Corpus-Based Analysis of Verb/Noun Collocations in Interdisciplinary Registers

Monica Holtz[1]

1. Introduction

This paper reports on a corpus-based analysis of verb/noun collocations in interdisciplinary registers. The research presented here is part of the project 'Linguistic Profiles of Interdisciplinary Registers'[2] which is being carried out at the Darmstadt University of Technology. The ultimate goal of this project is to linguistically analyse and profile emerging registers at the boundaries of computer science and some other discipline, such as computational linguistics, bioinformatics, and computational engineering, in order to investigate recent change in language as well as assess the influence of computer science on other disciplines.

This research is rooted in Systemic Functional Linguistics (SFL) (Halliday, 2004), and register linguistics (Halliday and Hasan, 1989; Biber, 1988, 1995; Biber et al., 1998; Biber et al., 1999). An integral part of register analysis is the linguistic characterisation of the field of discourse. The field of discourse reflects the domain-specificity of texts and is linguistically realised in terms of lexis, grammar, specialised terminology, and collocations. This work focuses on verbal complementation patterns, more specifically, verb/noun collocations.

The corpus under study in this paper consists of English texts from full research articles of the domains of computer science, computational linguistics, and linguistics and comprises over 6.28 million running words. The ultimate goal of this work is to identify similarities and differences between these three disciplines in terms of field of discourse. Hence, the aim of this work is the identification, classification, and analysis of collocations for verb/noun expressions of frequent words in these domains.

This paper is organised as follows. Section 2 presents a brief survey of the theoretical underpinnings of this research, corpus-based register analysis, and Systemic Functional Linguistics. Section 3 focuses on the corpus under study describing its compilation, preprocessing, annotation, and the tools used. The results of this work are discussed in Section 4 followed by conclusions in Section 5.

2. Corpus-Based Register Profiling

The theoretical and methodological underpinnings of this research are SFL, register linguistics and corpus linguistics (CL) (McEnery and Wilson, 2001).

SFL considers the functional variation of language and the context of situation in which this variation takes place, thereby providing an analytical framework for lexical and grammatical qualitative and quantitative analysis of linguistic features of this

[1] Department of Linguistic and Literary Studies, English Linguistics, Darmstadt University of Technology
e-mail: holtz@linglit.tu-darmstadt.de

[2] URL: http://www.linglit.tu-darmstadt.de/index.php?id=dfg_projekt (accessed: 11 May 2007).

variation. The texts under study are instances of different domain-specific registers and thus they require a linguistic characterisation in terms of register specific features. These features are described according to the principles of SFL including the parameters of field, tenor, and mode of discourse (Quirk et al., 1985). The parameter of field characterises texts in terms of their domain-specificity, being described in terms of lexis, specialised terminology, collocations, etc. The parameter of tenor characterises texts in terms of the interaction between the participants involved in the interaction, e.g., expert-to-expert for research articles. The parameter of mode refers to the realisation of the communication process in terms of channel and medium. For the texts under study, the channel is indirect, i.e., non-face-to-face communication, and the medium used in the communication is written-to-be-read. Hence, all texts are uniform in mode and tenor of discourse. Register variation is therefore mostly to be expected in terms of field of discourse, reflecting linguistic variation in terms of domain-specific terminology, e.g., nouns and collocations.

The parameter of field reflects as well in the transitivity of linguistic structures, which is one aspect of the ideational metafunction (Halliday, 2004). Transitivity describes the fact that experience is construed as a set of different process types, linguistically realised through verbs or verb groups, with different participants involved, mostly realised as nouns or noun groups, and associated with different types of circumstances. There are principally six process types in transitivity: material processes describing actual physical actions; mental processes describing the inner, mental experience; relational processes which are processes of identification and classification; behavioural processes representing “outer manifestations of inner workings, the acting out of processes of consciousness” (Halliday, 2004: 171); verbal processes of saying; and existential processes, which are processes concerned with existence in which phenomena are recognised ‘to be’. The co-occurrence of nouns and verbs is expected to vary among different domains, being evidence for register variation, as reflex of the situational parameter of field. This work focuses thus on the identification, classification and analysis of collocations for verb/noun expressions in these domains.

