CORPUS ANALYSIS
SERENA SOTGIA

QUANTITATIVE ANALYSIS

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

In the list of the 100 most frequent words the most present items are :

1) Articles
2) Prepositions
3) Conjunctions
4) Verbs
5) Nouns
6) Adverbs
7) Adjectives
8) Determinatives

ARTICLES	**DEFINITE**	**FREQUENCY**	INDEFINITE	FERQUENCY	TOT
	The	636	A	165	843
			An	42	

PREPOSITIONS	TYPE	FREQUENCY	TOT
	Of	239	808
	Between	38	
	With	49	
	In	177	
	To	119	
	For	52	
	At	37	
	From	32	
	On	26	
	By	39	

CONJUNCTIONS	TYPE	FREQUENCY	TOT
	And	158	255
	Or	25	
	That	60	
	Than	12	

ADVERBS	TYPE	FREQUENCY	TOT
	When	20	191
	So	11	
	There	22	
	Therefore	14	
	More	12	
	As	74	
	Since	13	
	Not	14	
	Around	11	

	TYPE		FREQUENCY	TOT
VERBS	BE	Is	133	336
		Are	62	
		Be	34	
		Was	13	
	HAVE	Have	15	
		Had	11	
		Has	18	
	MODALS	Will	18	
		Can	32	

	TYPE	FREQUENCY	TOT
NOUNS	Electron/s	79	734
	Atom/s	76	
	Energy	33	
	Ion/s	12	
	Dipole	34	
	Figure	31	
	Interaction	27	
	Force	25	
	Bond	24	
	State	24	
	Behaviour	23	
	Quasiparticle/s	37	
	Field	22	
	Nucleus	21	
	Attraction	19	
	Bonding	18	
	Charge	18	
	Temperature/s	31	
	Materials	17	
	Experiments	15	
	Moment	15	
	Subshells	15	
	System	14	
	Carbon	13	
	Number	12	
	Theory	12	
	Valence	12	
	Length	11	
	Molecule	11	
	Net	11	
	Shells	11	
	Solid	11	

	TYPE	FREQUENCY	TOT
ADJECTIVES	Other	26	183
	Superconducting	17	
	Found	16	
	High	16	
	Strong	16	
	Electric	14	
	Negative	14	
	Normal	14	
	All	12	
	Induced	11	
	Very	15	
	Only	12	

	TYPE	FREQUENCY	TOT
DETERMINATIVES	This	41	145
	Each	24	
	Two	41	
	our	14	
	These	13	
	Such	12	

2. What are the **most significant items** in that list? (for example, verbs, tenses, aspect, gender, personal pronouns, auxiliaries, modals, etc.)

Among the various items composing the most frequent 100-word list an important role is played by articles (843), prepositions (808), nouns (734), verbs (336) and conjunctions (262).

3. If it is the case, enlarge your range of frequency, and consider the first 200 / 400 / 500 or 1000 most frequent words in your corpus in order to be able to carry on the following task.

4. Create a **core vocabulary** of the target specialised language for each of the four word classes, i.e. nouns, verbs, adverbs and adjectives. (Just copy, paste the list, then delete all but the verbs, nouns, etc.)

Core Vocabulary

[6]	is	133
[11]	are	62
[12]	electrons	59
[15]	atoms	47
[22]	be	34
[23]	dipole	34
[24]	energy	33
[25]	can	32
[28]	figure	31
[29]	atom	29
[31]	interaction	27
[33]	other	26
[34]	force	25
[36]	bond	24
[38]	state	24
[39]	behavior	23
[40]	quasiparticle	23
[41]	field	22
[44]	nucleus	21
[45]	electron	20

[47]	attraction	19
[48]	bonding	18
[49]	charge	18
[50]	has	18
[51]	temperature	18
[53]	materials	17
[54]	superconducting	17
[55]	found	16
[56]	high	16
[57]	strong	16
[58]	experiments	15
[59]	have	15
[60]	moment	15
[61]	subshells	15
[64]	electric	14
[65]	negative	14
[66]	normal	14
[69]	pairing	14
[70]	quasiparticles	14
[71]	spin	14
[72]	system	14
[74]	carbon	13
[76]	temperatures	13
[78]	was	13
[79]	all	12
[81]	ions	12
[83]	number	12
[85]	shown	12
[88]	theory	12
[90]	together	12
[91]	valence	12
[92]	around	11
[93]	had	11
[94]	induced	11
[95]	length	11
[96]	molecule	11
[97]	net	11
[98]	shells	11
[99]	so	11
[100]	solid	11

5. The **minimal core vocabulary**, that is a limited number of items that are essential in the target variety of language, is the final outcome of the quantitative analysis carried on so far. **It tells you what is used, but not how it is used**.

