From: *Proceedings of the International Conference "Corpus Linguistics–2017"*, Saint Petersburg: Saint Petersburg State University.

A. Stefanowitsch

A LOT OF DATA: TEXTUALLY DISTINCTIVE COLLEXEMES IN A CORPUS OF SCIENTIFIC ENGLISHES

Abstract. Associations between words and grammatical patterns have been studied under various labels Such studies have consistently shown that grammatical structures are typically associated with an above-chance frequency with sets of lexical items that are often functionally or semantically motivated. The stability of such associations across text types is less clear: since vocabulary differs quite strongly depending on text type, the same would be expected of lexicon-grammar associations. In this paper, I show that such variation exists and can be used to investigate domain-specific functions of grammatical patterns as well as the functional relationship between text types.
Keywords. Collocational frameworks, collostructional analysis, text types, Scientific English, quantitative corpus linguistics.

1. Introduction

In this paper, I combine the logic of keyword analysis, a method for uncovering associations between words and text types, and collostructional analysis — specifically, distinctive collexeme analysis, a collocational method for investigating associations between words and alternating grammatical constructions. I will apply this combination to a well-studied collocational framework, [*a*(*n*) N *of*] in a corpus of Scientific English and a general corpus of (American) English, in order to determine the extent and quality of variation in lexicon-grammar associations across text types.

2. Descriptive and methodological background

By *text type* I mean here varieties defined externally by situation and topic area — roughly, what is referred to in applied contexts as «language for specific purposes», such as Business English, Academic English etc. Such text types have been investigated through keyword analysis (cf. Scott 1997 and the work building on it). They have specific vocabulary associated with them, which is unsurprising in the case of content words. However, function words also show such associations, pointing to grammatical differences between text types. That such grammatical differences exist is, of course, also known, it has been demonstrated impressively, for example, in the research tradition started in Biber (1985), where bundles of lexico-grammatical features are used to identify and categorize text types.

Distinctive collexeme analysis is one of a family of collocational methods that focus on statistical associations between words and grammatical structures (collocational frameworks, grammar patterns, constructional

idioms, constructions etc., cf. Stefanowitsch and Gries 2009 for an overview). Specifically, distinctive collexeme analysis compares association of lexical items to a functionally equivalent slot of two related constructions (for example, verbs in the ditransitive and the prepositional dative). Words that are statistically significantly associated with one of the constructions are referred to as distinctive collexemes of that construction.

In this paper, I combine this procedure with the idea of keyword analysis such that I compare the associations of lexical items to a slot in a single given construction in two (or more) text types. Specifically, I investigate the nouns associated with the collocational framework [*a*(*n*) N *of*] in Scientific English as compared to general usage; since there is growing evidence that Scientific English itself is not a monolithic text type (e.g. Biber and Gray 2016), I also investigate the nouns associated with this framework in different scientific domains. Words that are associated with a construction in one text type as opposed to another are referred to as textually distinctive collexemes of that construction in that text type.

The patterns [*a* N *of*] and [*an* N *of*], treated here as a single pattern, are two examples of sequences of two function words interrupted by variable slot for a content word called *collocational frameworks* by Renouf and Sinclair (1991). The content-words collocates (or, in terms of collostructional analysis, collexemes) of these frameworks typically come from a small number of semantic fields, suggesting that many such frameworks are (parts of) functionally motivated linguistic units. Specifically, Renouf and Sinclair find that the words in the pattern(s) [*a*(*n*) N *of*] tend to be measurements or partitives (although other possessive relations are also found) — one of the central functions of the framework seems to be quantification.

The corpora used in this paper are COCA, a 400-million-word corpus of spoken and written general American English, and FUSE-F, a 100-million+ word corpus of open access scientific research papers under development at the Freie Universität Berlin.

3. Case studies

3.1. [a(n) N of] in Scientific English

The collocational framework [*a* N *of*] (without the variant *an*) is one of three patterns investigated in Marco (2000) with respect to their occurrence in a small proprietary corpus of medical research papers. Using relative frequency as an association measure, he finds domain-specific associations that differ in their specifics from those in general usage, but that partially

conform to it in that they fall in the domain of measurements (*dose*, *group*, *measure*); in addition, he finds words that express quantifiable properties (*specificity*, *sensitivity*, *accuracy*).