Corpus based linguistic analysis of language is inherent to SFL, as for SFL real texts are ‘fundamental to the enterprise of theorising language’ (Halliday, 2004: 34). However, SFL and CL attempt to describe language very differently. While SFL is a very complex theory for describing language, CL, in contrast, is a methodology that can be applied in almost any theoretical framework (Thompson and Hunston, 2006: 2). Nonetheless, they have some aspects in common. They both are concerned with naturally occurring language, with language as text and with the contexts in which language is used. For these reasons, CL was chosen as methodological background for this research.

3. Corpus Under Study

The corpus under study is a closed corpus compiled from 496 full scientific papers in English of the disciplines computer science, computational linguistics, and linguistics, available on-line, and comprises over 6.28 million running words. This corpus was compiled from a larger corpus of scientific papers used in the project ‘Linguistic Profiles of Interdisciplinary Registers’. Table 1 illustrates the text sources used in the compilation of the corpus under study.

The texts of the corpus are originally in PDF[3]-format. This format however does not allow any further annotation and querying of linguistic information of texts. Thus, all texts of the corpus were primarily converted to plain text format using the AnnoLab suite (Eckart, 2006; Eckart and Teich, 2007). UTF-8[4] encoding was used to assure that as many as possible of the original characters remain intact.

Sub-corpus	Source	Year	Tokens
Computer science	J. of Algorithms J. of Computer and System Science	2004–2006 2005–2007	2,765,412
Computational linguistics	J. of Computational Linguistics Machine Translation J. of Natural Language Engineering	2003–2006 1998–2004 2006 (12:1)	1,550,769
Linguistics	Language J. of Linguistics Functions of Language Linguistic Inquiry	2003–2006 2006 (42:1) 2005–2006 2005–2006	1,964,583

Table 1: Text sources of the corpus.

Metadata information for bibliography and for the situational parameters of field, tenor, and mode of discourse is provided for all texts. It is managed via JabRef[5], an open source bibliography reference manager. The native file format used by JabRef is BibTeX, the standard LaTeX bibliography format. Figure 1 shows the typical metadata information for a text in the corpus with field, tenor, and mode information.

```
@ARTICLE{McShane2005,
  author = {McShane, Marjorie and Nirenburg, Sergei and Beale, Stephen},
  title = {An NLP Lexicon as a Largely Language-Independent Resource},
  journal = {Machine Translation},
  year = {2005},
  volume = {V19},
  pages = {139–173},
  number = {2},
  month = {June},
  abstract = {This paper describes salient aspects of the OntoSem  lexicon of English, a lexicon whose semantic descriptions can either be grounded in a language-independent ontology, rely on extra-ontological […].},
  field.experientialdomain = {see keywords},
  field.goalorientation = {exposition},
  field.socialactivity = {hierarchic},
  keywords = {Lexicon – Ontological Semantics – Semantics – Language-independent resources – NLP},
  mode.channel = {graphic},
  mode.medium = {written},
  pdf = {archive\B1\McShane2005.pdf},
  tenor.agentiveroles = {expert-to-expert},
  timestamp = {2006.11.24},
  url = {http://dx.doi.org/10.1007/s10590–006–9001–y}
}
```

Figure 1: Example of meta-annotation.

[3] Adobe Portable Document Format; URL: http://www.adobe.com/de/products/acrobat/adobepdf.html (accessed: 11 May 2007).

[4] UTF-8 is defined by the Unicode Standard [UNICODE]. Descriptions and formulae can also be found in Annex D of ISO/IEC 10646-1 [ISO.10646]; URL: http://www.iso.ch/ (accessed: 11 May 2007).

[5] URL: http://jabref.sourceforge.net/ (accessed: 17 May 2007).

