Corpus Analysis
Roberta Meloni

QUANTITATIVE ANALYSIS

1) MOST FREQUENT 100 WORDS

[1]	the	473	[35]	lumbar	50	[69]	baseline	25
[2]	of	423	[36]	mass	50	[70]	purging	25
[3]	in	382	[37]	up	50	[71]	BN	24
[4]	and	345	[38]	follow-	49	[72]	bulimic	24
[5]	with	227	[39]	this	48	[73]	bmi	23
[6]	a	174	[40]	from	47	[74]	have	23
[7]	An*	157	[41]	function	46	[75]	menstrual	22
[8]	to	144	[42]	than	45	[76]	our	22
[9]	at	142	[43]	eumenorrheic	42	[77]	total	22
[10]	for	142	[44]	had	41	[78]	age	21
[11]	patients	141	[45]	normal	41	[79]	duration	21
[12]	bone	130	[46]	low	39	[80]	more	21
[13]	were	124	[47]	amenorrheic	38	[81]	percent	21
[14]	leptin	118	[48]	between	36	[82]	also	20
[15]	nervosa	110	[49]	compared	36	[83]	decreased	20
[16]	was	90	[50]	groups	36	[84]	studies	20
[17]	women	87	[51]	on	36	[85]	these	20
[18]	weight	83	[52]	mean	35	[86]	who	20
[19]	eating	78	[53]	not	35	[87]	anorexic	19
[20]	that	77	[54]	significantly	35	[88]	measured	19
[21]	bmd	76	[55]	loss	34	[89]	osteopenia	19
[22]	body	76	[56]	or	34	[90]	recovered	19
[23]	group	72	[57]	subjects	34	[91]	turnover	19
[24]	by	66	[58]	disorder	33	[92]	type	19
[25]	as	63	[59]	may	32	[93]	which	19
[26]	bulimia	63	[60]	density	31	[94]	analysis	18
[27]	study	63	[61]	are	30	[95]	associated	18
[28]	levels	62	[62]	all	29	[96]	controls	18
[29]	serum	60	[63]	amenorrhea	29	[97]	criteria	18
[30]	spine	60	[64]	data	27	[98]	differences	18
[31]	anorexia	58	[65]	higher	27	[99]	found	18
[32]	fat	58	[66]	osteoporosis	27	[100]	illness	18
[33]	is	54	[67]	binge	26			
[34]	be	50	[68]	disorders	26			

*** The frequency list doesn't distinguish "an", indefinite article (58 occurrences), from "AN", which stands for "*Anorexia Nervosa*" (99 occurrences).**

2) MOST SIGNIFICANT ITEMS

Excluding articles, prepositions and conjunctions, the most significant categories of my corpus are: **nouns** (38 items), **verbs** (13 items), **adjectives** (22 items), **pronouns** (5 items) and **adverbs** (2 items).

NOUNS	**(38 items; 1816 nouns)**		
1	[7] /[31]	AN / anorexia	157
2	[11]	patients	141
3	[12]	bone	130
4	[23]/ [50]	group(s)	108
5	[26]/ [71]	bulimia/ BN	87
6	[17]	women	87
7	[27]/ [84]	study(-ies)	83
8	[18]	weight	83
9	[22]	body	76
10	[28]	levels	62
11	[29]	serum	60
12	[30]	spine	60
13	[58]/ [68]	disorder(s)	59
14	[32]	fat	58
15	[36]	mass	50
16	[41]	function	46
17	[52]	mean	35
18	[55]	loss	34
19	[57]	subjects	34
20	[60]	density	31
21	[63]	amenorrhea	29
22	[64]	data	27
23	[66]	osteoporosis	27
24	[69]	baseline	25
25	[78]	age	21
26	[79]	duration	21
27	[81]	percent	21
28	[89]	osteopenia	19
29	[91]	turnover	19
30	[92]	type	19
31	[94]	analysis	18
32	[96]	controls	18
33	[97]	criteria	18
34	[98]	differences	18
35	[100]	illness	18
36	[19]	eating	10 (out of 78)
37	[70]	purging	5 (out of 25)
38	[77]	total	2 (out of 22)

