ad-hoc approach is more practical in commercial settings because of its task-oriented focus. Van Campenhoudt boldly noted that the only terminologies that truly adhere to the ideals of the GTT are those developed by university students in projects designed to reflect those ideals. He also revealed their unsustainable cost: several hundred hours to produce a few dozen entries (2006: 6). Again, these observations are further evidenced by surveys (Lommel and Ray 2007: 31).

One cannot criticise approaches adopted by the four companies in this study that do not comply fully with the GTT or with any of the recognised terminology theories for that matter. Doing so would only serve to further alienate commercial terminography from mainstream terminology theory and practise. On the contrary, the theoretical foundations of terminology need to adapt to modern applications. An *application-oriented terminology theory and methodology* is needed. A new paradigm for terminological resources needs to take shape, one that is less constrained by fixed semantic models, and provides for sufficient flexibility to embrace different linguistic contexts, communicative goals, and end-users of terminological resources.

Along with Fuertes-Olivera, Arribas-Bano, and Van Campenhoudt (see Section 2.5.2), we see similarities between commercial terminography and LSP lexicography, such as in their use of the semasiological approach. Both make texts and lexical units the object of investigation and processing, while at the same time focusing on a semantically-restricted domain or field of application that departs from LGP. However, we also join Roche (2012), among others, in recognising that there are significant differences. For instance, the concept-oriented approach used in terminography is essential for developing multi-purpose terminological resources for companies. Concept-orientation enables the resources to be used for both controlled authoring and controlled translation. This fundamental principle is the basis for termbase design and drives several practises such as the documentation and ranking of variants and lexical synonyms. Further, one of the main aims of LSP lexicography is to *define* terms (Bowker 2003: 162-163), whereas, terminographers in commercial environments undertake this task quite infrequently. We suggest, therefore, that a new theoretical and methodological framework for production-oriented terminography could take inspiration from specialised lexicography while preserving some of the practises of terminography that have proven effective, including some traditional (such as concept orientation) and some textual and corpus-based.

We have shown that the use of corpora for selecting terms would greatly increase the value and repurposability of commercial termbases. Given the potential uses of terminological

resources, repurposability (for different purposes and in different applications) as an overarching objective in the development of terminology resources is well motivated, yet methodological approaches to achieve this objective have been rarely discussed in the literature (an exception is Bourigault and Jacquemin 2000).

Using concordancing tools and keyword identification software would benefit all terminologists working in commercial settings. Such tools are commonplace in lexicography (Prinsloo 2009: 184; Van Campenhoudt 1999), giving further weight to our proposition that commercial terminography would benefit from a rapprochement with lexicography. Other scholars and practitioners have recommended that terminologists use such technologies on a routine basis, notably Bourigault and Slodzian (1999: 31), Bourigault and Jacquemin (2000), Condamines (2010: 40), L'Homme (2004: 16-17, 225), and Van Campenhoudt (1999). Indeed, L'Homme's entire monograph (2004) describes a methodology based on the use of such tools which she calls "terminotique" in French (terminotics, in English), a neologism formed by a blend of "terminologie" (terminology) and "informatique" (computer science). Bourigault and Slodzian, as well as Bourigault and Jacquemin (2000: section 9.2.3) add, however, that the effective use of such tools requires an appropriate methodological framework defining when, where, and how to use them.

Teubert notes that a corpus used for terminology work must be dynamic, and continually enriched by new texts documenting technological innovation linguistically (2005: 103). Adopting the terminology of John Sinclair (1991: 24), Teubert refers to such corpora as "monitor corpora" (2005: 103). Terminographers working in companies that are continually developing new products would greatly benefit from having access to the totality of up-to-date product information as a research corpus. Such a corpus should be part of the standard terminologist's toolset.

Nearly 15 years ago, Rogers stated, "the precise role which documentation[92] plays in terminology work remains unclear in so far as the processes by which the terminologist is supposed to extract information on concepts and their relations from the text remain inexplicit

92 Rogers uses "documentation" to refer to the texts containing terminology.

for the large part and empirically under-researched" (2000: 5). We maintain that the role of documentation for term selection is still not well understood by practising terminologists, who therefore fail to leverage it to its full potential. We hope that the current research contributes to the growing yet still limited body of empirical research Rogers refers to.

8.5 Further reflections

This issue of the scale of the work raised earlier leads to a number of other questions. Was Minitab's smaller size as a company a factor in its better performance? Can we conclude that the number of terminologists hired in a company should be based on some quantifiable metric (number of employees, revenue, corpus size, number of languages, and so forth)? What risks are companies taking when they do not employ a reasonable number of terminologists given this metric?

On the other hand, what technological factors contributed to performance outcomes? The Minitab terminologist, for instance, stated that she has direct access to crossTank, the company's translation memory repository, in which she can carry out concordances to verify terms. How does this tool compare to what the other terminologists use? How often are such tools used? And let us not forget that the four companies use different systems to manage their terminology: crossTerm, TermWeb, WorldServer and MultiTerm.

As stated earler, we have seen cases where the company failed to observe certain best practices or norms in the field of terminography. In some cases, this can be attributed to limitations in human resources, in other cases, limitations in technology, and yet in others, simply a lack of awareness of those norms. What are the risks and financial costs of these shortcomings? Put in the specific terms of a few examples, what are the financial costs of reduced strategic repurposability when TM segments are imported into a termbase, when termbases lack key pieces of metadata such as the part-of-speech, or when term autonomy and concept orientation are violated?

While beyond the scope of this research, such questions are fundamental to raising awareness of the value, and costs, of terminology management in commercial settings. We simply hope that the current research will trigger further enquiry in these directions.

8.6 Limitations and further research

Termbases in general are not specifically developed to be representative of any particular well-delineated corpus. Furthermore, ensuring that a corpus supplied by a company is complete may not even be possible, particularly for large companies. Files tend to be distributed around the company in different locations and file systems, and often fall under the responsibility of different people. For security reasons, some SAS files could not even be provided. Although the omission of these files likely accounts for part of the gap between the termbase and the corpus, it was acknowledged that this part would not be significant enough to affect the overall findings, given that the terms in the security-sensitive files were not routinely included in the termbase. Nevertheless, in all cases, but particularly for HP, it is likely that the corpora were incomplete, which could explain some cases of nonextant or infrequent termbase terms.

The large differences in the sizes of our four corpora necessitated that we normalise the frequency counts. To calculate the normalisation factor, we chose a base of normalisation close to the size of Minitab's corpus (4 million), since it is in the middle range of the four. However, this introduced the possibility of over- or under- factorisation for the other three companies, which could affect the accuracy of the counts. Given the range of corpus sizes, such a limitation is unavoidable no matter what base of normalisation is chosen.

Since the corpora in our study were not part-of-speech tagged, some of the statistics we gathered are imprecise. In particular, it is not possible to separate the concordance statistics for the different word classes presented by homographs, such as *attribute* (noun or verb) and *constant* (noun and adjective). This limitation has been recognised in the literature from corpus-based lexicography (for example, Prinsloo 2009: 189). In addition, the words in the corpus were not lemmatised prior to frequency being measured; this adds another degree of

imprecision to the statistics. Again, this has been recognised in the literature: “lemmatisation makes frequency counts more meaningful” (Landau 2001: 336). On the other hand, lemmatising individual words would have affected the appearance and subsequent analysis of MWTs, which would have had a more serious effect on the research given their importance. It should also be pointed out that due the range of file formats, some proprietary, it would have been very challenging to use a part-of-speech tagger and lemmatiser.