All metadata information and linguistic annotations of the corpus are stored separately in different layers. They are represented in a stand-off format. These annotations will be linguistically queried over strings and multiple annotation layers, which will allow various types of linguistic analysis of the corpus.

All texts in the corpus were automatically annotated for part-of-speech and lemmatised through AnnoLab. AnnoLab incorporates TreeTagger[6] (Schmid, 1994a, 1994b), a language independent part-of-speech tagger. TreeTagger's English parameter file was trained on the PENN[7] Treebank (Markus et al., 1993).

4. Results and Discussion

In order to decide on which verb/noun collocations to study, firstly a frequency list of nouns was generated for the whole corpus under study, comprising all three sub-corpora from computer science, computational linguistics, and linguistics, using Oxford WordSmith Tools 4.0[8] (Scott, 2004).

The obtained frequencies of nouns were divided into the following frequency bands: < 300, 300–600, 600–900, and > 900. Nouns occurring less then 300 times in the whole corpus were not considered relevant for this study. From each of the other frequency bands, three nouns were chosen for the verb/noun (i.e., verb-noun and noun-verb) collocation analysis. These nouns are: 'algorithm', 'system', and 'model' from the frequency band higher than 900; 'structure', 'process', and 'analysis' from the frequency band between 600 and 900; and finally 'information', 'meaning', and 'parameter' from the frequency band between 300 and 900. These nouns were chosen according to their frequency and their terminology potential, i.e., the representativeness or typicality of a word for a certain discipline (e.g., the word 'algorithm' can be assumed representative or typical for the domain of computer science). This representativeness can be supported by the observed high occurrence frequencies of nouns supposed to be typical for certain domains or disciplines. Table 2 displays the nouns chosen for the verb/noun collocation analysis and their frequency of occurrence in the whole corpus and in each of the sub-corpora. All frequencies of occurrence are normalised as per million words.

Nouns	**Frequency (per million words)**			
	whole corpus	**computer science**	**computational linguistics**	**linguistics**
algorithm	2,255	4,620	830	51
system	1,525	1,753	2,298	594
model	1,103	1,046	1,833	607
structure	898	468	974	1,444
process	787	671	1,096	705
analysis	677	366	643	1,142
information	509	298	974	438
meaning	417	27	628	800
parameter	324	389	429	151

Table 2: Nouns chosen for verb/noun collocation analysis and their occurrence frequency (per million words).

[6] URL: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html (accessed: 17 May 2007).

[7] URL: http://www.cis.upenn.edu/ treebank/ (accessed: 17 May 2005).

[8] URL: http://www.lexically.net/wordsmith/index.html (accessed: 17 May 2005).

These results indicate that the noun ‘algorithm’ is more frequent and hence typical, in the domain of computer science, while the nouns ‘system’, ‘model’, and ‘process’ are more frequent in the domain of computational linguistics, and finally that the nouns ‘structure’, ‘analysis’ and ‘meaning’ occur more frequently in the domain of linguistics.

Computational linguistics can be seen as a mixed discipline at the border of the plain disciplines computer science and linguistics having therefore a predisposition to assimilate and use terminology from both these domains (e.g., ‘algorithm’ is a term originally from the source domain of computer science, and ‘meaning’ is formerly a term from the source domain of linguistics). Although ‘algorithm’ does occur in the domain of linguistics and ‘meaning’ does occur in the domain of computer science, their frequencies are much lower than in the source domains or in the mixed domain, computational linguistics. On the other hand, other nouns, e.g., ‘parameter’ and ‘process’, show similar distribution in all three domains, probably because these words do not encode strong domain-specific meanings.

Concordances of the chosen nouns, both in singular and in plural form, and on the basis of the part-of-speech tagged sub-corpora, were generated using the query `{chosen noun} <w V*>*` (e.g., `algorithm/algorithms <w V*>*`) by Oxford WordSmith Tools 4.0. The resulting frequency of collocates, i.e., any verbs (node) collocating with the chosen nouns to either three positions left or right, are displayed in Table 3 (computer science), Table 4 (computational linguistics), and Table 5 (linguistics).