VERBS			(13 items; 543 verbs)
1	[13]	were	124
2	[16]	was	90
3	[33]	is	54
4	[34]	be	50
5	[44]	had	41
6	[49]	compared	36
7	[59]	may	32
8	[61]	are	30
9	[74]	have	23
10	[88]	measured	19
11	[99]	found	18
12	[95]	associated	17(out of 18)
13	[90]	recovered	9 (out of 19)

2.1) VERB ANALYSIS

AUXILIARIES		(5 ITEMS)	
1	[13]	were	72 (out of 124)
2	[16]	was	59 (out of 90)
3	[34]	be	19 (out of 50)
4	[33]	is	18 (out of 54)
5	[44]	had	12 (out of 41)
6	[74]	have	8 (out of 23)
7	[61]	are	7 (out of 30)

MODALS		(1 ITEM)	
1	[59]	may	32

ACTIVE VOICE

PRESENT TENSES (130 out of 543)

Present Simple	
is	36 (out of 54)
may	32
be	31 (out of 50)
are	23 (out of 30)
have	8 (out of 23)

PAST TENSES (204 out of 543)

	P. Simple	P. Cont.	Perfect	P. Part.
were	51 (out of 124)	1 (out of 124)		
was	31 (out of 90)			
had	29 (out of 41)		12 (out of 41)	
compared	3 (out of 36)		1 (out of 36)	24(out of 36)
have			15 (out of 23)	
measured				7 (out of 19)
associated			1 (out of 18)	10(out of 18)
found	9 (out of 18)		1 (out of 18)	
recovered	3 (out of 9)		3 (out of 9)	3 (out of 9)

PASSIVE VOICE

PRESENT TENSES (44 out of 543)

Present Simple		
be	19 (out of 50)	could be gathered may be mediated should be noted ……
is	18 (out of 54)	is characterized is related to is reflected by ……
are	7 (out of 30)	are presented are correlated are summarized ……

PAST TENSES (166 out of 543)

Past Simple		
were	72 (out of 124)	were studied were tested for were observed ……
was	59 (out of 90)	was performed by was compared with was measured by ……
measured	12 (out of 19)	were measured by was measured ……
compared	8 (out of 36)	were compared by were compared with was compared between ……
found	8 (out of 18)	were found was found ……
associated	7 (out of 18)	were associated with may be associated has been associated ……

ADJECTIVES (22 items; 788 verbs)			
1	[15]	nervosa	233*
2	[19]	eating	68 (out of 78)
3	[35]	lumbar	50
4	[39]	this	43 (out of 48)
5	[43]	eumenorrheic	42
6	[45]	normal	41
7	[46]	low	39
8	[47]	amenorrheic	38
9	[65]	higher	27
10	[67]	binge	26
11	[72]	bulimic	24
12	[32]	fat	22 (out of 58)
13	[75]	menstrual	22
14	[76]	our	22
15	[77]	total	20 (out of 22)
16	[83]	decreased	20
17	[70]	purging	20 (out of 25)
18	[85]	these	20
19	[87]	anorexic	19
20	[90]	recovered	10 (out of 19)
21	[95]	associated	1 (out of 19)
22	[20]	that	1 (out of 77)

* The number shown is the sum of the number of times the adjective "nervosa" appears alone (110) and the number of times it appears combined with "anorexia" and "bulimia", as "AN" (99) and "BN" (24).

2.2) ADJECTIVE ANALYSIS

QUALIFYING ADJ.	DEMONSTRATIVE ADJ.	COMPARATIVE ADJ.	POSSES. ADJ.
nervosa 110	this 43 (out of 48)	higher 27	our 22
eating 68 (out of 78)	these 20 (out of 20)		
lumbar 50	that 1 (out of 77)		
eumenorrheic 42			
normal 41			
low 39			
amenorrheic 38			
binge 26			
bulimic 24			
fat 22 (out of 58)			
menstrual 22			
total 20 (out of 22)			
decreased 20			
purging 20 (out of 25)			
anorexic 19			
recovered 10 (out of 19)			
associated 1 (out of 19)			

PRONOUNS (5 items; 60 pronouns)			
1	[86]	who	20
2	[93]	which	19
3	[20]	that	11 (out of 77)
4	[39]	this	5 (out of 48)
5	[80]	more	5 (out of 21)

