BIO Web of Conferences **2**, 06001 (2014)
DOI: 10.1051/bioconf/20140206001

How Can Linguistics Help to Structure a Multidisciplinary Neo Domain such as Exobiology?

A. Condamines[1]

[1]CLLE-ERSS, UMR 5263, CNRS and University of Toulouse

Abstract. This chapter examines the contribution that corpus linguistics can make towards stabilising definitions within exobiology. By using the clues provided by natural language processing tools, linguists build an interpretation of contexts and try to show how the meaning is constructed, taking into account the different points of view of the disciplines involved in exobiology. Examples of contextualised interpretations are proposed.

1 Introduction

This chapter attempts to show how corpus linguistics may help to structure a new and interdisciplinary domain such as exobiology. Over the last few decades, linguistics has developed various methods to study text corpora as systematically as possible. It is now easy to identify the most specific words in a corpus and, by using contextual information, to better understand their meaning. However, exploring a corpus can sometimes lead to discovering that variations in meaning may occur for the same word in different parts of the corpus, thus invalidating the hypothesis of a stable meaning even within a scientific discipline. Many linguists therefore consider that meaning is not given once and for all but that it has to be constructed and that, if the aim is to define a concept, it is necessary to take into account all the variations of meaning appearing in a corpus, a point of view shared by most epistemologists concerning the creation of concepts. In a discipline such as Exobiology, which emerged from several disciplines, each with their own point of view, linguistics may help to identify these viewpoints and contribute to the definition of concepts. The following section presents the characteristics of exobiology from a linguistic point of view and lists the relevant elements of corpus linguistics for studying a multidisciplinary corpus. The third part shows how it is possible to construct and analyse an exobiology corpus in order to throw light on its semantic categories. The chapter closes with the study of examples of these categories.

2 Characteristics of exobiology and linguistics for specialized corpora

This section presents the epistemological context of the study. It focuses first on the characteristics of exobiology and then on the features of corpus linguistics that are relevant for the study of specialized texts. Taking these characteristics into account, it discusses how corpus linguistics can analyse a multidisciplinary corpus and spot differences and similarities between the four disciplines involved in exobiology.

Article available at http://www.bio-conferences.org or http://dx.doi.org/10.1051/bioconf/20140206001

2.1 Exobiology

From a linguistic point of view, exobiology presents two main characteristics. First, it is an emerging discipline and second, it is a multidisciplinary field. These two points are very important and they entail several consequences:

- This situation may generate complementary or conflicting viewpoints, concerning the lexicon in particular, as different disciplines may attribute either a similar or a different meaning to certain terms.

- The concepts of the discipline are still evolving. When experts from different fields agree to confront their points of view, it is clear that concepts will change under the influence of multidisciplinarity.

- Definitions are not stable. This is a consequence of the evolution of the concepts: meanings themselves fluctuate and, at least temporarily, definitions are destabilised.

This situation seems very close to the one described by Kuhn as a paradigm shift [1]. Nevertheless, there is perhaps a difference in that with exobiology, most of the experts from each domain involved in the construction of exobiology are aware that it is collectively necessary to throw light on the same object: extraterrestrial life.

From a linguistic point of view, the study of exobiology is extremely interesting because it is possible to analyse a discipline as it is being constructed. The issue of establishing definitions in such a situation enables linguists to question the role of definition both as a way to constitute referents for a discipline and also as a potential risk, that of limiting creation. As Rey pointed out [2], in the word *definition* as well as in the word *term,* there is the meaning of limitation (finite). When you confine a term, you stabilize its meaning but you also limit variations in use. So building definitions is possible only when a high level of consensus is reached among the different experts. As Fourez comments [3]: “Les pratiques interdisciplinaires peuvent être considérées comme des négociations entre des points de vue et des intérêts différents, dans un contexte et selon un projet”[a].

2.2 Corpus Linguistics

Atkins et al. give the following definition of a corpus: “a subset of [an Electronic Text Library] built according to explicit design criteria for a specific purpose […]” [4]. We will see that, in our case, the specific purpose will be the spotting of differences between words and terms used in texts from different disciplines.

It is generally agreed that four features are characteristic of corpus linguistics ([5], [6]):

a) Corpus linguistics necessarily requires the use of specific tools.

The corpora analyzed by linguists (even in specialized fields) may contain several hundreds of thousands of word forms, making it very difficult to study and memorize all the contexts of a word. Moreover, Natural Language Processing (NLP) tools propose very sophisticated methods that can enlighten us in original ways on linguistic phenomena (see below).

b) Corpus linguistics is doubly situated.

This point refers to the fact that in order to understand how a meaning may be apprehended, it is necessary to take into account both the situation in which the text was written (by whom - in the

[a] Interdisciplinary practices can be considered as the negotiation of different points of view and different interests in a particular context and with a specific purpose. (All translations from French are by the author)

present case, the discipline of the expert -, the date of publication, etc.) and the aim for which the text was written (in our case, to contribute to constructing a common point of view).

c) Corpus linguistics is often based on a comparative approach : comparison between different corpora or between several parts of a single corpus.

A good way to understand the meaning of words is to compare them and their functioning within several sub-corpora. The organization of a corpus into sub-corpora plays a very important role in such an approach because differences between meanings will be interpreted with regard to the point of view chosen for the corpus organization.

d) Corpus linguistics involves interpretation.