Examining the termbase terms based on a more granular part-of-speech attribution at the level of the word components of MWTs could reveal additional morphosyntactic patterns to explain the gap. However, we were not able to do so for the simple reason that it is not possible to run a part-of-speech tagger on the termbase terms. The termbase terms are lexical units without the syntactic context that is needed to accomplish part-of-speech tagging. The only way to mark the part of speech of the word components is manually. With a total of nearly 17,000 terms to analyse, most of which are MWTs, this would have involved manually tagging upwards of 100,000 individual words.

Finally, we acknowledge that findings obtained from a study of four IT companies, one of which was eliminated midway through the research, cannot be generalised across the entire commercial sector. We therefore view these findings as a springboard for further research and validation.

Our findings could be validated by studying additional IT companies or companies in other sectors. Conducting a study with a termbase that was developed from the outset based on a specific, well-delineated corpus would help to eliminate the risk of statistical inaccuracy caused by the limitations of our corpus compilation methods. Using a corpus that could be part-of-speech tagged and lemmatised would reveal more about the morphosyntactic properties of productive and un-productive terms.

This research demonstrates that further exploration of the methods of LSP lexicography and corpus linguistics will lead to more effective practises for production-oriented terminography. We hope that it can serve as a catalyst for further debate about a methodological

framework for terminology management in production environments. We propose that such a framework would include the following elements:

- adopting more statistically-based critera for term selection
- using the organisation's monitor corpus as the primary source of terms
- using corpus analysis technologies such as concordancers, keyword identifiers, and collocate relationship calculators
- adopting a termbase data model that ensures that the terminological resource can be repurposed in a range of NLP applications

The current research has also stimulated many ideas for further invesigation including:

- conducting an empirical study to determine the effectiveness of mining multi-word terms from keyword-based concordances versus addressing the noise and silence of automatically-extracted term candidates
- determining the relation between terminological variation and genre in commercial content
- describing and measuring the impacts of limitations in technology, such as the limitations of sub-segment-level matching of TM in CAT tools
- describing and measuring the impacts of certain adopted practices, such as importing TM segments into termbases

BIBLIOGRAPHY

AeroSpace and Defence Industries Association of Europe (2010). *Simplified Technical English. Specification ASD-STE100. International Specification for the Preparation of Maintenance Documentation in a Controlled Language*. ASD.

Ahmad, Khurshid (2001). The Role of Specialist Terminology in Artificial Intelligence and Knowledge Acquisition. *Handbook of Terminology Management.* V. 2. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.809-844.

Ahmad, Khurshid and Margaret Rogers (2001). Corpus Linguistics and Terminology Extraction. *Handbook of Terminology Management.* V. 2. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.725-760.

Anick, Peter (2001). The automatic construction of faceted terminological feedback for interactive document retrieval. *Recent Advances in Computational Terminology*. Didier Bourigault, Christian Jacquemin, Marie-Claude L'Homme, Eds. Amsterdam. John Benjamins Publishing Company, p.29-52.

Amparo, Alcina (2009). Teaching and learning terminology. New strategies and methods. *Terminology*. Amsterdam. John Benjamins Publishing Company, V. 15, No. 1, p.1-9.

Baker, Paul (2006). *Using Corpora in Discourse Analysis*. London, U.K. Continuum.

Barriere, Caroline (2006). Semi-automatic Corpus Construction from Informative Texts. *Lexicography, Terminology, and Translation. Text-based studies in honour of Ingrid Meyer.* Lynne Bowker, Ed. Ottawa, Canada. University of Ottawa Press.

Bawarshi, Anis and Mary Jo Reiff (2010). *Genre. An Introduction to history, theory, research, and pedagogy*. West Lafayette, Indiana, USA. Parlor Press LLC.

Bellert, Irena and Paul Weingartner (1982). *Sublanguage. Studies of Language in Restricted Semantic Domains*. Richard Kittredge and John Lehrberger, Eds. Berlin. Walter de Gruyter.

Biber, Douglas, Susan Conrad and Randi Reppen (1998). *Corpus Linguistics. Investigating Language Structure and Use.* Cambridge, U.K. Cambridge University Press.

Bouma, Gerlof (2009). Normalized (Pointwise) Mutual Information in Collocation Extraction. *Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference.* Chiarcos, Eckart de Castilho and Stede, Eds. Tübingen. Gunter Narr Verlag, p.31-40. Available from: https://svn.spraakdata.gu.se/repos/gerlof/pub/www/Docs/npmi-pfd.pdf

Bourigault, Didier and Monique Slodzian (1999). Pour une terminologie textuelle. *Terminologies nouvelles*, V. 19, p.29-32.

Bourigault, Didier and Christian Jacquemin (2000). Construction de ressources terminologiques. In *Ingénierie des langues*. J.M. Pierrel, Ed. Paris. Hermès.

Bowker, Lynne (2002). An empirical investigation of the terminology profession in Canada in the 21st century. *Terminology*, V. 8, No. 2, p.283-308.

Bowker, Lynne and Ingrid Meyer (1993). *Beyond 'textbook' concept systems: Handling multidimensionality in a new generation of term banks.* In Terminology and Knowledge Engineering (TKE'93), Schmitz, K.D., Ed. Frankfurt. Indeks Verlag, p.123-137.

Bowker, Lynne and Jennifer Pearson (2002). *Working with Specialized Language. A practical guide to using corpora*. London: Routledge.

Bowker, Lynne (2003). Specialized lexicography and specialized dictionaries. In *A Practical Guide to Lexicography*. Piet van Sterkenburg, Ed. Amsterdam. John Benjamins Publishing Company, p.154-164.

Bowman, Catherine, Diane Michaud, and Heidi Suonuuti (1997). Do's and Don'ts of Terminology Management. *Handbook of Terminology Management*, V. 1. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.215-217.

Buchan, Ronald (1993). Quality indexing with computer-aided lexicography. *Terminology: Applications in interdisciplinary communication*. Helmi B. Sonneveld and Kurt L. Loening, Eds. Amsterdam. John Benjamins Publishing Company.

Budin, Gerhard (2001). A critical evaluation of the state-of-the-art of terminology theory. *Terminology Science and Research*: Journal of the International Institute for Terminology Research, IITF. Vienna, TermNet, V. 12, No. 1-2, p.7-23.

Cabré, Maria Teresa (1995). On diversity and terminology. *Terminology*. Amsterdam. John Benjamins Publishing Company, V. 2, No. 1, p.1-16.

Cabré, Maria Teresa (1996). Terminology today. In *Terminology, LSP and Translation. Studies in language engineering in honour of Juan C. Sager.* Amsterdam. John Benjamins Publishing Company, p.15-33.

Cabré, Maria Teresa (1999-a). *La terminología: Representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos.* Barcelona, Institut Universitari de Lingüística Aplicada.

Cabré, Maria Teresa (1999-b). *Terminology – Theory, methods and applications*. Amsterdam. John Benjamins Publishing Company.

Cabré, Maria Teresa (2000). Elements for a theory of terminology: Towards an alternative paradigm. *Terminology.* Amsterdam. John Benjamins Publishing Company, V. 6, No.2, p.35-57.

Cabré, Maria Teresa (2003). Theories of terminology. Their description, prescription and explanation. *Terminology.* Amsterdam. John Benjamins Publishing Company,V. 9, No.2, p.163-199.

Cabré, Maria Teresa, Carme Bach, Rosa Estopa, Judit Feliu, Gemma Martinez, Jorge Vivaldi (2004). The GENOMA-KB project: towards the integration of concepts, terms, textual corpora and entities. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC).* Lisbon, Portugal.