Sub-corpus of computer science										
Noun	**Total Left**	**Total Right**	**Left/Right**	**L3**	**L2**	**L1**	**Node**	**R1**	**R2**	**R3**
algorithm	1,826	938	1.95	441	733	653	<w V*>*	243	324	371
system	220	109	2.02	60	81	78	<w V*>*	29	34	46
model	469	224	2.09	140	159	170	<w V*>*	44	81	99
structure	183	115	1.59	51	65	67	<w V*>*	9	27	79
process	197	112	1.76	63	79	56	<w V*>*	14	43	54
analysis	157	104	1.52	50	60	47	<w V*>*	9	44	51
information	105	139	0.75	34	40	31	<w V*>*	27	60	52
meaning	8	11	0.71	5	1	2	<w V*>*	3	3	5
parameter	100	98	1.03	47	36	17	<w V*>*	30	38	29

Table 3: Sub-corpus of computer science: frequency (per million words) and position of collocates for the query {chosen noun} <w V*>*.

In the sub-corpus of computer science, all chosen nouns collocate with verbs both to their left and right positions (see Table 3). All nouns, but ‘information’ and ‘meaning’, collocate more frequently with verbs to the left rather than to the right. And the nouns ‘algorithm’, ‘system’, and ‘model’ even collocate roughly twice more often to the left than to the right of a verb.

Sub-corpus of computational linguistics										
Noun	**Total Left**	**Total Right**	**Left/Right**	**L3**	**L2**	**L1**	**Node**	**R1**	**R2**	**R3**
algorithm	393	165	2.38	86	141	166	<w V*>*	46	50	68
system	871	358	2.43	235	325	310	<w V*>*	48	130	180
model	694	288	2.41	172	257	265	<w V*>*	51	92	146
structure	211	130	1.63	56	74	81	<w V*>*	12	46	71
process	264	132	2.01	77	103	84	<w V*>*	23	40	68
analysis	232	149	1.56	76	86	70	<w V*>*	19	48	81
information	383	341	1.12	112	128	143	<w V*>*	82	131	128
meaning	148	156	0.95	53	54	41	<w V*>*	29	68	59
parameter	90	65	1.39	41	32	17	<w V*>*	16	23	26

Table 4: Sub-corpus of computational linguistics: frequency (per million words) and position of collocates for the query {chosen noun} <w V*>*.

In English, the left position to a verb is most probably occupied by the subject, whereas the right position to a verb may be taken by an object, a complement, or an adjunct in a sentence. A ratio left/right higher than one indicates therefore a preference of nouns for occupying the subject position in sentences. Thus, ‘algorithm’, ‘system’, and ‘model’ tend strongly to assume subject positions, while ‘information’ and ‘meaning’ occur most likely as objects, complements or adjuncts in sentences of the domain of computer science. This left/right ratio can also be seen as indication for the typicality of a word in a domain, since the subject position is a very prestigious one in English sentences. Hence, ‘algorithm’, ‘system’, and ‘model’ are more typical than ‘meaning’ and ‘information’ in the domain of computer science.

All chosen nouns collocate with verbs to both left and right positions in the sub-corpus of the domain of computational linguistics as well (see Table 4). The left/right ratio for ‘system’, ‘model’, and ‘process’ is higher than two, which corroborates the initial assumption, that these are typical nouns in the domain of computational linguistics. Although ‘algorithm’ has only the sixth higher occurrence frequency in this domain (see Table 2) its left/right ratio is 2.38 (see Table 4). This indicates that ‘algorithm’, while being formerly not from the domain of computational linguistics, has been strongly assimilated in this domain, thereby occupying preferably subject positions. In contrast, nouns not initially assumed for being typical for this domain, e.g., ‘meaning’, collocate preferably with verbs to non-subject positions.