Minimal Core Vocabulary

	TYPE		FREQUENCY	TOT
VERBS	BE	Is	133	336
		Are	62	
		Be	34	
		Was	13	
	HAVE	Have	15	
		Had	11	
		Has	18	
	MODALS	Will	18	
		Can	32	

	TYPE	FREQUENCY	TOT
NOUNS	Electron/s	79	734
	Atom/s	76	
	Energy	33	
	Ion/s	12	
	Dipole	34	
	Figure	31	
	Interaction	27	
	Force	25	
	Bond	24	
	State	24	
	Behaviour	23	
	Quasiparticle/s	37	
	Field	22	
	Nucleus	21	
	Attraction	19	
	Bonding	18	
	Charge	18	
	Temperature/s	31	
	Materials	17	
	Experiments	15	
	Moment	15	
	Subshells	15	
	System	14	
	Carbon	13	
	Number	12	
	Theory	12	
	Valence	12	
	Length	11	
	Molecule	11	
	Net	11	
	Shells	11	
	Solid	11	

	TYPE	FREQUENCY	TOT
ADVERBS	When	20	191
	So	11	
	There	22	
	Therefore	14	
	More	12	
	As	74	
	Since	13	
	Not	14	
	Around	11	

	TYPE	FREQUENCY	TOT
ADJECTIVES	Other	26	183
	Superconducting	17	
	Found	16	
	High	16	
	Strong	16	
	Electric	14	
	Negative	14	
	Normal	14	
	All	12	
	Induced	11	
	Very	15	
	Only	12	

Qualitative Analysis

- Nouns

The most significant nouns the text is made up of are tightly related to the subject it deals with. As a results, they form a specific vocabulary whose items are not likely to be found in non specialized texts. Just to quote some of them, the most frequently recurring ones are (both in the singular and plural): ***electron, atom, quasiparticle, temperature, interaction, energy, dipole, force, bond, nucleus, field, attraction, moment, charge***.

As is apparent, most of them carry just one typically scientific meaning, which inevitably needs to be explained within the linguistic context it belongs to. Not immediately accessible, these nouns are not easily found in standard language and thus have the virtue of being unambiguous.

On the contrary, apparently easier words such as ***field, force, attraction, bond and moment***, might cause misunderstanding since they are here used with a meaning which is completely different from the one they carry in common language.

- Adjectives

A type of adjectives is present such as ***induced, found, normal, all, negative strong, electric, superconducting, only*** and so on, which simply serve as linguistic tools fit for descriptions, qualifications, definitions and restrictions.

Concerning the aspect of the adjectives present throughout the text (so, not only in the 100 most frequent words list), it must be said that most of them are superlatives and comparatives. We can actually find ***outer, outermost***, ***farthest***, ***nearest***, ***hardest***, ***highest, larger, lower, strongest, closest, more negative*** etc. . The possible reason for such a high frequency of superlatives and comparatives may be that continuous comparisons are one of the leading features of the corpus.

As for the disposition, this corpus contains (relatively) long strings of (compound) adjectives. Most of them actually limit to two/three juxtaposed adjectives, e.g. ***nearest previous inert element*** ; ***overall potential energy*** and

spherical negative-charge cloud. However, it is also possible to find longer strings of compound adjectives such as ***highly anisotropic effective magnetic interaction***.

Finally, a strong tendency to avoid the use of the relative pronoun followed by the infinitive is easily detectable throughout the corpus. Some exemplary instances are: ***the shells obeying to***........ instead of ***the shells that obey to.......***, and ***this induced dipole moment, depends on the electric field causing it*** instead of ....***that causes it***. Such a choice, entails a significant drop in the use of the relative pronoun *that/which* and of the infinitive of verbs as well, thus increasing the number of gerunds which are actual substitutes for more than one word at the time and function as real adjectives.

- Adverbs

Quite commonly used adverbs such as ***when, therefore, since***, ***so*** and ***as*** are broadly spread throughout the corpus. Their function is exactly the one they usually carry in non specialized texts: they actually complete the text, help it flow fluently, introduce simile, comparisons, and serve as complementary linguistic items to some concepts.

For instance, ***when*** always introduces a new scientific concept or environment, while ***therefore*** is always used to express the outcome resulting from the new state of affairs introduced by *when,* or to reinforce the content and meaning conveyed by a previous statement, e.g. : ***when the shells overlap, the electrons are shared by both the atoms [.......] ; they therefore cross the overlap region more frequently [......] .***

Furthermore, some other adverbs like ***similarly, generally, relatively, equally, thus, consequently, furthermore, although*** and ***however*** are also of great importance in the whole corpus since they all convey very important linguistic functions. Specifically, ***thus*** and ***consequently*** are used to come to final conclusions or considerations following a previous statement, while ***first, initially*, *then*** and ***furthermore*** mark the steps in the flow of a process. ***Although*** and ***however*** introduce a premise which lays the foundations of a following conceptual turning point, whereas ***similarly*** and ***equally*** provide a link backwards thus allowing parallels and comparisons.