Before zooming in on specific domains like medicine, I will attempt a broader and statistically more stringent replication of his study. Based on a comparison of the collocates in the framework [*a*(*n*) N *of*] in the FUSE-F and the COCA, I identified textually distinctive Scientific English and those more strongly associated with general usage. Table 1 shows the textually distinctive collexemes of the framework in Scientific English.

Table 1. **Textually distinctive collexemes of [*a*(*n*) N *of*] in Scientific vs. General English**

Collexeme	FUSE (O:E)	COCA (O:E)	Coll. Str.
function	(18008:5403)	(3062:15667)	33921.30
subset	(6063:1683)	(500:4880)	13330.10
number	(25815:15133)	(33202:43884)	9462.88
total	(10015:4219)	(6437:12233)	9168.40
consequence	(5383:1757)	(1468:5094)	8448.28
variety	(16314:8863)	(18251:25702)	7607.35
range	(7971:3342)	(5062:9691)	7357.52
set	(9533:4316)	(7298:12515)	7332.22
effect	(2527:709)	(236:2055)	5417.00
measure	(5075:2033)	(2855:5897)	5173.96
role	(2133:569)	(87:1651)	5131.59
reduction	(2428:712)	(349:2065)	4725.85
increase	(2963:983)	(870:2850)	4488.15
combination	(6518:3101)	(5575:8992)	4394.52
decrease	(1524:423)	(127:1228)	3332.13
marker	(1497:422)	(148:1223)	3171.24
overview	(2435:873)	(968:2530)	3145.88
model	(2992:1192)	(1655:3455)	3083.83
result	(9673:5867)	(13208:17014)	3034.56
inhibitor	(1173:305)	(17:885)	3027.24

Interestingly, the textually most distinctive collexemes of the framework are not overwhelmingly quantifying expressions. There are a few cases (*number*, *total,* and arguably *variety* and *range,* though these stress diversity rather than pure quantity); however, most collexemes are best characterized as relatively abstract possessive uses encoding causality (*function*, *consequence*, *effect*, *reduction*, *increase*, *decrease*, *result*) or categorization (*subset, set*). In addition, there are individual items relating to the scientific process in general (*measure*, *model, overview*) or specific scientific concepts (*marker, inhibitor*).

In contrast, as Table 2 shows, the textually distinctive collexemes of the pattern in general usage are mainly the kind of quantifying and/or partitive

Table 2. **Textually distinctive collexemes of [*a*(*n*) N *of*] in General vs. Scientific English**

Collexeme	COCA (O:E)	FUSE (O:E)	Coll. Str.
lot	(143005:107256)	(1237:36986)	78626.65
couple	(35816:27058)	(572:9331)	17163.17
kind	(16085:12592)	(849:4342)	5147.64
bit	(9466:7097)	(78:2447)	4935.59
bunch	(8179:6103)	(28:2104)	4563.59
sense	(15776:12657)	(1245:4365)	3861.93
piece	(11171:8693)	(520:2998)	3802.28
matter	(16621:13802)	(1940:4759)	2724.68
professor	(4111:3069)	(16:1058)	2273.66
friend	(3856:2878)	(15:993)	2132.47
sort	(6004:4707)	(326:1623)	1881.99
handful	(7726:6207)	(622:2141)	1852.13
way	(7100:5676)	(533:1957)	1799.63
man	(3093:2304)	(5:794)	1774.24
bottle	(3148:2366)	(34:816)	1583.73
cup	(3441:2607)	(65:899)	1571.09
glass	(3094:2342)	(56:808)	1425.40
part	(12987:11193)	(2066:3860)	1289.59
act	(3023:2307)	(79:795)	1272.20
pile	(2376:1782)	(20:614)	1232.30

expressions (*lot*, *couple*, *bit*, *bunch*, *piece*, *handful*, *bottle*, *cup, glass*, *part*, *pile*) that Renouf and Sinclair (1991) found; additionally, there are type expressions (*kind*, *sort*) and various possessive constructions from the social domain (*professor*, *friend*, *man*, *act*) — the latter being completely absent from the textually distinctive collexemes of Scientific English.

Thus, while the pattern is used for quantification in Scientific English, it is used in this way much less frequently than in general usage. This result, which may appear somewhat surprising at first glance, given that quantification plays a crucial role in scientific discourse, makes sense once we take into account the *kind* of quantification that the pattern is used for: it is used for relatively imprecise quantities like *lot*, *couple*, *bunch*, etc., which are unlikely to be used in reporting scientific results.