Sub-corpus of linguistics										
Noun	**Total Left**	**Total Right**	**Left/Right**	**L3**	**L2**	**L1**	**Node**	**R1**	**R2**	**R3**
algorithm	25	10	2.50	6	10	9	<w V*>*	4	3	3
system	189	63	3.02	50	74	65	<w V*>*	7	19	36
model	204	121	1.68	61	61	81	<w V*>*	23	40	58
structure	357	212	1.69	120	118	120	<w V*>*	24	83	105
process	135	71	1.91	36	57	43	<w V*>*	9	29	34
analysis	532	268	1.98	150	167	215	<w V*>*	34	88	147
information	214	165	1.30	59	72	84	<w V*>*	39	63	64
meaning	233	210	1.11	94	75	64	<w V*>*	39	78	93
parameter	48	18	2.61	13	19	16	<w V*>*	3	6	9

Table 5: Sub-corpus of linguistics: frequency (per million words) and position of collocates for the query {chosen noun} <w V*>*.

The query results for the last discipline, linguistics, also show a tendency for preferred occurrence of the chosen nouns at subject positions, since all left/right ratio are higher than one (see Table 5). Based on the frequency of occurrence, ‘structure’, ‘analysis’, and ‘meaning’ were assumed to be typical for the domain of linguistics (see Table 2). The nouns ‘structure’ and ‘analysis’ comply with the profile shown for typical nouns in the two other domains with clear preference for subject positions. In contrast, ‘meaning’ is found almost equally at both positions. Even though ‘meaning’ is a typical noun in the domain of linguistics, it shows a versatile profile of collocation with verbs, assuming almost equally subject and non-subject positions. On the other hand, the occurrence frequency of ‘system’ in the domain of linguistics is just 594 per million words (see Table 2), while its left/right ratio is higher than three. This indicates that ‘system’ has been strongly assimilated into this domain assuming thereby mainly subject positions.

The next step in this study is to investigate in more detail the actual variety of lexical verbs occurring in such verb/noun collocations. For this purpose, the nouns ‘algorithm’, ‘system’, and ‘meaning’ are considered. The first two nouns are from the highest frequency band (> 900) and the last noun is from the lowest frequency band (600–300), so that eventual differences in collocation patterns are expected to be predominant. These nouns collocate predominantly with the verbs ‘be’ and ‘have’ in all three sub-corpora, following the collocation pattern: ‘{chosen noun} is / has x’ or ‘x is / has {chosen noun}’. Figure 2 shows some examples of this collocation pattern for ‘algorithm’.

```
   algorithm , respectively . Our algorithm is   then analyzed in Section 4
       location algorithm Theorem 3 . There is   a nice algorithm for the
algorithm , since the accelerated algorithm is   very efficient in the
   later iterations ) ; thus this algorithm is   a 2 – approximation
         s algorithm . The latter algorithm has  an approximation factor 3k
       satisfy as desired . Suppose that we have an algorithm which is able
         . Proof . Assume that an adversary has  an algorithm A that takes N
```

Figure 2: Examples of collocations ‘‘algorithm’ is / has x’ or ‘x is / has ‘algorithm’’.

However, the repertoire of other lexical verbs collocating with the chosen nouns varies considerably both qualitatively and quantitatively within the three sub-corpora. The first noun, ‘algorithm’, collocates with a larger variety of lexical verbs in the domain of computer science compared to the other two domains. Moreover, the frequency of collocates with other verbs than ‘be’ and ‘have’ is notably higher in computer science than in computational linguistics and linguistics domains. Table 6 illustrates some of these verb-collocates with ‘algorithm’ both at the left and right position to a lexical verb for the three domains under study, and Figure 3 shows some examples of these collocations.

```
more general algorithm . Our algorithm works    for the more general
moderate . Extensions to our algorithm allow    one to predict the
        Also , the algorithm of [ 12 ] assume   that the processors have
this problem , since the algorithm can decide   to go up or down along th
           The first four intervals We design   an algorithm Slow – LPT
         Z ( ai ) . The algorithm will estimate Z ( am ) by estimating
 / Kloks algorithm would substantially improve  the performance
  membership query algorithm which can learn    any polynomial size
bounds the ability of any algorithm to predict  elements of a sequence of
      Thus , the above algorithm might require  exponentially many steps
             . 1 , we will show how to run      the algorithm without the
problem . However , the algorithm will use      certain structural
           e target concept, and there exists   a learning algorithm that
```