Cabré, Maria Teresa, Anne Condamines and Fidelia Ibekwe-SanJuan (2005). Introduction: Application-driven terminology engineering. *Terminology.* Amsterdam. John Benjamins Publishing Company, V. 11, No.1, p.1-19.

Campo, Angela and Monique Cormier (2005). The Role of the Communicative Approach in the Development of Terminology. *Meta: Translators' Journal.* Montreal, Canada. Les Presses de l'Université de Montréal. V. 50, No. 4.

Cao, Jing (2011). *A corpus-based study of adjectives in contemporary English*. PhD Thesis. City University of Hong Kong.

Cermak, Frantisek (2003). Source materials for dictionaries. In *A Practical Guide to Lexicography*. Piet van Sterkenburg, Ed. Amsterdam. John Benjamins Publishing Company, p.18-25.

Champagne, Guy (2004). *The Economic Value of Terminology. An Exploratory Study.* Ottawa. Translation Bureau of Canada.

Chomsky, Noam (1965). *Aspects of the Theory of Syntax*. Cambridge, MA. MIT Press.

Chung, Teresa Mihwa (2003). A corpus comparison approach for terminology extraction. *Terminology.* Amsterdam. John Benjamins Publishing Company, V. 9, No.9, p.221-245.

Church, Kenneth and Patrick Hanks (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics,* V. 16, No. 1, p.22-29.

Collet, Tanja (2004). What's a term? An attempt to define the term within the theoretical framework of text linguistics. *Linguistica Antverpiensia New Series, NS3 - The Translation of Domain Specific Languages and Multilingual Terminology Management.* Manchester, U.K. St. Jerome Publishing, V. 3, p.99-111.

Condamines, Anne (1995). Terminology: New needs, new perspectives. *Terminology.* Amsterdam. John Benjamins Publishing Company, V. 2, No. 2, p.219-238.

Condamines, Anne (2005). Linguistique de corpus et terminologie. *Langages, 157,* L. Depecker, Ed. p.36-47.

Condamines, Anne (2007a). Corpus et terminologie. *La redocumentarisation du monde*. R.T. Pédauque, Ed. Toulouse: Cepadues Editions, p.131-147.

Condamines, Anne (2007b). L'interprétation en sémantique de corpus: le cas de la construction de terminologies. *Revue Française de Linguistique Appliquée*: *Corpus: état des lieux et perspectives.* V. XII-1, p.39-52.

Condamines, Anne (2008a). Taking genre into account when analysing conceptual relation patterns. *Corpora*. V. 8, p.115-140.

Condamines, Anne (2008b). Peut-on prévenir le risque langagier dans la communication écrite en entreprise? *Langage et Société : Le risque du langage en situation de travail*, N. 125, p.77-91.

Condamines, Anne (2010). Variations in terminology. Application to the management of risks related to ranguage use in the workplace. *Terminology.* Amsterdam. John Benjamins Publishing Company, V. 16, No. 1, p.30-50.

Corbolante, Licia and Ulrike Irmler (2001). Software Terminology and Localization. *Handbook of Terminology Management.* V. 2. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.516-535.

Daille, Béatrice (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Doctoral Thesis. Université Paris 7.

Daille, Béatrice (2005). Variations and application-oriented terminology engineering. *Terminology*. Amsterdam. John Benjamins Publishing Company, p.181-197.

Daille, Béatrice (2007). Variations and application-oriented terminology engineering. In *Application-Driven Terminology Engineering*. Fidelia Ibekwe-SanJuan, Anne Condamines and M. Teresa Cabré Castellvi, Eds. Amsterdam. John Benjamins Publishing Company, p.163-177.

Daille, Béatrice, Benoit Habert, Christian Jacquemin and Jean Royauté (1996). Empirical observation of term variations and principles for their description. *Terminology.* Amsterdam. John Benjamins Publishing Company, V. 3, No.2, p.197-257.

De Bessé, Bruno and Donatella Pulitano (1996). Which terms should firms or organisations include in their terminology banks? The case of the Canton of Berne. In *Terminology, LSP and Translation. Studies in language engineering in honour of Juan C. Sager*. Amsterdam. John Benjamins Publishing Company, p.35-46.

De Bessé, Bruno (1997). Terminological Definitions. *Handbook of Terminology Management.* V. 1. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.63-74.

De Saussure, Ferdinand (1916). *Cours de linguistique générale.* Paris. Éditions Payot et Rivages. (Republished in 1995).

Ditlevsen, Marianne Grove (2011). Towards a Methodological Framework for Knowledge Communication. In *Current Trends in LSP Research*. Margrethe Petersen and Jan Engberg, Eds. Bern. Peter Lang, p.187-208.

Drouin, Patrick (2002). *Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés*. Doctoral Thesis. Montreal. Université de Montréal.

Drouin, Patrick (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology.* Amsterdam. John Benjamins Publishing Company. V. 9, No. 1, p.99-115.

Drouin, Patrick, M.C. L'Homme and C. Lemay (2005). Two methods for extracting "specific" single-word terms from specialized corpora: Expirimentation and evaluation. *International Journal of Corpus Linguistics.* Amsterdam: John Benjamins Publishing Company, V. 10, No. 2, p.227-255.

Dubuc, Robert (1992). *Manuel pratique de terminologie.* Quebec, Canada. Linguatech.

Dubuc, Robert (1997). *Terminology: A Practical Approach*. Quebec, Canada. Linguatech.

Dubuc, Robert and Andy Lauriston (1997). Terms and Contexts. *Handbook of Terminology Management.* V. 1. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.63-74.

Engwall, Gunnel (1994). Not Chance but Choice: Criteria in Corpus Creation. *Computational Approaches to the Lexicon*. B.T.S. Atkins and A. Zampolli, Eds.Oxford, U.K. Oxford University Press, p.49-82.

Evert, Stefan (2004). *Association Measures*. http://www.collocations.de/AM/

Evert, Stefan (2008). Corpora and collocations. *Corpus Linguistics: An International Handbook*. A. Ludeling and M. Kyoto, eds. Berlin. Mouton de Gruyter.

Felber, Helmut. (1984). *Terminology Manual*. Vienna. Infoterm.

Fidura, Christie (2013). *Terminology Matters*. White paper published by SDL Inc. Available from: http://www.sdl.com/search/?query=terminology+matters

Flowerdew, Lynne (2004). The argument for using English specialized corpora. Discourse in the Professions. Perspectives from corpus linguistics. Ulla Connor and Thomas A. Upton, Eds. Amsterdam. John Benjamins Publishing Company, p.11-33.

Freixa, Judit (2006). Causes of denominative variation in terminology. A typology proposal. *Terminology.* Amsterdam. John Benjamins Publishing Company. V. 12, No. 1, p.51-77.

Fuertes-Olivera, Pedro and Ascension Arribas-Bano (2008). *Pedagogical Specialised Lexicography*. Amsterdam. John Benjamins Publishing Company.

Galinski, Christian (1994). Exchange of standardized terminologies within the framework of the standardized terminology exchange network. *Standardizing and Harmonizing Terminology: Theory and Practice*. ASTM STP 1223. Sue Ellen Wright and Richard A. Strehlow, Eds. Philadelphia, USA. American Society for Testing and Materials, p.141-149.

Gilreath, Charles (1994). The semantic valence of terms: a systematic treatment of multi-meaning terms. *Standardizing and Harmonizing Terminology: Theory and Practice.* Philadelphia, USA. ASTM STP 12233. American Society for Testing and Materials.

