CREATION OF A CORPUS OF SPECIALISED LANGUAGE

Rosa Derosas

I have analysed two different corpus concerning biochemistry: one composed of some articles about Cannabis (corpus A) and one of some different abstract about Dna, Proteins, Vitamins(corpus B). All the articles and abstracts of both corpus are taken from the web.

The analysis of two different corpus about the same area has allowed me to detect better the most common features of bio-chemical language

Checklist

QUANTITATIVE ANALYSIS

1.What are the most frequent 100 words used in your corpus?

The most frequent 100 words are:

Corpus A

	Items	Frequency
[1]	The	444
[2]	Of	332
[3]	And	279
[4]	A	213
[5]	In	212
[6]	To	185
[7]	Thc	103
[8]	marijuana	101
[9]	that	75
[10]	As	68
[11]	Is	65
[12]	by	64
[13]	driving	63
[14]	On	63
[15]	for	62
[16]	with	60
[17]	from	57
[18]	this	54
[19]	were	54
[20]	was	50
[21]	have	45
[22]	are	44
[23]	Or	44
[24]	these	44
[25]	cannabis	42
[26]	Be	38
[27]	not	36
[28]	been	34

Corpus B

	Items	Frequency
[1]	The	649
[2]	Of	416
[3]	And	257
[4]	In	244
[5]	A	201
[6]	To	184
[7]	Is	152
[8]	Are	117
[9]	As	102
[10]	That	100
[11]	Dna	77
[12]	which	73
[13]	For	72
[14]	It	71
[15]	vitamin	68
[16]	By	67
[17]	Or	65
[18]	One	60
[19]	protein	60
[20]	Be	57
[21]	proteins	56
[22]	With	50
[23]	From	48
[24]	cholesterol	47
[25]	This	47
[26]	On	46
[27]	Can	45
[28]	S	45

[29]	plant	34	[29]	Was	43
[30]	subjects	33	[30]	Acid	42
[31]	may	32	[31]	structure	40
[32]	study	31	[32]	Other	39
[33]	cbd	30	[33]	They	38
[34]	has	29	[34]	An	35
[35]	their	29	[35]	Not	35
[36]	but	28	[36]	Their	35
[37]	performance	28	[37]	strand	34
[38]	test	27	[38]	amino	33
[39]	dose	26	[39]	More	31
[40]	other	26	[40]	Only	30
[41]	S	26	[41]	Form	27
[42]	they	26	[42]	Acids	26
[43]	An	24	[43]	Two	26
[44]	At	24	[44]	All	25
[45]	It	24	[45]	Also	25
[46]	alcohol	23	[46]	Have	25
[47]	also	23	[47]	These	25
[48]	production	23	[48]	known	24
[49]	after	22	[49]	May	23
[50]	cannabinoid	22	[50]	sequence	23
[51]	cannabinoids	22	[51]	But	22
[52]	drug	21	[52]	At	21
[53]	effects	21	[53]	called	21
[54]	cbc	20	[54]	Has	21
[55]	Kg	20	[55]	many	21
[56]	studies	20	[56]	Some	21
[57]	traffic	20	[57]	double	20
[58]	which	20	[58]	strands	20
[59]	content	19	[59]	Had	19
[60]	both	18	[60]	Ldl	19
[61]	more	18	[61]	Were	19
[62]	only	18	[62]	Each	18
[63]	many	17	[63]	Hdl	18
[64]	some	17	[64]	He	18
[65]	doses	16	[65]	Liver	18
[66]	found	16	[66]	particles	18
[67]	its	16	[67]	carbohydrates	17
[68]	produced	16	[68]	Helix	17
[69]	results	16	[69]	Its	17
[70]	placebo	15	[70]	Than	17
[71]	plants	15	[71]	because	16
[72]	under	15	[72]	Cells	16
[73]	Uv	15	[73]	chemical	16
[74]	between	14	74	Into	12
[75]	greater	14	[75]	Diet	16
[76]	material	14	[76]	disease	16
[77]	same	14	[77]	First	16
[78]	smoking	14	[78]	Same	16
[79]	drivers	13	[79]	foods	15