Just to have an overview of the adverbs functioning in the corpus, it could be said that concepts are introduced and developed through stages, where adverbs (serving as real sequencers) follow one another according to the order below....... :

- Since/Due to → so/therefore → consequently/however
- First → thus → furthermore → Therefore
- Initially → then

.... where **Due to + Therefore** express a cause and effect relationship.

A final consideration about the adverb ***since*** is to be made. As stated before, in this corpus, *since* is mostly found as an introductory item at the beginning of main sentences, to the detriment of *because* which, on the contrary, is almost never used for this function. There are actually just very few cases in the whole corpus, in which ***because*** is used in main sentences instead of since: ***Because the systems of interest display near antiferromagnetic behaviour, one finds...........***; ***Because the systems of interest display near antiferromagnetic behaviour....*** .

- Verbs

Among all the verbs used in the corpus, ***to be*** and ***to have*** play a very important role both as main verbs and auxiliaries. As shown in the quantitative analysis table, their full conjugation is exploited in the present and past tense as well. However, ***is*** is the most recurring item, while ***have*** is evidently less frequent.

As stated before, they serve both as main verbs and auxiliaries. As for the former case, instances abound throughout the text: ***the result of this interaction is.....; the net force is the sum of....; it is instructive to consider....; the BCS theory had a significant impact..... etc***. .

Extremely more frequent, however, is the use of **to be** as an auxiliary to form the passive, both in the present and past tense, e.g. : ***the superconducting state is thus characterized by...; this distance is called.....; the presence of neutron and proton superfluids [...] was invoked in.....*** .

The same can be said about the use of ***to have*** concerning the formation of the Present Perfect Simple, the Past Perfect Simple and, in very few cases, the Present Perfect Continuous. Here follow some instances: ***this pseudogap behaviour, which has been calculated..........; the symmetry of the pairing state has been established; many different experiments had been used......; [....] as had been suggested by.....; such systems have been receiving.....*** .

Though the Simple Past is also employed in the corpus, Present and Past Perfect tenses use seems to prevail throughout the text. The explanation for their extensive use could be that, according to scientific texts general tendency, the phenomenon of " hedging " is here confirmed too.

Concerning other kinds of verbs present in the corpus, it must be said that phrasal verbs also abound. We can actually find expressions such as ***worked out; split up; make up; carried out; bring about; turned out***; ***come about.***

Further, *modal auxiliary verbs* are also used very often. It is possible to find almost all of them and recognize their traditional and specific function:

1)The model of the atom that we must use to understand its general behaviour......; 2) [....] over a long time period the electron would appear as a spherical negative charge....; 3) it predicts that any system of interacting

fermions <u>could</u> undergo [....] a superfluid transition; 4) we promised to withdraw our theory, <u>should</u> subsequent experiments show anything other than....... ; 5) The BCS superconducting transition is fundamentally different from what <u>might</u> happen if the pairs had formed well above ..... ; 6) we <u>can</u> show that a dipole moment creates an electric field for which... .

As is apparent, each modal conveys a specific meaning according to the sentence context it belongs to. That is, in sentence 1, **must** expresses the necessity of doing something because it is important in order to achieve an expected result. In sentence 2, **would** is used to say what is expected to happen given a base condition, that is " **over a long time period ".** In sentence 3, **could** introduces a future possibility in quite a formal way (more formal than **can**, which could have been used as well). In sentence 4, contrary to its usual function, **should** is not used to give an advice but to talk about something that may possibly happen. In sentence 5, **might** express a possibility too. In sentence 6, **can** expresses the concrete ability to do something.

Needless to say that modals are here used in an appropriate as much as effective way, according to the intent. Yet, it must also be pointed out that the text proceeding by hypotheses, assumptions and possible related results, some caution in statements is unavoidable. That's why most of the subtleties and shades and of meaning find a reason for existence.

As for the verb tenses, the array offered by the corpus includes: Simple Present, Simple Past, Present Perfect Simple and Continuous, Past Perfect and the Future, uniquely expressed by means of ***Will***.

A considerable presence of infinitives and passive forms must also be pointed out.

<u>Some more significant language features</u>

Some further linguistic features are equally of great importance within the Corpus.