In sum, while the pattern serves the same range of functions both in Scientific English and in general usage, Scientific English places a greater emphasis on the relational exploits the pattern in different ways. One crucial difference to Marco's (2000) results is that the collocates identified are less domain-specific, but this is due to the fact that our corpus includes text from a broader range of disciplines, so that collexemes have a higher chance of becoming textually distinctive if they are used *across* these disciplines — they really are typical of Scientific English in general rather than any particular discipline-specific English.

3.2. [a(n) N of] across Scientific Englishes

Let us turn to a more direct (if still quantitatively more rigorous) replication of Marco's (2000) and similar studies, focusing on individual disciplines. The subcorpora for these disciplines were constructed by grouping the journals in the FUSE-F corpus into five broad categories — medicine, neurosciences, life sciences (biology and biochemistry), physical sciences (physics, chemistry, engineering) and psychology. Each subcorpus was individually compared against the COCA. Table 3 lists the top 5 textually distinctive collexemes of each discipline (this limit is due to length restrictions, see the section Data and Software below for a link to more extensive supplementary materials).

The direct comparison of individual discipline-specific Englishes with the general usage represented by COCA shows clear differences between these text types. In small part, this is due to domain-specific terminology becoming textually distinctive, as in the case of *inhibitor* for Medicine or *solution* (in the sense of «liquid mixture of a substance and a solvent») for the Physical Sciences. However, most of the textually most distinctive

Table 3. **Textually distinctive collexemes of [*a(n)* N *of*] in five Scientific Englishes vs. General English**

Collexeme	Sci. Engl. (O:E)	COCA (O:E)	Coll. Str.
Medicine			
subset	(1090:73)	(500:1517)	4811.33
variety	(3567:996)	(18210:20781)	4421.44
number	(4785:1736)	(33161:36210)	4057.96
consequence	(908:109)	(1464:2264)	2594.35
inhibitor	(422:20)	(17:419)	2464.21
Neurosciences			
function	(9008:1454)	(3001:10556)	25741.66
subset	(2703:388)	(500:2815)	8809.51
set	(4263:1397)	(7281:10147)	4738.66
consequence	(1999:419)	(1464:3044)	4120.90
total	(3599:1214)	(6434:8819)	3805.53
Life Sciences			
subset	(1534:160)	(500:1874)	5638.74
total	(3325:767)	(6434:8992)	5522.84
function	(2188:408)	(3001:4781)	4593.05
number	(7256:3176)	(33161:37241)	4491.47
consequence	(1577:239)	(1464:2802)	4071.37
Physical Sciences			
function	(720:23)	(3001:3698)	3765.48
factor	(100:6)	(814:908)	395.90
solution	(55:1)	(170:224)	310.86
decrease	(46:1)	(127:172)	268.68
increase	(76:6)	(869:939)	254.61
Psychology			
function	(5479:532)	(3001:7948)	20066.17
effect	(1073:82)	(236:1227)	4751.77
measure	(1641:282)	(2849:4208)	3588.19
set	(2263:599)	(7281:8945)	3060.81
total	(1935:525)	(6434:7844)	2527.53

collexemes are still from the semantic field Science in general, the disciplines differing in the importance that these collexemes play (for example, *subset* plays a very important role in Medicine, the Life Sciences and Neuroscience, but not the Physical Sciences or Psychology, and *decrease/increase* play a very important role in the Physical Sciences but not the other disciplines). When more than just the top five collexemes are included, the overlap of course becomes greater, but the differences in importance remain and could provide interesting insights into the relative role of particular scientific concepts in different disciplines.

To get at the domain-specific vocabulary, a more direct comparison of the texts from the different scientific disciplines *amongst each other* rather than to general usage is necessary. There are different ways in which such multiple comparisons can be achieved, in the collostructional literature, no single method has become the standard. Here, I use a method proposed by Oakes and Farrow (2007), who simply create a large two-dimensional contingency table of all lexical items and their frequencies in all corpora involved and calculate the contribution of each cell to the overall chi-square value. These chi-square components are then used as association measures. Table 4 lists for each variety the five *attracted* collexemes with the largest chi-square component (i. e. the ones significantly more frequent than expected) and the five *repelled* collexemes with the largest chi-square components (i. e. the ones significantly less frequent than expected). This tells us not only what vocabulary is preferred in each discipline as opposed to the others, but also what vocabulary is avoided.