2.3) PRONOUN ANALYSIS

RELATIVE PRON.	DEMONSTRATIVE PRON.	INDEFINITE PRON.
who 20	this 5 (out of 48)	more 5 (out of 21)
which 19	that 7 (out of 77)	
that 7 (out of 77)		

ADVERBS (2 items; 45 adverbs)			
1	[54]	significantly	35
2	[80]	more*	10 (out of 21)

2.3) ADVERB ANALYSIS

MANNER	QUANTITY
Significantly 35	more 10 (out of 21)

*** "More" occurs 21 times in my frequency list, but I my analysis takes into consideration only the occurrences as a pronoun and an adverb, excluding the 6 times it occurs as an indefinite determiner.**

3) MINIMAL CORE VOCABULARY

NOUNS (38 items; 1816 nouns)			
1	AN / anorexia	20	density
2	patients	21	amenorrhea
3	bone	22	data
4	group(s)	23	osteoporosis
5	bulimia / BN	24	baseline
6	women	25	age
7	study(-ies)	26	duration
8	weight	27	percent
9	body	28	osteopenia
10	levels	29	turnover
11	serum	30	type
12	spine	31	analysis
13	disorder(s)	32	controls
14	fat	33	criteria
15	mass	34	differences
16	function	35	illness
17	mean	36	eating
18	loss	37	purging
19	subjects	38	total

VERBS (13 items; 543 verbs)	
1	To be
2	To have
3	To compare
4	May
5	To measure
6	To find
7	To associate
8	To recover

ADJECTIVES (22 items; 788 verbs)			
1	nervosa	12	fat
2	eating	13	menstrual
3	lumbar	14	our
4	this	15	total
5	eumenorrheic	16	decreased
6	normal	17	purging
7	low	18	these
8	amenorrheic	19	anorexic
9	higher	20	recovered
10	binge	21	associated
11	bulimic	22	that

ADVERBS (2 items; 45 adverbs)	
1	significantly
2	more

QUALITATIVE ANALYSIS

1) CONCORDANCE OF THE MOST RELEVANT WORDS

NOUNS	
Anorexia	• **In anorexia nervosa, leptin values are…….** • **…females with anorexia or bulimia nervosa.**
Patients	• **…..patients were studied during a phase…..** • **……a mixed sample of patients……** • **Anorexic patients…….** • **….bulimic patients……..**
Study(-ies)	• **The goal of this study was to assess…..** • **This study prospectively investigated…** • **Studies were also conducted…….**
Weight	• **body weight** • **weight loss** • **low weight**
Data	• **data were collected….** • **data indicate that…….** • **data demonstrate…..**
Eating	• **recurrent binge eating** • **abnormal eating**

These are some of the most frequent nouns in my corpus.

- **"Anorexia"** is the most frequent of all (157 occurrences), as it is strictly connected to the main topic of these abstracts: Eating Disorders. It is often associated with the terms "nervosa" (99), which defines a specific kind of this illness, and "bulimia", an eating disorder that presents similar features.
- **"Patients"/ "Study"/ "Data"** are very common terms, representative not just of the topic of my corpus, but of scientific abstracts as a whole. Observing their concordances, we can deduce that they are strictly correlated, as patients are often subject of studies and experiments carried out to collect data useful to gather information. The term "Data" is, therefore, typical of scientific fields, while the term "Study", which comes from everyday language, acquires here the specific meaning of "scientific investigation, research".
- The frequency of the term **"Weight"** is justified by the subject of the abstract. Its concordances ("body", "loss" and "low") point out the obsession of eating disorder patients for weight loss.
- **"Eating"** is another frequent word as a noun as well as an adjective. It is always associated with the terms "binge" and "abnormal", to define the symptomatic disorder that affects bulimic patients.