[80]	following	13	[80]	His	15
[81]	fungi	13	[81]	Large	15
[82]	high	13	[82]	molecular	15
[83]	highway	13	[83]	Most	15
[84]	however	13	[84]	total	15
[85]	influence	13	[85]	between	14
[86]	most	13	[86]	enzymes	14
[87]	No	13	[87]	food	14
[88]	present	13	[88]	such	14
[89]	resin	13	[89]	when	14
[90]	cigarettes	12	[90]	about	13
[91]	group	12	[91]	bases	13
[92]	all	11	[92]	been	13
[93]	associated	11	[93]	body	13
[94]	demonstrated	11	[94]	cell	13
[95]	distance	11	[95]	chain	13
[96]	glands	11	[96]	function	13
[97]	low	11	[97]	nucleotides	13
[98]	opportunistic	11	[98]	phenylalanine	13
[99]	reported	11	[99]	So	13
[100]	use	11	[100]	them	13

1. **What are the most significative items in that list?**

The most significative items are:

NOUN

Corpus A		**Corpus B**	
Items	**Frequency**	**Items**	**Frequency**
marijuana	101	dna	77
cannabis	42	vitamin	68
Plant	34	one	60
subjects	33	protein	60
Study	31	proteins	56
performance	28	cholesterol	47
Test	27	acid	42
Dose	26	structure	40
alcohol	23	strand	34
production	23	amino	33
cannabinoid	22	form	27
cannabinoids	22	acids	26
Drug	21	sequence	23
Effects	21	strands	20
kg	20	ldl	19
studies	20	liver	18
traffic	20	particles	18
content	19	carbohydrates	17
doses	16	helix	17
results	16	cells	16

results	16	cells	16
placebo	15	diet	16
plants	15	disease	16
material	14	foods	15
drivers	13	enzymes	14
fungi	13	food	14
highway	13	bases	13
influence	13	body	13
resin	13	cell	13
cigarettes	12	chain	13
group	12	function	13
distance	11	nucleotides	13
glands	11	phenylalanine	13

Acronyms	**Frequency**	**Acronyms**	**Frequency**
Thc	103	Hdl	18
Cbd	30		
Cbc	20		
Uv	15		

VERBS

Corpus A

Auxiliar	**Frequency**
Is	65
Are	44
Have	45
Has	29
Were	54
Was	50
Infinitive	**Frequency**
Be	38
Modal	**Frequency**
May	32
Past participle	**Frequency**
been	34
Found	16
Produced	16
Associated	11
Demonstrated	11
Reported	11

Corpus B

Auxiliar	**Frequency**
Is	152
Are	117
Have	25
Has	21
Were	19
Was	43
Had	19
Infinitive	**Frequency**
Be	57
Modal	**Frequency**
Can	45
May	23
Past participle	**Frequency**
known	24
called	21
Been	13
Associated	11
Demonstrated	11
Reported	11

ADJECTIVES

Corpus A

Demonstrative	**Frequency**
That	75
These	44

Corpus B

Demonstrative	**Frequency**
That	100
These	25

These	44
Possessive	**Frequency**
Their	29
Its	16
Qualifier	**Frequency**
More	18
Many	17
Greater	14
Following	13
High	13
Most	13
Present	13
Low	11
Opportunistic	11

These	25
Possessive	**Frequency**
Their	35
Its	17
His	**15**
Qualifier	**Frequency**
Other	39
Many	21
Some	21
Double	20
Same	16
First	16
Chemical	16
Most	15
Large	15
Molecular	15
Total	15

ADVERBS

Corpus A	
Adverbs	**Frequency**
By	64
Not	36
Also	23
Only	18
Between	14
However	13

Corpus B	
Adverbs	**Frequency**
Or	65
Not	35
More	31
Only	30
All	25
Also	25
Such	14
When	14
So	13

PREPOSITIONS

Corpus A	
Prepositions	**Frequency**
Of	332
In	212
To	185
as	68
On	63
For	62
With	60
From	57
But	28
After	22
At	24
Between	14
Under	15