Using this method yields an abundance of domain-specific terminology, such as *panel, dose* and *GOR* (*grade of recommendation*) for Medicine, *network* and *threshold* for Neuroscience, *homolog* and *MOI (multiplicity of infection*) for the Life Sciences, *LOD (limit of detection*), *MAAT (mean annual air temperature*) and *solution* for the Physical Sciences. Interestingly, Psychology does not have such domain-specific vocabulary among the very strongest collexemes, suggesting that it favors a more broadly accessible style of scientific writing. Of course, if we include more data, there will be domain-specific vocabulary for all fields, e. g. *illusion* and *representation* for Psychology (ranked 18th and 19th). Even the direct comparison of different Scientific Englishes against each other, however, shows that general scientific vocabulary is associated with different disciplines to different degrees. For example, the word *function* plays a very important role in Neuroscience, Physical Sciences and Psychology, but not in the other two disciplines.

Table 4. **Textually distinctive collexemes of [*a(n)* N *of*] in five Scientific English as a subtypes of Scientific English**

Attracted	Coll. Str.	Repelled	Coll. Str.
Medicine			
variety	740.01	*function*	1481.65
panel	670.87	*set*	476.91
dose	464.48	*measure*	245.84
GOR	435.94	*sequence*	157.51
inhibitor	409.05	*pair*	136.35
Neuroscience			
function	458.38	*member*	225.70
history	174.99	*variety*	156.91
network	173.80	*panel*	125.66
train	156.39	*source*	107.85
threshold	115.77	*homolog*	98.14
Life Sciences			
member	494.07	*function*	1152.75
homolog	450.91	*sense*	261.65
MOI	313.76	*measure*	249.64
total	289.24	*effect*	249.43
suite	271.31	*sequence*	178.28
Physical Sciences			
LOD	581.81	*total*	83.36
function	470.26	*subset*	72.39
MAAT	317.35	*group*	50.37
factor	286.53	*history*	31.10
solution	271.12	*role*	29.51
Psychology			
sense	1222.95	*variety*	264.26
function	1126.17	*member*	228.51
effect	688.34	*number*	228.34
sample	453.34	*inhibitor*	216.26
measure	437.67	*concentration*	189.92

Among the repelled textual collexemes in the different disciplines, we find, unsurprisingly, domain-specific vocabulary from other disciplines, for example, *history* in the Physical Sciences and *inhibitor* in Psychology. Again, however, we also find general scientific vocabulary that is avoided in particular disciplines, such as *measure* in Medicine and Life Sciences and *total* in Physics.

Interestingly, the prominent function of quantification, which was already weakly represented in Scientific English as a whole (cf. Table 1 above), is almost completely absent from the domain-specific collocates in Table 4, the only exceptions being *variety* and *dose* in Medicine. The obvious and most likely explanation is that this function is evenly distributed across disciplines, but as a consequence, the domain-specific phraseological patterns of the framework [*a*(*n*) N *of*] in Scientific Englishes are radically different from general usage not just with respect to domain-specific vocabulary, but also with respect to the dominant meaning(s) of the pattern.

3.3. Collexemes of *[a(n) N of]* as indicators of text type

To get a more general idea as to how the function of the framework [*a*(*n*) *N of*] differs across general usage and various Scientific Englishes, we can cluster text types by the distribution of collexemes within this framework in the spirit of Biber's research mentioned above. Here, I selected 1000 collexemes on an *n*-th line basis from each text type represented in COCA and each discipline in FUSE-F that had at least 1000 occurrences. These were used as a basis for a distance matrix that was submitted to a hierarchical cluster analysis.

The results are surprisingly consistent: the first main difference is between spoken English and all written varieties, pointing to differences that are not unexpected but that have not, to my knowledge, been investigated. The next split is between the non-academic text types in COCA and all Scientific Englishes, including those represented in COCA as «academic». Among the Scientific Englishes, there are various well-motivated clusters of sub-disciplines from medicine, biology and chemistry: for example, pediatrics and public health cluster together, as do neurology and psychiatry, as do immunology, oncology, endocrinology and pharmacology, which are joined by chemistry one level up. The only unexpected cluster is the one containing Physics and Psychology, which may simply be due to the fact that these two disciplines are relatively distant from the others, which form a sort of continuum from chemistry over biology to neuroscience.