Gopferich, Susanne (2000). Analysing LSP Genres (Text Types): From Perpetuation to Optimization in Text(-type) Linguistics. In *Analysing Professional Genres*. Anna Trosborg, Ed. Amsterdam. John Benjamins Publishing Company, p.227-247.

Greenwald, Susan (1994). A construction industry terminology database developed for use with a periodicals index. *Standardizing and Harmonizing Terminology: Theory and Practice*. ASTM STP 1223. Sue Ellen Wright and Richard A. Strehlow, Eds. Philadelphia, USA. American Society for Testing and Materials, p.115-125.

Hajutin, A.D. (1978). *Les diverses orientations du travail terminologique.* Translation of: *O razlicnyh napravlenijah v terminologiceskoj rabote*. Quebec, Canada. GIRSTERM, Université Laval.

Halliday, M.A.K. (1985). *Spoken and written language*. Oxford, U.K. Oxford University Press.

Halliday, M.A.K. and Ruqaiya Hasan (1976). *Cohesion in English*. London and New York. Longman Publishing Group.

Halliday, M.A.K., and Webster, Jonathan, Eds. (2009). *Continuum Companion to Systemic Functional Linguistics.* London, U.K. Continuum International Publishing Group.

Hanks, Patrick (2013). *Lexical Analysis - Norms and Exploitations.* London, U.K. The MIT Press.

Hoffman, Lothar. (1979). Towards a theory of LSP. Elements of a methodology of LSP analysis. Vienna, Austria. *Fachsprache, International Journal of Specialized Communication,* V. 1, No 2, p.12-17.

Hoffmann, Sebastian, Stefan Evert, Nicholas Smith, David Lee, Ylva Berglund Prytz (2008). *Corpus Linguistics with BNCWeb – a Practical Guide*. Peter Lang GmbH, Frankfurt, Germany.

Hunston, Susan (2002). *Corpora in Applied Linguistics*. Cambridge, U.K. Cambridge University Press.

Ibekwe-SanJuan, Fidelia (1998). *Terminological variation, a means of identifying research topics from texts*. COLING '98 Proceedings of the 17th International Conference on Computational Linguistics, V1, p.564-570

Ibekwe-SanJuan, Fidelia, Anne Condamines and M. T. Cabré Castellvi (2007). *Application-Driven Terminology Engineering*. Fidelia Ibekwe-SanJuan, Anne Condamines and M. Teresa Cabré Castellvi, Eds. Amsterdam. John Benjamins Publishing Company.

ISO TC37, SC3 (1999). *ISO 12620:1999. Terminological Data Categories.* Geneva. International Organization for Standardization. This standard is currently being replaced by the ISO Data Category Registry, ISOCat (http://www.isocat.org/).

ISO TC37, SC1 (2000). *ISO 1087-1:2000. Terminology work - Vocabulary - Part 1: Theory and application.* Geneva. International Organization for Standardization.

ISO TC37, SC3 (2004). *ISO 16642:2004. Terminological Markup Framework (TMF)*. Geneva. International Organization for Standardization.

ISO TC37, SC3 (2008). *ISO 30042:2008. TermBase eXchange (TBX)*. Geneva. International Organization for Standardization.

Jacquemin, Christian (2001). *Spotting and Discovering Terms through Natural Language Processing.* Cambridge, U.K. The MIT Press.

Justeson, John and Slava Katz (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*. Cambridge, U.K. Cambridge University Press, V. 1, No. 1, p.9-27.

Kageura, Kyo (1995). Toward the theoretical study of terms. *Terminology*. Amsterdam. John Benjamins Publishing Company, V. 2, No. 2, p.239-257.

Kageura, Kyo (2002). *The Dynamics of Terminology. A descriptive theory of term formation and terminological growth*. Amsterdam. John Benjamins Publishing Company.

Karsch, Barbara, Alma Kharrat, Robin Lombard and Masaki Itagaki (2008). *TKE Workshop: Development of Enterprise-level Ontology and Lexical Resource Solutions*. Copenhagen, Denmark.

Kelly, Natalie and Donald DePalma (2009). *The Case for Terminology Management*. Common Sense Advisory, Inc.

Kennedy, Chris and Rod Bolitho (1984). *English for specific purposes*. London. Macmillan.

Kerremans, Koen (2010). A Comparative Study of Terminological Variation in Specialised Translation. *Reconceptualizing LSP*. Online proceedings of the XVII European LSP Symposium 2009. Heine, Carmen and Jan Engberg, Eds. Aarhus.

Kit, Chunyu and Xiaoyue Liu (2008). Measuring mono-word termhood by rank difference via corpus comparison. *Terminology*. Amsterdam. John Benjamins Publishing Company. V. 14, No. 2, p.204-229.

Kittredge, Richard and John Lehrberger (1982). *Sublanguage. Studies of Language in Restricted Semantic Domains*. Berlin. Walter de Gruyter.

Knops, Eugenia and Gregor Thurmair (1993). Design of a Multifunctional Lexicon. *Terminology: Applications in interdisciplinary communication.* Sonneveld, Helmi B. and Kurt L. Loening, Eds. Amsterdam. John Benjamins Publishing Company, p.87-109.

Kocourek, Rostislav (1982). *La langue française de la technique et de la science*. La Documentation Française, Paris. Weisbaden. Oscar Brandstetter Verlag Gmbh & Co.

Korkas, Vassilis and Margaret Rogers (2010). How much terminological theory do we need for practice? An old pedagogical dilemma in a new field. *Terminology in Everyday Life*. Thelen, Marcel and Steurs, Frieda, Eds. Amsterdam. John Benjamins Publishing Company, p.123-136.

Kubler, Natalie and Cecile Frérot (2003). Verbs in specialised corpora: from manual corpus-based description to automatic extraction in an English-French parallel corpus. *Technical Papers - University Centre for Computer Corpus Research on Language*. V. 16, p.429-438.

Lam Kam-mei, Jacqueline (2001). *A Study of Semi-Technical Vocabulary in Computer Science Texts, with Special Reference to ESP Teaching and Lexicography*. Hong Kong. Language Centre. The Hong Kong University of Science and Technology.

L'Homme, Marie-Claude (1998). Le statut du verbe en langue de spécialité et sa description lexicographique. *Cahiers de lexicologie. Institut de linguistique française.* Paris, France. V. 73, No. 2, p.61-84.

L'Homme, Marie-Claude (2002). What can verbs and adjectives tell us about terms? *Proceedings of the Terminology and Knowledge Engineering conference*. Nancy, France, p.65-70.

L'Homme, Marie-Claude (2003). Capturing the Lexical Structure in Special Subject Fields with Verbs and Verbal Derivatives: A model for specialized lexicography. *International Journal of Lexicography*, V. 16, No. 4, p.403-422.

L'Homme, Marie-Claude (2004). *La terminologie : principes et techniques*. Montreal, Canada. Les Presses de l'Université de Montréal.

L'Homme, Marie-Claude (2005). Sur la notion de “terme.” *Meta: Translators' Journal.* Montreal, Canada. Les Presses de l'Université de Montréal, V. 50, No. 4, p.1112-1132.

L'Homme, Marie-Claude (2006). The processing of terms in dictionaries: New models and techniques. *Terminology*. Amsterdam. John Benjamins Publishing Company, V. 12, No.2, p.181-188.

L'Homme, Marie-Claude and Elizabeth Marshman (2006). Terminological Relationships and Corpus-Based Methods for Discovering Them: An Assessment for Terminographers. *Lexicography, Terminology, and Translation. Text-based studies in honour of Ingrid Meyer.* Lynne Bowker, Ed. Ottawa, Canada. University of Ottawa Press.