Corpus B	
Prepositions	**Frequency**
Of	416
In	244
To	184
As	102
For	72
By	67
With	50
From	48
On	46
But	22
At	21
Between	14

ARTICLES

Corpus A		Corpus B	
definite	**Frequency**	**Definite**	**Frequency**
The	444	The	649
Indefinite	**Frequency**	**Indefinite**	**Frequency**
A	213	A	201
An	**24**	**An**	**35**

Corpus A Core vocabulary:

Nouns	Verbs	Adjectives	Adverbs
marijuana	is	their	by
cannabis	were	other	not
plant	was	its	also
subjects	have	greater	however
study	are	following	only
performance	be	high	all
test	been	low	between
dose	may	opportunistic	
alcohol	has	only	
production	found	all	
cannabinoid	produced	present	
cannabinoids	associated	that	
drug	demonstrated	these	
effects	Reported	both	
kg	test	many	
studies		some	
traffic		same	
content		driving	
doses			
results			
placebo			
plants			
material			
drivers			
fungi			
highway			
influence			
resin			
cigarettes			
group			
distance			
glands			
THC			
CBD			
CBC			

UV			

Corpus B Core vocabulary

Nouns	Verbs	Adjectives	Adverbs
Dna	Is	That	Or
Vitamin	Are	Other	Not
One	Be	Their	More
Protein	Can	These	Only
Proteins	Was	Its	All
Cholesterol	Have	Chemical	Also
Acid	Known	His	Such
Structure	May	Large	When
Strand	Called	Molecular	So
Amino	Has	Total	About
Form	Had	Two	
Acids	Were	Many	
Sequence	Been	Double	
Strands	Form	First	
Ldl		Each	
Liver		Some	
Particles		Same	
Carbohydrates		Most	
Helix			
Cells			
Diet			
Disease			
Foods			
Enzymes			
Food			
Bases			
Body			
Cell			
Chain			
Function			
Nucleotides			
Phenylalanine			

QUALITATIVE ANALYSIS

The qualitative analysis allows us to see in what ways biochemistry language is different from everyday language.

1. **Focus on some of the most frequent words of your list, for example technical or semi-technical vocabulary, modals, verbs, connectors, etc.**

- **Some examples of technical or semi- technical words:**

In www.edict.com.hk I found that the total number of words parsed in **Corpus A** is 8490, the words in the 2000 most frequent list are 5065 (59.66 %) and Number of words in the 2-5K List is 989 (11.65 %). Finally, total number of words not in either list is 2436 (28.69 %). While in **Corpus B**, the total number of words parsed in this text is 9056, the words in the 2000 Most Frequent List are 5871 (64.83 %), the number of words in the 2-5K List is 917 (10.13 %) and the total number of words not in either list is 2268 (25.04 %).
In both corpus, the words not present in either list are typical of a specific language of biochemistry i.e:
Corpus A: ecology , Cannabis, Chemical, Cannabinoids, compounds

Corpus B: Myoglobin, alpha, helices, X-ray, crystallography, compounds

As we can see the item *Compound* is present in both corpus, (in Corpus A, number of hits = 14; in Corpus B, number of hits = 5). In Edict dictionary we find the following meaning:

as ***adjective***
1. compound *adjective* (botany) especially of leaf shapes; composed of several similar parts or lobes
2. compound *adjective* (zoology) composed of many distinct individuals united to form a whole or colony: "coral is a compound organism"
See also: colonial
3. compound *adjective* consisting of two or more substances or ingredients or elements or parts: "soap is a compound substance"; "housetop is a compound word"; "a blackberry is a compound fruit"
as ***noun***
4. compound *noun* a whole formed by a union of two or more elements or parts
5. compound *noun* (chemistry) a substance formed by chemical union of two or more elements or ingredients in definite proportion by weight
See also: chemical compound
6. compound *noun* an enclosure of residences and other building (especially in the Orient)
as ***noun (countable)***
7. compound *noun (countable*) If something is a compound of different things, it consists of those things.: Honey is basically a compound of water, two types of sugar, vitamins and enzymes.
See also: combination , composition , blend , mixture , union , fusion