Figure 3: Examples of verb/noun collocations for ‘algorithm’.

Algorithm								
Computer science			**Computational linguistics**			**Linguistics**		
Verb	**Left**	**Right**	**Verb**	**Left**	**Right**	**Verb**	**Left**	**Right**
accept	13.0	6.9	describe	9.7	4.5	adopt	0.0	0.5
achieve	20.3	2.5	learn	1.3	0.0	assume	0.5	0.0
allow	6.9	1.8	operate	2.6	0.0	begin	0.5	0.0
apply	4.0	11.2	outperform	1.9	0.0	deduce	0.5	0.0
approximate	10.8	4.7	produce	2.6	3.2	extract	0.5	0.0
assume	5.8	1.1	rely	2.6	0.0	generate	2.5	0.0
Call	2.2	8.7	require	3.2	0.0	identify	1.0	0.0
compare	5.1	1.1	resolve	3.9	0.0	predict	0.5	0.0
compute	34.7	4.0	run	0.0	2.6	runs	0.5	0.0
consider	8.3	10.1	suggest	0.6	0.0	takes	0.5	0.0
construct	4.7	5.1	take	1.3	0.0	use	1.0	1.0
design	3.6	3.6	use	23.2	6.4			
exist	5.4	8.7	works	3.2	0.0			
Find	37.6	5.1						
generate	7.6	0.0						
make	13.4	6.5						
need	15.5	3.3						
perform	14.1	2.5						
produce	11.9	4.7						
provide	5.8	4.7						
require	13.4	1.4						
return	19.9	4.0						
Run	65.1	25.7						
solve	22.4	1.4						
take	17.4	0.0						
work	25.3	2.9						

Table 6: Frequency of some verbs collocating with the noun ‘algorithm’ (per million words).

In all three domains ‘algorithm’ occupies predominantly subject positions. Additionally, the verb-collocates for ‘algorithm’ in the domain of computer science realise a broad variety of process types, i.e., material (e.g., ‘run’, ‘achieve’, ‘generate’), relational (‘allow’, ‘need’), mental (‘assume’, ‘accept’), verbal (‘call’, ‘require’), and existential (‘exist’). In the domains of computational linguistics and linguistics the amount of different verb-collocates decreases considerably and also the range of different process types is narrowed, i.e., no quantitatively relevant existential processes in the computational linguistics and linguistics corpora, and no quantitatively relevant verbal processes in the linguistics corpus. These observations reinforce the initial supposition that ‘algorithm’ is originally a term from the domain of computer science migrating into the domain of computational linguistics and eventually reaching the domain of linguistics.

The second chosen noun, ‘system’, shows a similar profile for verb/noun collocations. However, reflecting the fact that ‘system’ occurs more often in the domain of computational linguistics (see Table 2), it also collocates with a larger variety of different lexical verbs either to the left or right position in the domain of computational linguistics compared to the other two domains. Table 7 shows some examples of lexical verbs collocating with ‘system’ and their frequency of occurrence for the three sub-corpora under study.