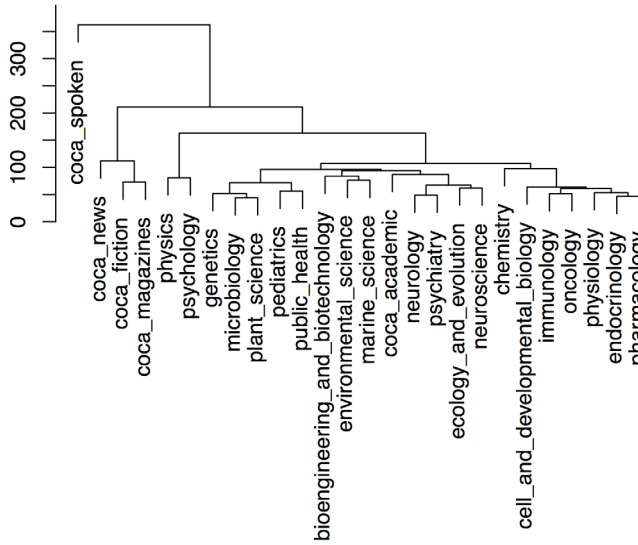

Fig. 1. **Text types in COCA and FUSE-F clustered by collexemes in the collocational framework [*a(n)* N *of*]**

4. Conclusion

The case studies in this paper have show that even highly entrenched collocational frameworks like [*a*(*n*) N *of*] may vary across text types in two ways. First, in their specific lexical associations, which differ due to domain-specific vocabulary and due to domain-specific preferences for general vocabulary. Specifically, [*a*(*n*) N *of*] is used for quantification in general usage but serves a wider range of functions in Scientific English. The studies also show that while there is good reason to assume a broad category of Scientific English that differs clearly from non-academic varieties, there are considerable differences between scientific disciplines, so that Scientific English is best thought of as a cluster of varieties that share a general scientific vocabulary but are differentiated by their specific terminology and that these differences interact with grammatical patterns systematically.

Supplementary materials

The data sets for the case studies reported here may be downloaded from www.stefanowitsch.de/data/2017alod.zip

Data and Software

1. *Davies M.* (2008-) The Corpus of Contemporary American English (COCA), 2016 ed. (commercial version). Provo (Utah), 2016.
2. *Flach S.* (2017), {collostructions}. An R implementation for the family of collostructional methods, v 0.0.10., www.bit.ly/sflach
3. *R Development Core Team* (2017), R: A language and environment for statistical computing, v. 3.3.3. www.R-project.org.
4. *Stefanowitsch A, Flach S.* (2017), The Frontiers Free University Scientific English corpus (FUSE-F), Beta. Berlin, 2016.

References

1. *Biber D.* (1985), Investigating macroscopic textual variation through multi-feature/multi-dimensional analyses. In: Linguistics 23, pp. 337–360.
2. *Biber D., Gray B.* (2016). *Grammatical complexity in Academic English: Linguistic change in writing.* Cambridge, 2016.
3. *Marco M. J. L.* (2000), Collocational frameworks in medical research papers: a genre-based study. In: English for Specifc Purposes 19 (1), pp. 63–86.
4. *Oakes M. P., Farrow M.* (2007), Use of the Chi-Squared test to examine vocabulary differences in English language corpora representing seven different countries. In: Literary and Linguistic Computing 22 (1), pp. 85–99.
5. *Renouf A., Sinclair J.* (1991), Collocational frameworks in English. In: Aijmer K., Altenberg B. (eds.), English corpus linguistics. London, 1991, pp. 128–143.
6. *Scott M.* (1997), PC analysis of key words — And key key words. In: System 25 (2), pp. 233–245.
7. *Stefanowitsch A., Gries St. Th.* (2009), Corpora and grammar. In: Lüdeling A., Kytö M. (eds.), Corpus linguistics: An international handbook, vol. 2. Berlin, New York, 2009, pp. 933–952.

Anatol Stefanowitsch
Freie Universität Berlin (Germany)
E-mail: anatol.stefanowitsch@fu-berlin.de