In both corpus the item *Compound* is used as a noun*: a substance formed by chemical union of two or more elements or ingredients in definite proportion by weight*. But as we can notice from our search it can have different meaning according to a specific area, i.e. in zoology it is an adjective (n. 2 of the table above)

- **Some examples of a same item used in different forms:**

In both corpus there are some examples of items used in different forms , as noun and verbs:
In Corpus A *Test*
In corpus B ***Form***

- **Collocation of Cause effect items: since, because, as a result of and therefore:**

In both corpus there is a low use of cause effect items

Frequency

Item	Corpus A	Corpus B
Since	**1**	**6**
Because	**4**	**11**
As a result of	**0**	**2**
Therefore	**2**	**2**

Their use seems not different from that in general common English , some examples:
CORPUS A.:
- This is because road-tracking is primarily…..
- They further ventured that since the production of terpeness is….
- …..processing and are therefore more accessible for compensation…..

CORPUS B:
-The name arises because it was once considered a sing….
-However, since there are just four possible….
-It is therefore not a vitamin for them.

2. **What are the most common forms of nominalization (if any)?**

In both corpus there are few nominalization (adjectives, verbs and adverbs transformed into nouns)

Corpus A	Corpus B
Driving (63) **Production (23)** **Smoking (14)**	**Function (13)** **Living(7)** **Production (6)** **Arrangement (5)**

5. **Can you detect compound noun phrases that are typical of this variety, (i.e. nominal phrases)?**

Some examples of compound noun phrases:
Corpus A):
- **…..**flower-associated bracts…(in general English: bracts that are associated with flower
-….insensitive to drug-induced changes(…changes caused by drug);

Corpus B):
-Animal-derived foods contain all of….(….Foods that are derived from animals…)
-….when faced with life-threatening disease (…..disease that threatens our life…)

6. **Are there any passive forms? How often are they used and when?**

In Corpus A there is a large presence of **passive form** of the **perfect tenses (have been …ed (15) / has been …ed (11)), and of simple past (were…ed (48)) while in Corpus B there is a large presence of the simple present passive form** [**is/are …ed (51)**]**. In Corpus A t**hey are used to explain the results or observations about studies or testing. For

example: *Stalked glands have been observed; all products which have been found in analyses of Cannabis*
In **Corpus B** they are used to focus the attention on the object and not on the subject, for example: *The primary structure is held together ; its native state is often described.*

7. **What are the most common verbal tenses to be found?**

In both corpus the most common verbal tenses are passive forms and modal verbs:

- The modal **May** appears 32 times in Corpus A and 22 times in Corpus B.

In both corpus May expresses possibility, not different from general English language. This large use of may suggests us that the writers of the article are making supposition on the basis of experiments done. I.e.:

-The cannabinoids may also serve as a purely….
-Proteins may shift between several similar…….

- The modal **Can** appears only 5 times in corpus A, and 42 times in corpus B and expresses possibility and capability.

- there can be many different secondary motifs present in one single protein molecule..(possibility); Regulation can involve a protein's shape or concentration.(capability)

8. **Some more information …**

In both corpus there is a large use of acronyms (nouns formed by the initials of other words) and compound words, for example:

Acronyms: In corpus A: THC(103); CBD(30) ; CBC(2o) UV(15);
In corpus B: DNA (77); HDL(18) drug-placebo; immuno-compromised; oral-mucosal

Compound words: in Corpus A: drug-placeb; active-drug;
In corpus B: double-helical; straight- chain; sub-structures-alpha

9. **Conclusion**

From this analysis we can notice that biochemistry language, like other scientific languages, tends to simplify the syntax: short sentences, taken by nouns that becomes the nucleus of the sentence, compound noun phrases, compound nouns, acronyms. A large use of passive forms gives to the subject a sense of abstractness, only focusing on the objects without pointing out the presence of the writer.

Rosa Derosas