In the domains of computational linguistics and computer science, ‘system’ occupies mostly subject positions, while in the domain of linguistics it takes very often non-subject positions. The diversity of process types linguistically realised with

System								
Computer science			**Computational linguistics**			**Linguistics**		
Verb	**Left**	**Right**	**Verb**	**Left**	**Right**	**Verb**	**Left**	**Right**
allow	1.4	0.0	achieve	5.2	0.0	acquire	1.5	0.0
apply	0.0	0.7	adapt	0.0	2.6	allow	2.0	0.0
assume	0.7	0.0	allow	7.1	3.2	become	2.0	0.0
check	0.7	0.7	ask	1.9	0.0	describe	1.0	1.5
consider	0.4	5.1	become	2.6	0.6	develop	1.0	1.5
consist	2.5	0.0	build	2.6	3.9	find	1.5	0.5
converge	1.1	0.0	choose	2.6	0.0	follow	2.0	1.0
establish	1.1	0.4	consist	3.2	0.0	implement	0.0	1.5
find	1.1	0.0	create	0.6	1.9	integrate	0.5	0.0
generate	0.7	0.0	determine	0.6	0.6	propose	1.0	0.0
minimise	1.8	0.0	develop	12.3	4.5	seem	3.1	0.0
model	5.4	5.8	evaluate	2.6	7.1	underline	1.5	0.0
need	3.3	0.4	extract	2.6	0.0	use	2.0	1.0
obtain	1.8	1.1	generate	5.2	1.3	work	0.5	0.0
operate	0.0	1.4	handle	3.9	0.0			
reach	1.8	0.0	help	1.3	3.9			
represent	0.4	0.0	identify	3.9	0.0			
require	1.8	0.0	implement	5.2	1.9			
satisfy	5.8	1.1	improve	3.2	3.9			
solve	2.5	0.7	include	3.2	0.0			
use	5.8	1.4	learn	11.0	0.0			
			make	5.2	6.4			
			obtain	1.3	3.2			
			perform	4.5	0.0			
			produce	10.3	9.7			
			propose	5.2	7.1			
			recognise	4.5	0.0			
			rely	3.2	0.0			
			train	12.3	4.5			
			use	29.0	14.8			
			work	8.4	2.6			

Table 7: Frequency of some verbs collocating with the noun ‘system’ (per million words).

‘system’ varies also domain-specifically. For instance, mental processes (e.g., ‘learn’, ‘evaluate’, ‘recognise’ ‘acquire’, ‘consider’) occur in all three domains; however their frequency of occurrence is higher and the variety of mental verbs is bigger in the domain of computational linguistics. These observations corroborate the initial assumption that ‘system’ is a typical noun in the domain of computational linguistics.

Finally, ‘meaning’ shows the most differentiated profile of verb/noun collocations within the three domains under study compared to the former two nouns. Table 8 shows some examples of verbs collocating with ‘meaning’ and their frequency of occurrence for the three sub-corpora under study.

While ‘meaning’ does not collocate with many different verbs in the computer science domain assuming thereby practically only non-subject positions, it exhibits a higher freedom degree both in the variety of verbs and in the choice for syntactic function in the sentences in the domains of linguistics and computational linguistics. Such observations comply with the initial assumption that ‘meaning’ is a very typical term in the domain of linguistics, being also common in the domain of computational linguistics and not very likely in the computer science domain.

Meaning								
Computer science			**Computational linguistics**			**Linguistics**		
Verb	**Left**	**Right**	**Verb**	**Left**	**Right**	**Verb**	**Left**	**Right**
Call	0.0	0.4	agree	0.6	0.0	become	2.0	0.0
describe	0.0	0.4	capture	1.3	6.4	build	0.5	2.6
Give	0.0	0.4	carry	0.0	3.2	carry	0.0	4.5
understand	0.0	0.4	change	3.2	0.6	change	0.5	5.8
			clarify	0.0	0.6	characterise	0.0	3.2
			comprise	1.3	0.0	combine	0.0	1.9
			compute	0.6	0.6	complement	1.5	1.3
			convey	1.3	6.4	consider	1.5	1.3
			declare	0.0	0.6	constitute	0.0	1.3
			encode	0.0	1.3	convey	0.5	1.3
			express	2.6	2.6	describe	0.5	2.6
			express	0.0	3.9	distinguish	0.0	1.9
			model	0.0	0.6	encode	1.0	5.8
			represent	1.3	7.1	express	0.0	3.2
			require	0.6	0.6	identify	0.0	1.9
			share	0.0	1.9	involve	2.0	1.3
			specify	0.0	1.3	occur	1.5	0.0
						provide	1.0	0.6
						require	2.5	0.0
						share	0.5	6.4
						show	0.0	1.3
						undergo	0.5	0.0
						understand	0.0	1.9
						use	7.1	1.9

Table 8: Frequency of some verbs collocating with the noun ‘meaning’ (per million words).