Landau, Sidney (2001). *Dictionaries. The Art and Craft of Lexicography*. Cambridge, U.K. Cambridge University Press.

Lara, Luis Fernando (1998). Concepts and Term Hierarchy. *Terminology.* Amsterdam. John Benjamins Publishing Company, V. 5, No. 1, p.59-76.

Li, H (2011). *A Computational Approach to English Core Vocabulary Based on the British National Corpus*. PhD Dissertation, Department of Chinese, Translation and Linguistics, City University of Hong Kong.

Localization Industry Standards Association (2005). *Translation Memory Exchange (TMX)*. Available from: http://www.gala-global.org/oscarStandards/tmx/tmx14b.html

Localization Industry Standards Association (2009). *TBX-Basic*. Available from: http://www.gala-global.org/oscarStandards/tbx/tbx-basic.html

Lombard, Robin (2006). Managing source language terminology. In *Perspectives on Localization*. Keiran J. Dunne, Ed. Amsterdam. John Benjamins Publishing Company, p.155-171.

Lommel, Arle (2005). *LISA Terminology Management Survey. Terminology Management Practices and Trends.* Geneva. Localization Industry Standards Association.

Lommel, Arle and Rebecca Ray (2007). *Creating Global Content.* Geneva. Localization Industry Standards Association.

Marshman, Elizabeth and Patricia Van Bolderen (2008). *Towards an integrated analysis of aligned texts in terminology: The CREATerminal approach.* First International Workshop on Terminology and Lexical Semantics. Montreal, Canada. Observatoire de linguistique Sens-Texte (OLST), p.42-53. Available from: http://olst.ling.umontreal.ca/pdf/ProceedingsTLS09.pdf

Martin, Ronan (2011). *Term Inclusion Criteria.* Internal SAS document, SAS Inc., Cary, N.C.

Martin, Willy and Hennie van der Vliet. Design and production of terminological dictionaries. *A Practical Guide to Lexicography*. Piet van Sterkenburg, Ed. Amsterdam. John Benjamins Publishing Company, p.333-349.

Maynard, Diana and Sophia Ananiadou (2001). Term extraction using a similarity-based approach. *Recent Advances in Computational Terminology*. Didier Bourigault, Christian Jacquemin, Marie-Claude L'Homme, Eds. Amsterdam. John Benjamins Publishing Company, p.261-278.

Maurais, Jacques (1993). Terminology and Language Planning. *Terminology: Applications in interdisciplinary communication*. Helmi B. Sonneveld and Kurt L. Loening, Eds. Amsterdam. John Benjamins Publishing Company, p.111-125.

McEnery, Tony and Andrew Hardie (2012). *Corpus Linguistics: Method, Theory and Practice.* Cambridge, U.K. Cambridge University Press.

Mel'čuk, I., A. Clas and A. Poliguère (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve, Belgium. Duculot.

Meyer, Ingrid (1993). Concept management for terminology: A knowledge engineering approach. *Standardizing Terminology for Better Communication: Practice, Applied Theory, and Results*. Richard Alan Strehlow, Sue Ellen Wright, Eds. Philadelphia. American Society for Testing and Materials. ASTM STP 1166, p.140-151.

Meyer, Ingrid and Kristen Mackintosh (1996). The Corpus from a Terminographer`s View-point. *International Journal of Corpus Linguistics*. Amsterdam. John Benjamin's Publishing Company, V. 1, no 2, p.257-285.

Meyer, Ingrid and Kristen Mackintosh (2000). When terms move into our everyday lives: An overview of de-terminologization. *Terminology*. Amsterdam. John Benjamin's Publishing Company, V. 6, no 1, p.111-138.

Myking, Johan (2007). No fixed boundaries. *Indeterminacy in Terminology and LSP.* Bassey Edem Antia, Ed. Amsterdam. John Benjamin's Publishing Company, p.73-91.

Nagao, Makoto (1994). A Methodology for the Construction of a Terminology Dictionary. *Computational Approaches to the Lexicon*. B.T.S. Atkins and A. Zampolli, Eds. Oxford, U.K. Oxford University Press, p.397-411.

Nation, I.S.P. (1990). *Teaching and learning vocabulary*. New York, N.Y. Newbury House.

Nazarenko, Adeline and Touria Ait El Mekki (2007). Building back-of-the-book indexes? *Application-Driven Terminology Engineering*. Fidelia Ibekwe-SanJuan, Anne Condamines and M. Teresa Cabré Castellvi, Eds. Amsterdam. John Benjamins Publishing Company, p.199-224.

Nkwenti-Azeh, Blaise (2001). User-specific Terminological Data Retrieval. *Handbook of Terminology Management.* V. 2. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.600-613.

Oakes, Michael and Chris Paice (2001). Term extraction for automatic abstracting. *Recent Advances in Computational Terminology*. Didier Bourigault, Christian Jacquemin, Marie-Claude L'Homme, Eds. Amsterdam. John Benjamins Publishing Company, p.353-370.

Park, Youngja, Roy J. Byrd and Branimir K. Boguraev (2002). Automatic Glossary Extraction: Beyond Terminology Identification. *Proceedings of the 19th international conference on computational linguistics*, V. 1. Pennsylvania, USA. Association for Computational Linguistics.

Pavel, Sylvia (1993). Neology and Phraseology as Terminology-in-the-Making. *Terminology: Applications in interdisciplinary communication*. Helmi B. Sonneveld and Kurt L. Loening, Eds. Amsterdam. John Benjamins Publishing Company, p.21-33.

Pearson, Jennifer (1998). *Terms in Context – Studies in Corpus Linguistics*. Amsterdam. John Benjamins Publishing Company.

Picht, Heribert and Jennifer Draskau (1985). *Terminology: An Introduction.* Denmark. LSP Centre, Copenhagen Business School.

Picht, Heribert and Carmen Acuna Partal (1997). Aspects of Terminology Training. *Handbook of Terminology Management.* V. 1. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.63-74.

Prinsloo, D.J. (2009). The role of corpora in future dictionaries. In Lexicography in the 21st Century. Sandro Nielsen and Sven Tarp, Eds. Amsterdam. John Benjamins Publishing Company, p.181-206.

Pozzi, Maria (1996). Quality assurance of terminology available on the international computer networks. In *Terminology, LSP and Translation. Studies in language engineering in honour of Juan C. Sager.* Amsterdam. John Benjamins Publishing Company, p.67-82.

Rey, Alain (1995). *Essays on Terminology.* Amsterdam. John Benjamins Publishing Company.

Riggs, Fred (1989). Terminology and lexicography: their complementarity. *International Journal of Lexicography*, V. 22, p.89-110.

Riggs, Fred (1994). The Representation of concept systems. *Standardizing and Harmonizing Terminology: Theory and Practice.* Philadelphia. ASTM STP 12233. American Society for Testing and Materials.

Riggs, Fred, Matti Malkia and Gerhard Budin (1997). Descriptive Terminology in the Social Sciences. *Handbook of Terminology Management.* V. 1. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.184-196.

Rinaldi, Fabio, James Dowdall, Michael Hess, Kaarel Kaljurand, and Magnus Karlsson (2003). *The Role of Technical Terminology in Question Answering*. In Proceedings of TIA-2003 - Terminologie et Intelligence Artificielle, Strasbourg.

Roche, Christophe (2012). Should terminology principles be re-examined? *Proceedings of the 10th Terminology and Knowledge Engineering Conference* (TKE 2012). Aguado de Cea et al., Eds., p.17-32.

Roche, Christophe (2012). Terminologie conceptuelle versus Terminologie textuelle. *Repères*. France. Équipe Condillac. No. 1.