VERBS

ACTIVE VOICE

PRESENT TENSES (130 out of 543)

Present Simple	
is	36 (out of 54)
may	32
be	31 (out of 50)
are	23 (out of 30)
have	8 (out of 23)

PAST TENSES (204 out of 543)

	P. Simple	P. Cont.	Perfect	P. Part.
were	51 (out of 124)	1 (out of 124)		
was	31 (out of 90)			
had	29 (out of 41)		12 (out of 41)	
compared	3 (out of 36)		1 (out of 36)	24(out of 36)
have			15 (out of 23)	
measured				7 (out of 19)
associated			1 (out of 18)	10(out of 18)
found	9 (out of 18)		1 (out of 18)	
recovered	3 (out of 9)		3 (out of 9)	3 (out of 9)

PASSIVE VOICE

PRESENT TENSES (44 out of 543)

Present Simple		
be	19 (out of 50)	could be gathered may be mediated should be noted ……
is	18 (out of 54)	is characterized is related to is reflected by ……
are	7 (out of 30)	are presented are correlated are summarized ……

PAST TENSES (166 out of 543)

Past Simple		
were	72 (out of 124)	were studied were tested for were observed ……
was	59 (out of 90)	was performed by was compared with was measured by ……
measured	12 (out of 19)	were measured by was measured ……
compared	8 (out of 36)	were compared by were compared with was compared between ……
found	8 (out of 18)	were found was found ……
associated	7 (out of 18)	were associated with may be associated has been associated ……

As shown in the table above, the verbs used in the abstract appear in the active as well as in the passive form, in both present and past tenses.

- Among the most frequent verbs there are "**To be**"(347) and "**To have**"(64), both used as full verbs(172 and 37, respectively) and as auxiliaries (175 and 27, respectively). The Past tense is more common then the present, and the Past Simple is the most frequent tense used, more common than the Present Perfect because it refers to experiments and studies already finished.
- There is only one Continuous form ("was having") in the list analysed, and this can be justified by the fact that this form occurs rarely in scientific abstracts.

- The frequent use of the passive is typical of scientific abstracts, and its aim is to describe standard procedures impersonally. It is found in alternative to the active "We-form", used by groups of scholars who describe their own procedural choices or the results of their experiments.
  Passive forms are accompanied by the auxiliaries "**To be**" and "**To have**", and often by the modals **could/should/may** (see the concordances in the table).
- Other important verbs are "**measured**"(19), referred to body weight/fat, and "**compared**"(36), which refers to data comparisons.
- The most frequent modal verb is "**may**" (32), used to express "possibility" rather than certainty, (eg. when talking about results), and as *hedging*, when writers distance from what they state. Some examples are: "**may explain**", "**fat mass may be important for preservation**", "**body fat may play a central role**".

ADJECTIVES	
Nervosa	• **Bulimia Nervosa** • **Anorexia Nervosa**
Low	• **low caloric requirements** • **low weight**
Bulimic	• **bulimic subject** • **bulimic women**
Fat	• **fat mass**
Anorexic	• **anorexic patients** • **anorexic women**

The collocations reported, show us that the choice and the frequency of adjectives depends on the frequency of the noun they refer to. All the adjectives in the list above are associated with the most frequent words of the corpus, and are specific of the topic they describe, rather than of scientific abstracts. "**Low**", for instance, occurs 39 times, and it is often associated with "**weight**", which occurs 83 times.

2) TYPICAL LINGUISTIC PATTERNS

Indicate that Suggest that Found that

All these linguistic patterns are typical of scientific abstracts, and refer to data and results of studies and research: "**This study found that**"; "**These results suggest that**"; "**Our data indicate that**".
Although in some cases they are used as synonyms, we should notice that while "**indicate that**" expresses certainty, "**suggested that**" conveys a certain degree of uncertainty, and can express "hedging". "**Found that**" occurs always in the past tense, as a sort of *invariable formula* referred to experiments or studies already carried out.

Compared with

This pattern occurs 28 times, frequently collocated with the terms "values" ("**Values were compared with**"), and "patients" ("**anorexic patients […]compared with**"). It refers to the data comparison typical of scientific hypotheses and studies.

Associated with

The pattern is quite frequent (17 times), and is used to express relationships of causes and effects, as shown in the following examples: "**this does not appear to be associated with leptin gene mutations**"; "**Recovery from illness was associated with near-normal bone size…**".

3) NOMINALIZATIONS

Very few nominalizations are present in this corpus. The most relevant is "**eating**" (see noun concordances).

4) NOMINAL PHRASES

The most common nominal phrases define the type of patients analysed.