5. Conclusion

This paper explored verb/noun collocations in the registers of computer science, computational linguistics, and linguistics on the basis of corpus data, and is theoretically rooted in SFL. For this purpose, a corpus of research articles in English from these three registers was compiled. Based on a selection of typical nouns from each domain, verb/noun collocations were identified and analysed.

It was found that the chosen nouns occur more often in their original source domain than in the other domains. Moreover, the more far away a domain is from the source domain of these nouns, the less they occur in this domain. It was also observed that in all three registers the selected nouns collocate primarily with the verbs ‘be’ and ‘have’ both at left, i.e. subject, and right, i.e. non-subject, positions. Additionally, these nouns collocate with a greater variety of lexical verbs in their source domain than in the other domains. Again, the more distant a domain is from the source domain of a certain noun, the less it collocates with different verbs. In addition, the more a noun is assimilated in a new domain, the more versatile it becomes with regard to syntactic function. Finally, the more a noun is incorporated in a register other than its original one, the greater the variety of processes.

Future work will include similar and further lexical (e.g., adjective/noun collocation) and grammatical analysis (e.g., voice, agency, theme) on large corpora of other scientific registers.

Acknowledgments

I would like to thank the *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) for supporting the project ‘Linguistic Profiles of Interdisciplinary Registers’. Thanks to Elke Teich, Richard Eckart, and Sabine Bartsch for their valuable comments on earlier versions of this paper.

References

Biber, D. (1988) Variation across speech and writing. Cambridge: Cambridge University Press.

Biber, D. (1995) Dimensions of register variation: A cross-linguistic comparison. Cambridge: Cambridge University Press.

Biber, D., S. Conrad, and R. Reppen (1998) Corpus linguistics: Investigating language structure and use. Cambridge: Cambridge University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finnegan (1999) Longman Grammar of Spoken and Written English. Essex: Pearson Education.

Eckart, R. (2006) A framework for storing. managing and querying multi-layer annotated corpora. Diploma thesis. Darmstadt: Darmstadt University of Technology.

Eckart, R. and E. Teich (2007) An XML-based data model for flexible representation and query of linguistically interpreted corpora. in Rehm. G., A. Witt, and L. Lemnitzer eds. *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen. Data Structures for Linguistic Resources and Applications*. Proceedings of the Biennial GLDV Conference 2007. Tübingen: Gunter Narr.

Halliday, M.A.K. (2004) An Introduction to Functional Grammar. 3rd edition. revised by C.M.I.M. Matthiessen. London: Edward Arnold.

Halliday, M.A.K. and R. Hasan (1989) Language, context and text: aspects of language in a social-semiotic perspective. Oxford: Oxford University Press.

Marcus, Mitchell P., B. Santorini, and M. A. Marcinkiewicz (1993) Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

McEnery, T. and A. Wilson (2001) Corpus Linguistics. Edinburgh: Edinburgh University Press.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik (1985) A Comprehensive Grammar of the English Language. London: Longman.

Schmid, H. (1994a) Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing.

Schmid, H. (1994b) Part-of-Speech Tagging with Neural Networks. Proceedings of the 15th International Conference on Computational Linguistics (COLING–94).

Scott, M. (2004) WordSmith Tools version 4. Oxford: Oxford University Press.

Thompson. G. and S. Hunston eds (2006) System and Corpus: exploring connections. London: Equinox.