Rogers, Margaret (1997). Synonymy and equivalence in special-language texts. In *Text Typology and Translation*. Anna Trosborg, Ed. Amsterdam. John Benjamins Publishing Company, p.217-246.

Rogers, Margaret (2000). Genre and Terminology. In *Analysing Professional Genres*. Anna Trosborg, Ed. Amsterdam. John Benjamins Publishing Company, p.3-21.

Rogers, Margaret (2007). Lexical chains in technical translation. A case study in indeterminacy. In *Indeterminacy in Terminology and LS*P. Bassey Edem Antia, Ed. Amsterdam. John Benjamins Publishing Compmany, p.15-35.

Rondeau, Guy (1981). *Introduction à la terminologie*. Montreal, Canada. Centre educatif et culturel Inc.

Sager, Juan, David Dungworth and Peter F. McDonald (1980.) *English Special Languages. Principles and Practice in Science and Technology*. Wiesbaden. Brandstetter-Verlag.

Sager, Juan (1990). *A Practical Course in Terminology Processing*. Amsterdam. John Benjamins Publishing Company.

Sager, Juan (1994). What's wrong with “terminology work” and “terminology science”? *Terminology*. Amsterdam. John Benjamins Pulishing Company, V. 1, No. 2, p.375-381.

Sager, Juan (2001). Terminology Compilation: Consequences and Aspects of Automation. *Handbook of Terminology Management.* V. 2. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.761-771.

Sanchez, Maribel Tercedor (2011). The cognitive dynamics of terminological variation. *Terminology*. Amsterdam. John Benjamins Publishing Company, V. 17, No. 2, p.181-197.

Schmitz, Klaus-Dirk and Daniela Straub (2010). *Successful terminology management in companies*. Stuttgart. TC and more GmbH.

Shreve, Gregory (2001). Terminological Aspects of Text Production. *Handbook of Terminology Management.* V. 2. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.772-787.

Schubert, Klaus (2011). Specialized Comunication Studies. *Current Trends in LSP Research*. Margrethe Petersen and J.Engberg, Eds. Bern, Switzerland. Peter Lang AG, p.19-58.

Seomoz (2012). *The Beginner's Guide to SEO*. Available at: http://www.seomoz.org/beginners-guide-to-seo

Sinclair, John (1991). *Corpus, Concordance, Collocation*. Oxford, U.K., Oxford University Press.

Sinclair, John (2003). Corpora for lexicography. In *A Practical Guide to Lexicography.* Piet van Sterkenburg, Ed. Amsterdam. John Benjamins Publishing Company, p.167-178.

Sinclair, John, S. Jones, R. Daley and R. Krishnamurthy (2004). *English Collocation Studies: The OSTI Report*. London, U.K. Continuum.

Sinclair, John, S. Jones and R. Daley (1970). *English Lexical Studies*. OSTI Report. University of Birmingham, U.K.

Slodzian, Monique (2000). L'émergence d'une terminologie textuelle et le retour du sens. In *Le Sens en Terminologie*. Henri Béjoint, Philippe Thoiron, Eds. Lyon, France: Presses Universitaires Lyon, p.61-85.

Strehlow, Richard (2001-a). Terminology and Indexing. *Handbook of Terminology Management*. V. 2. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company. p.419-425.

Strehlow, Richard (2001-b). The Role of Terminology in Retrieving Information. *Handbook of Terminology Management*. V. 2. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.426-444.

Swales, John (1990). *Genre Analysis. English in academic and research settings*. Cabridge, UK. Cambridge University Press.

Temmerman, Rita (1997). Questioning the univocity ideal. The difference between socio-cognitive Terminology and traditional Terminology. *Hermes, Journal of Linguistics*. Aarhus Universitet. No. 18, p.51-90.

Temmerman, Rita (2000). *Towards New Ways of Terminology Description: The socio-cognitive approach*. Amsterdam. John Benjamins Publishing Company.

Temmerman, Rita and Marc Van Campenhoudt (2001). The dynamics of terms in specialized communication: An interdisciplinary perspective. *Terminology.* Amsterdam. John Benjamins Publishing Company, V. 17, No. 1, p.1-8.

TerminOrgs (2014). *TBX-Basic*. Available at www.terminorgs.net.

Temmerman, Rita, Peter De Baer, and Koen Kerremans (2010). Competency-based job descriptions and Termontography. The case of terminological variation. In *Terminology in Everyday Life*. Thelen and Frieda Steurs, Eds. Amsterdam. John Benjamins Publishing Company, p.179-191.

Teubert, Wolfgang (2005). Language as an economic factor: the importance of terminology. *Meaningful Texts*. Geoff Barnbrook, Pernilla Danielsson and Michaela Mahlberg, Eds. London, U.K. Continuum, p.96-106.

Thomas, Patricia (1993). Choosing headwords from language-for-special-purposes (LSP) collocations for entry into a terminology data bank (term bank). *Terminology: Applications in interdisciplinary communication*. Helmi B. Sonneeld and Kurt L. Loening, Eds. Amsterdam. John Benjamins Publishing Company, p.43-68.

Thurow, Shari (2006). *The Most Important SEO Strategy.* Available from: http://www.clickz.com/clickz/column/1717475/the-most-important-seo-strategy

Tong, K.S.T. (1993). *Investigating the sub-technical vocabulary in Computer Science text through computer corpus building and concordancing techniques*. Paper presented at the 10th World Congress of the International Association of Applied Linguistics, Amsterdam.

Trimble, Louis (1985). *English for science & technology: A discourse approach*. Cambridge. Cambridge University Press.

Van Campenhoudt, Marc (1999). *Terminologie descriptive: Petite initiation à l'exploitation de corpus*. Communication présentée dans le cadre de la 8e Université d'automne en terminologie Université de Rennes II.

Van Campenhoudt, Marc (2002). Lexicographie vs terminographie : quelques implications théoriques du projet DHYDRO. In *Travaux du Lilla*. H. Zinglé, Ed. Université de Nice-Sophia Antipolis, No. 4, p.91-103.

Van Campenhoudt, Marc (2006). *Que nous reste-t-il d'Eugen Wuster?* Intervention dans le cadre du colloque international Eugen Wuster et la terminologie de l'École de Vienne. Paris. Université de Paris7.

Van Sterkenburg, Piet (Ed) (2003). *A Practical Guide to Lexicography*. Amsterdam. John Benjamins Publishing Company.

Varantola, Krista (2003). Linguistic corpora (databases) and the compilation of dictionaries. In *A Practical Guide to Lexicography*. Piet van Sterkenburg, Ed. Amsterdam. John Benjamins Publishing Company, p.228-239.

Warburton, Kara (2001a). *Terminology Management in the Localization Industry – Results of the LISA Terminology Survey*. Geneva. Localization Industry Standards Association.

Available from: http://www.terminorgs.net/downloads/LISAtermsurveyanalysis.pdf

Warburton, Kara (2001b). Globalization and Terminology Management. *Handbook of Terminology Management,* V. 2. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.677-696.

Wettengel, Tanguy and Aidan Van de Weyer (2001). Terminology in Technical Writing. *Handbook of Terminology Management,* V. 2. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.445-466.

Williams, Malcolm (1994). Terminology in Canada. *Terminology.* Amsterdam. John Benjamin's Publishing Company, Vol 1, No 1, p.195-201.

Woyde, Rick (2005). Introduction to SAE J1930: Bridging the Disconnect Between the Engineering, Authoring and Translation Communities. *Globalization Insider.* Geneva. Localization Industry Standards Association. Available from: http://www.translationdirectory.com/article903.htm

Wright, Sue Ellen (1997). Term Selection: The Initial Phase of Terminology Management. *Handbook of Terminology Management*, V. 1. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.13-23.