- For example, variations of certain status and contrast with a prior situation are very common in the text. The most frequent item used to introduce them is the adverb **instead**, e.g. : ***Due to the requirement of stable orbits, the electrons therefore do not randomly occupy the whole region around the nucleus. Instead, they occupy various well-defined spherical regions.***

- Another widespread adversative is also **whereas,** seldom replaced by the equivalent *while*: ***Protons are positively charged particles, whereas neutrons are neutral particles, and both have about the same mass.***

- Explanations and new concepts are never left alone: they are further developed by means of expressions such as ***for example...; that is... ; which means that....; this is the... ,*** e.g. : ***On the energy diagram, (.....) means (.....) , which means that the equilibrium of two atoms corresponds to the potential energy of the system acquiring its minimum value.***

- Moreover, most concepts and statements require some visual references preceded by the structures: ***as depicted in figure .... , as shown in figure...; as is apparent in (this) figure; as indicated in figure....*** .

- A very limited use of personal pronouns is also easily traceable in the whole text. Instead, the repetition of the same noun at even very short distances in the text or within the same sentence results in a repetitive language. Here follows an example:

" Although there is a coulombic repulsion between the protons, all the protons and neutrons are held together in the nucleus by the strong force, which is a powerful, fundamental, natural force between particles. This force has a very short range of influence, typically less than .... . When the protons and neutrons are brought together very closely, the strong force overcomes the electrostatic repulsion between the protons and keeps the nucleus intact. The number of protons in the nucleus is the atomic number .... of the element ".

- As for the use of the relative pronouns, **which** is widely used instead of **that**, whose use is limited to no more than six/seven cases in all the corpus. This is possibly due to the fact that the text typology requires continuous demarcation and narrowing of the items referred to.

- Possessive phrases are often expressed using the preposition **of,** while there is no evidence of the Saxon genitive. A significant instance is the following sentence: ***The formation of the bond means that the energy of the system of two atoms together must be less than that of the two atoms separated.***

- A considerable use of prefixes is present in the text. We can thus find words such as, ***superfluid, quasiparticle, superconducting, pseudogap, antiferromagnetic, superconductivity, underdoped***.

- Finally, it must be said that the corpus follows quite a linear syntactic structure. This means that, on the whole, sentences tend to be short, plain and clear. A paratactic rather than hypotactic combination of phrases is actually prevailing.
  However, just in quite isolated cases, the reader is likely to find excessively long sentences such as:

For the magnetically underdoped systems, the exceptionally strong interaction between the hot quasiparticles leads to a change in their character; below a characteristic temperature, which corresponds roughly to that temperature at which the correlation length is equal to twice the lattice spacing, and the uniform spin susceptibility takes on its maximum value, there is a transfer of the quasiparticle spectral weight from low to high frequencies, as though a gap had opened up in the hot quasiparticle spectrum.

CONCLUSIONS

To sum up the leading features of this corpus, I would first say that its prose and structure do not drift away from general/standard English more than a daily newspaper article would. The very characteristic feature being the content, difficulties could only stem from the vocabulary used. As a matter of fact, the verbal patterns, forms and tenses, and the textual frame do not differ from those we might find in non specialized texts.

Comprehension is thus easier than expected, except when some technical words are used in order to express strictly scientific concepts, which could not be expressed otherwise. In this case a paraphrase is unavoidable so as to access the meaning easily.

On the whole, the corpus seems to form an organic text, whose different constituents are organized within a coherent and cohesive frame. The flow of the general topic is marked by specific consequential steps which are: introduction → explanation → cause and effect relationship → possible comparisons/juxtapositions/changes → final conclusion.

The corpus dealing with scientific topics, it can be said that the text belongs to the **expositive** and **descriptive** typology. Formal and rhetorical structures consistent with such text typologies can actually be found throughout the corpus. More specifically, in the first part of the corpus, it can be noticed that the author's attitude towards the text is neutral and objective, that is, contents are conveyed in an extremely impersonal style where the writer is practically inexistent. As a result, facts are described just as they are, without any " active " intrusion of the writer. This also accounts for the only personal and relative pronouns used being the neutral **It** or **They** and **Which** respectively**,** obviously referring to chemical items, reactions, experiments, bonding and the like.

Contrary to this, in the part of the corpus where two theories are described**, I, We** and even several proper nouns such as **Aage Bohr**, **Ben Mottelson**, **Mal Ruderman** and many others, appear. They all show a clear intent of the author to list the scientists who took part in the scientific discoveries at issue but, whenever strictly scientific facts are dealt with, impersonality and objectivity come out again thus reasserting the scientific typology of the corpus.

As already mentioned above, an extensive use of the passive in descriptions and explanations, further shows a need for objectivity and hedging as well.

Moreover, continuous exemplifications and clarification following a previous statement, as well as " referring back to " and " referring forward to " sentences, contribute to the descriptive and expositive nature of the corpus.

This said, a syllabus on ESP for Chemistry could not obviously do without the textual and linguistic features analyzed so far.

Serena Sotgia