- **Eating disorder patients**
- **Low-weight women**
- **Binge-eating subtypes**
- **Anorexic/ Bulimic patients**

4) ACRONYMS

There are few acronyms in the corpus, but their frequency is very high. The most important are:

- **AN stands for Anorexia Nervosa**
- **BN stands for Bulimia Nervosa**
- **BMI stands for Body Mass Index**

5) STRUCTURE

The texts taken into consideration present all the same structure, typical of scientific articles, divided into three main sections:

A) *INTRODUCTION*

- **Title:** it sums up the topic treated in the article, with descriptive and enticing words. It's followed by full names and addresses of the authors.

- **Abstract:** it is an overview of experiment, results and discussion. They summarize the key findings of the article and introduce the problem in about 150 words.

- **Introduction:** It summarizes the general problem area, and focuses on the particular problem going to be investigated. It typically starts introducing the general problem and carries on with the discussion of the relevant literature (if any), the personal encounters with the problem and the review of previous experiments, to finish with mentioning relevant theories about the topic. (It is not always present).

B) *DESCRIPTION OF THE STUDY/EXPERIMENT*

- **Material/Subjects & Methods:** this part describes the subjects and the materials required for the research/study, indicate the methodology (eg how the experiment was run), and list any problems encountered.

- **Results:** this section reports the data gathered during the experiment/study in neat tables and their levels of significance, including figures and graphs.

C) *CONCLUSION*

- **Discussion:** it is an objective interpretation of the results, with a final summary of the report and references to further studies.

- **Acknowledgments:** in this part the author thanks for the help he received during his/her study. (It is not always present).

- **Footnotes:** they contain additional information.

- **References:** a list of all the papers cited in text.

6) CONCLUSIONS

The analysis carried out has been extremely useful to me because, although it was not the first time I approached ESP texts, I have become aware of many important aspects typical of scientific writings. Among these:

- The **language used** is **not very far from GE**, and the differences it presents can be exploited in teaching, focusing on the peculiar features of ESP (I would define my corpus language "SOFT ESP", as it is not too technical);
- The language is always very **objective and impersonal**, thanks to the use of the 3rd person and passive verbs;
- **Nominalizations**, **nominal phrases** and **linguistic patterns** are not necessarily many, but their frequency is very high;
- The most common tense is the **Past Simple**, used especially to describe the experiments carried out. This is followed by the **Present Simple**, also common in ESP. The most frequent person is the **3rd** (singular and plural), so as to convey impersonality;
- Though there is a high percentage of **passive verbs**, compared to GE, I realized they are not more frequent than active forms;
- **Modal verbs** are very frequent in ESP texts, though they **"lose"** most of the meanings and **functions** they have in **GE** (asking permission and polite requests, for instance), and **express** only probability/**possibility** and **hedging**;
- The most **frequent terms refer to the topic treated** in the articles and, as my tables show, there is a strict **connection** between the **frequency** of **nouns** and the occurrence of **adjectives**: the higher the frequency of a noun, the higher the occurrence of an adjective associated to it;
- These kind of articles present very **few connectors**. This might be explained with the aim of scientific papers: reporting and summarizing things very briefly;
- The **technical vocabulary** depends on the **theme** with which the article deals. Technical or sub-technical terms often come from GE, but they acquire a specific meaning, according to the field;
- Like literary genres, ESP texts present a fix structure that varies according to the different field or domain. Medicine articles, like almost all Scientific articles, are written following a **three-section structure**, each part with specific aims and respecting a standard length.

6) INFLUENCES OF THIS ANALYSIS ON MY SYLLABUS

The present corpus will be object of study in an ESP course for adults (Medicine domain: Endocrinology), during the second module, (second unit) with the purpose of studying, understanding and re-elaborating typical register and rhetorical features of Medical English like:

- Technical vocabulary, studied through concordances;
- Use of Simple Past and Present;
- Passive voices;
- Modal verbs and their meaning in ESP;
- Nominalizations/ Nominal Phrases;
- Typical linguistic pattern;
- Acronyms;
- Three-section-structure (and relative rhetorical devices)

Special attention will be paid to the typical structure of scientific articles, and to the impersonal techniques used by authors.
This study will allow students to familiarize with the language issues listed above, and to observe typical scientific concordances (through specific software and web sites), in order to enable them to plan, structure and write detailed articles on Medicine.

Roberta Meloni