Wright, Sue Ellen (2001a). Terminology and Total Quality Management. *Handbook of Terminology Management,* V. 2. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.488-502.

Wright, Sue Ellen (2001b). Terminology as an Organizational Principle in CIM Environments. *Handbook of Terminology Management,* V. 2. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.467-479.

Wright, Sue Ellen and Leland Wright (1997). Terminology Management for Technical Translation. *Handbook of Terminology Management*, V. 1. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.147-159.

Wright, Sue Ellen and Gerhard Budin (1997). Infobox No. 2: Terminology Activities. *Handbook of Terminology Management*, V. 1. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.327.

Wright, Sue Ellen and Gerhard Budin (2001). Infobox No. 31: Human Language Technologies and Language Engineering. *Handbook of Terminology Management*, V. 2. Sue Ellen Wright and Gerhard Budin, Eds. Amsterdam. John Benjamins Publishing Company, p.887.

Wüster, Eugen (1979). *Introduction to the General Theory of Terminology and Terminological Lexicography*. (Translation.) Vienna. Springer.

Wüster, Eugen (1967). *Grundbegriffe bei Werkzeugmaschinen*. London. Technical Press.

APPENDIX A – Words and expressions removed from the Minitab termbase

The following words and expressions were removed from Minitab's termbase. Most are general lexicon words and expressions but there are also some code strings and other non-terms.

# (symbol)	examine	observed value
% (symbol)	example	obtain
& (symbol)	exceed	obvious
* (symbol)	excellent	occur
+ (symbol)	except	often
/ (symbol)	exceptional	old
< (symbol)	exceptionally	omit
<= (symbol)	explain	on
<> (symbol)	extent	on hand
> (symbol)	extra	on the other hand
>= (symbol)	extremely	once
?	fairly	one
a	fall	one time
a bit	familiar	optionally
a level	farther	or
a lot	fast	order
a single	find	organization
ability	find out	organize
able	finish	otherwise
abnormal	finished	out
abnormally	first	outward
about	fit	outwards
absolutely	fits	over and over
accomplish	fix	p
accomplished	fixed	participate
according to	for example	particular
account for	for instance	particularly
activate it	for that matter	past
active	for the most part	perform
additional	for this reason	performed
additional considerations	former	period
additionally	frequently	period of time
adequate	full	permit
adequately	fully	permitted
adjacent	further	person
adjoining	generally	place

adjust	get	plus
advise	go	poor
affect	good fit	precision
afterward	good idea	predict
again and again	gray	preferable
agreement	great	preferably
ahead of time	greater	preparation
all	greater than	prescribed
allocate	greatest	present
allocation	greatly	pretty
allow	grey	prevent
allow you	happen	previous
allowable	help	primarily
allowed	hence	primary
almost	here are	prior
also	Hi	prior to
alternately	high	proceed
Alternative	higher	process
altogether	highest	produce
among	highlight	proper
amount	highlighted	properly
and/or	highly	put
another	i.e.	quantity
answer	idea	quickly
anytime	identification	quite
appear	identified	quite a
appropriate	identify	rapid
approximately	ie	rapidly
are	illustrate	rather
are related	Figure	rather than
arrange	image	reasonable
arrangement	immediately	reasonably
arrival	important considerations	recommended
arrive at	impose	reduce
as is	impossible	reflect
as long as	impression	regarding
as well as	improper	relate
assign	improve	relatively
assist	improve upon	reliable
assistance	in	remain
associated with	in addition	remember
assumption	in advance	represent
at most	in control	represented
at once	in general	respective
at the same time	in many cases	respectively
average	in most cases	rest

B/W	in order to	result
backward	in other words	result in
backwards	in place of	resume
badly	in practice	rough
basics	in some cases	roughly
be	in steps of	run the risk
been	in that case	safely
before	in this case	sales
beforehand	inadequate	search
begin	include	see
beginning	incorporate	seek
benefit	incorrect	seem
beside	increase	select
besides	indicate	selected
better	indicated	seriously
between	indicates	service
beyond	indication	set
big	individual	set up
bigger	information	setting
biggest	information about	settings
bring	information on	settle
by hand	initial	severely
can	instead	should
can not	instead of	show
cancel	insufficient	similar
cannot	intact	simply
carry out	inside	simultaneously
case	interfere	since
case in point	invalid	size
cause	investigate	small
center	investigation	so that you
centered	is advisable	so you
chance	is generally recommended	sometimes
change	is recommended	specific
changing	is strongly recommended	specified
chapter	it is advisable to	specify
check	just	spend time
choose	keep	standardize
choose to	keep in mind	start
clean it up	kind	start value
clean up	large	starting
common	larger	starting value
comparable	largest	stay
comparatively	last	stop
compare	later	sudden
comparison	leave	sufficient

complete	leave out	sufficiently
completed	left	suggest
completely	less than	take
complexity	let	take a sample
comprised of	let Minitab	take action
concentrate	let you	take advantage
concept	likewise	take appropriate action
concern	limit	take corrective action
concurrently	limitation	take effect
conduct	limits	take full advantage
conducted	little	take immediate steps
confirm	locate	take into account
confounded	location	take into consideration
confounding	long term	take on
consequently	long-term	take part
consider	look at	take place
considerable	main	take steps
considerably	main effect	take the absolute value
considerations	mainly	take the average
considered	maintain	take the diagonal
consist of	make	take the form
continue	make a decision	take the function of
continue to function	make certain	take the square root
correct	make conclusions	take time
correctly	make easier	takes time
could	make it active	tell
create a chart	make sense	test
create a cluster	make sure	that is
create a graph	makes sense	then
create a plot	manner	therefore
current	many	think of
current time	match	thought of
custom	maximize	thus
customer	maximum	time interval
customization	may	time period
customize	medium	time span
customized	middle	to some extent
decide	midpoint	topic
decrease	might	totally
decreasing	minimize	try
delete	minimum	turn on
designate	Minitab	turned off
designated	moderate	typical
desire	moderately	typically
determine	modification	unacceptable
determine if	modify	uncommon

determine whether	more	understand
did not	most	unusual
didn't	most of the time	unusual observation
difference	mostly	unusual value
discontinue	move	unusually
discover	much	update
display	n	upon
distinct	N/A	use
distinction	near	user
distinguish	nearly	usual
divide into	neither	usually
do	never	valid
document	next	value
documentation	next to	variability
done	No	variable
due to	normal	variation
e	not	verify
each	not acceptable	version
each week	not applicable	versus
early	not correct	very
eg	not enough	view
eliminate	not equal to	want
emphasize	not possible	way
employ	not sufficient	we
employees	not symmetrical	weekly
endure	not valid	when
enough	note	when possible
ensure	note that	whenever possible
entire	notify	whether
entirely	number	whole
especially	objective	will
etc	observation	wish
etc.	observations	within
every	observe	wonderful
evident	observed units	would

APPENDIX B – Words and expressions removed from the Symantec termbase

The following words and expressions were removed from Symantec's termbase. Most appear to be from the general lexicon but a few are obvious mistakes or non-English words.

- ????? (this is a term from a language using a non-ASCII character set)
- availability
- browse to (the term “browse” was kept)
- complete
- content
- copy
- delete
- deletion
- eindpuntbeveiliging
- enter
- exclusion
- identification
- information
- item
- location
- low
- manage
- management
- multiple
- Norton
- OK
- removal
- response
- safe
- select
- skip
- speed
- successfully
- support (2 instances)
- task
- unwanted
- verify
- αθ?ρυβη λειτουργ?α

APPENDIX C – Data category description for the Minitab termbase

Concept data categories:

Data category	Type	Values / Comment
ID	Autogenerated value	
Instance	Drop-down list	Default Across Server
Template	Drop-down list	Default
Definition	Text field	
Source of Definition	Drop-down list	Beginner's Dictionary of American Usage, Learnersdictionary.com, Minitab Author, Minitab Editor, Minitab Statistical Glossary, Other, Wikipedia
Other Source of Definition	Text field	
Domain	Drop-down list	computing, general, Minitab, statistics
Sub-Domain	Drop-down list	Minitab: Minitab Statistical Software, Minitab: Qeystone Software
Sub-Sub-Domain	Drop-down list	Minitab Statistical Software: commands, Minitab Statistical Software: user interface, Qeystone Software: user interface
Image	File upload	
Subject	Drop-down list	(many values – Minitab publications)
Note	Text field	

Term data categories:

Data category	Type	Values / Comment
Term	Text field	
Language	Drop-down list	
Sublanguage	Drop-down list	
Released	Check box	
Register	Drop-down list	Bench, Jargon, Qeystone, Minitab, Neutral, Slang, Technical, Vulgar

Data category	Type	Values / Comment
Type	Drop-down list	Do Not Translate, Idiom, Minitab Command, Minitab Dialog Item, Minitab Dialog Title, Minitab Message, Minitab Other, Minitab Output, Minitab Session Command, Proprietary, Stock Phrase – General, Stock Phrase – Minitab, Symbol, Transcription, Transliteration
Form	Drop-down list	Acronym, Full, Short, Surface
Part of Speech	Drop-down list	Adjective, Adverb, Noun, Other Part of Speech, Proper Noun, Verb
Grammatical Number	Drop-down list	Dual, Non-Count, Other Number, Plural – Count, Singular – Count
Grammatical Gender	Drop-down list	Feminine, Masculine, Neutral, Other Gender
Usage Status	Drop-down list	Allowed, Preferred, Rejected, Constrained
Usage Note	Text field	
Related Terms	Text field	(Soon to be replaced by the SKOS data categories.)
Note	Text field	
ESL Friendly Vocabularies	Drop-down list	STE Approved, STE Not Approved, STE Not Included, Beginner's Dictionary Included, Beginner's Dictionary Not Included
Context Sentence 1	Text field	
Source of Context Sentence 1	Drop-down list	Minitab Help, Minitab Training, Minitab Chooser (a publication), Minitab Statistical Glossary
Source of Context	Text field	
Context Sentence 2	Text field	
Source of Context Sentence 2	Drop-down list	Minitab Help, Minitab Training, Minitab Chooser (which is a publication), Minitab Statistical Glossary
Gender	Drop-down list	Feminine, Masculine, Neuter, Other Gender
Project	Drop-down list	(many values)
Subject	Drop-down list	(many values)

APPENDIX D – Software used in the research

The following software tools were used to carry out the research:

- Wordsmith Tools, version 6.0. (http://www.lexically.net/wordsmith/)
- UltraEdit Professional Text/HEX Editor, Version 19
- oXygen XML Editor, version 14
- TermWeb Professional, Version 3.8.16. (http://www.interverbumtech.com/)

APPENDIX E – List of abbreviations

- CAT – computer-assisted translation
- CTT – Communicative Theory of Terminology
- ESP - English for Specific Purposes
- GTT – General Theory of Terminology
- LGP – language for general purposes
- LSP – language for special purposes
- MWU - multi-word unit
- MWT - multi-word term
- NLP – natural language processing
- SL - source language
- termbase – terminology database
- TL - target language
- TM – translation memory
- TMS – terminology management system

APPENDIX F – Calculation formulae for the collocate relationship measures

This appendix provides the calculation formulae for the various collocate relationship measures discussed in Section 7.6.

1. Log likelihood

$$\textbf{log-likelihood} = 2\sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$

Where O_{ij} = observed frequencies and E_{ij} = expected frequencies

2. Z-score

$$\textbf{z-score} = \frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$$

O_{11} = observed frequency of collocate

E_{11} = expected frequency of collocate

3. Specific Mutual Information

$$\operatorname{pmi}(x;y) \equiv \log \frac{p(x,y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)}.$$

$$P(x) = \frac{f(x)}{n}$$

Where p is probability, calculated as the frequency of a word, x, divided by the corpus size n.

4. Dice coefficient

$$\text{Dice} = \frac{2O_{11}}{R_1 + C_1}$$

O_{11} = observed frequency of collocate
R_1 = frequency of first term
C_1 = frequency of second term

5. MI3

$$\text{MI}^3 = \log \frac{(O_{11})^3}{E_{11}}$$

Where:

O_{11} = observed frequency of collocate
E_{11} = expected frequency of collocate

$$E_{11} = \frac{\text{frequency of 1st term x frequency of 2nd term}}{\text{sum of bigrams containing either term}}$$

6. T-Score

$$\text{t-score} = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$$

Where, O_{11} = observed frequency of collocate

E_{11} = expected frequency of collocate

APPENDIX G – Sample legal agreement

The following text is provided only as an example of a legal agreement for using corpora and terminology for research purposes. Each company providing such data requires its own customised agreement. The text below is not a legal document.

CONFIRMATION OF NON-DISCLOSURE OF CONFIDENTIAL INFORMATION

1. DISCLOSURE

1.1 In consideration of your disclosure to me of language data and other information (whether or not contained in documents) relating to this (in the following named Protected Material) for the purposes of [my research for my PhD thesis for the City University of Hong Kong] (Purpose), I will keep the Protected Material confidential. Accordingly, for a period of [three] years from the date of this letter, I shall not, without your prior written consent, either:
(a) communicate or otherwise make available the Protected Material to any third party; or
(b) use the Protected Material for any purpose other than the Purpose.

1.2 I may disclose the Protected Material to the minimum extent required by:
(a) any order of any court of competent jurisdiction or any competent judicial, governmental or regulatory body; or
(b) the rules of any listing authority or stock exchange on which our shares are listed or traded; or
(c) the laws or regulations of any country with jurisdiction over our affairs (provided, in the case of a disclosure under the [name of appropriate freedom of information act], none of the exemptions to that Act applies to the Protected Material disclosed).

2. LIMITATIONS ON OBLIGATIONS

The obligations set out in paragraph 1 shall not apply, or shall cease to apply, to such of the Protected Material as I can show to your reasonable satisfaction:

2.1 has become public knowledge other than through disclosure by me in breach of this agreement; or

2.2 was already known to me prior to disclosure by you; or

2.3 has been received by me from a third party who did not to my knowledge acquire it in confidence from you or from someone owing a duty of confidence to you.

3. RETURN OF THE PROTECTED MATERIAL

We shall, whenever you so request, return to you all documents and other records of the Protected Material or any of it in any form and whether or not such document or other record was itself provided by you.

4. GOVERNING LAW AND JURISDICTION

This letter and any dispute or claim arising out of or in connection with it or its subject matter or formation (including non-contractual disputes or claims) shall be governed by and construed in accordance with the law of [named jurisdiction]. The parties irrevocably agree that the courts of [named juristiction] shall have exclusive jurisdiction to settle any dispute or claim that arises out of or in connection with this letter or its subject matter or formation (including non-contractual disputes or claims).

Yours faithfully,

.....
[signature of researcher]

We hereby acknowledge receipt and accept the contents of this letter

Signed .....
For and on behalf of [name of participating company]

Date .....