Starting from the premise that a meaning does not pre-exist but is constructed, and that NLP tools provide clues, the linguist's role is to propose an interpretation of linguistic phenomena, taking these clues into account (and, in some cases, invalidating them).

Generally speaking, corpus linguistics tools can provide three kinds of information, or clues.

The first one concerns the quantitative aspect of the lexicon. At the moment, this feature is the most fully developed one in NLP, which implements statistical methods in order to spot the most specific words or patterns in a corpus. This information is very often combined with the other two.

The second clue enables variations in forms to be compared. For example, it is possible to show that a given noun is systematically used with an adjective in one sub-corpus whereas this is not the case in another one.

The third type of information, also very widely used in NLP, is distributional. This was developed by Harris from a mathematical angle [7] and by Firth from a more sociological perspective [8]. As the latter said: *You shall know a word by the* company *it keeps*. From a semantic viewpoint, this has various consequences, among which the fact that the meaning of a word can only be grasped by considering the contexts in which it appears (or its "distribution") and the fact that two words will be synonyms if they appear in the same contexts. As we will see below, distributional information is very useful in specialized domains [9], [10], [11]. The main issue in such a distributional approach, however, is the problem of the similarity of contexts. One seldom finds two different words in exactly the same contexts. Many times, it is necessary to decide whether two contexts are similar or not, that is to say whether they may be categorized as belonging to the same semantic category or not. This process of categorization concerns the first stage of the interpretation.

Finally, a specific feature of corpus linguistics applied to specialized corpora is that linguists are not experts in the domain concerned by the corpus. In such kinds of study, therefore, we can speak of the "co-construction" of an interpretation as the result of two types of expertise: that of the domain experts and that of the language experts (linguists).

Like many philosophers of science [12], when adopting a corpus linguistics methodology, most linguists contend that meaning is not stable and given once and for all but rather that it is built ([13], [14]).

"The language of science demonstrates rather convincingly how language does not simply correspond to, reflect or describe human experience; rather, it interprets, or, as we prefer to say 'construes' it. A scientific theory is a linguistic construal of experience." [9].

"Whatever reality may mean, it always corresponds to an active intellectual construction. The description presented by science can no longer be disentangled from [scientists'] questioning activity" ([15])

This view is opposed to the traditional approach to terminology, which focuses mainly on standardisation:

"The new socio-cognitive theory of Terminology emphasises that Terminology should not be uniquely oriented towards standardisation and it questions the validity of objectivism as the theoretical underpinning of terminology". ([16])

We will see in the third part how corpus linguistics applied to specialized corpora can contribute to this process of construction.

To sum up this part, it can be said that when analysing an emerging discipline such as exobiology, one needs to adopt a constructivist point of view, shared both by some philosophers of science and corpus linguists.

3 Corpus linguistics on the EXOBIOLOGY corpus

This part describes how corpus linguistics methods may be applied to exobiology texts. By interpreting the information provided by NLP tools and comparing texts from the four main disciplines concerned by exobiology, it is possible to describe lexical and semantic phenomena and use them to spot both similarities and differences among the four disciplines [17].

3.1 The Exobiology corpus

The exobiology corpus was built by Nathalie Dehaut (PhD student) with the help of M. Gargaud (Laboratoire d'Astrophysique de Bordeaux). The main difficulty was that since the discipline is still under construction, there are no texts that really belong to exobiology. There are however some communicative situations which may be considered as representative of this construction. This is the case with the exobiology summer schools organised by the CNRS (Centre National de la Recherche Scientifique) where experts from the main disciplines constituting exobiology present to advanced students or even specialised colleagues from other disciplines the main issues of their original domain which are related to exobiology. This is a well known communicative situation (close to scientific popularisation) which is well adapted to linguistic exploration. The corpus contains the two books edited after these summer schools : *L'environnement de la Terre Primitive et l'Origine de la Vie (2001)* et *Les traces du vivant et l'origine de la Vie (2003)*. These two books are written in French. While this could be considered as a problem because the creation of concepts within exobiology concerns the international scientific community who writes mostly in English, it is easier for us who are French linguists to analyse texts in our language; moreover, it is well known that, concerning the organisation of concepts, there is no crucial difference within the same domain between different languages.

The corpus comprises 36 papers organised in four sub-corpora corresponding to the scientific origin of the 38 authors (see Table 1).

Table 1. Composition of the corpus.

Sub-corpus	Number of papers	Number of words
Astronomy	12	88,815
Biology	8	65,589
Chemistry	8	80,190
Geology	8	77,010
Total	36	311,604

3.2 Types of linguistic phenomena observed

The clues provided by corpus linguistic tools have been interpreted in order to characterize lexical, semantic and discursive phenomena identified within the corpus.

3.2.1 Quantitative results

Table 2 shows the 8 most frequent word forms in the whole corpus with their frequency per thousand words in each sub-corpus.

Table 2. The most frequent word forms in the whole corpus (per 1000 w.).

	Astro	Bio	Chem	Geo
Atmosphère	5.39	0.46	1.80	6.18
Eau	3.83	4.05	1.26	4.13
Temperature	2.57	2.67	1.13	4.43
Planète	5.26	0.61	0.37	3.24
Acide	0.45	1.83	6.42	0.18
Vie	2	2.65	1.73	2.19
Formation	0.72	0.77	2.43	2.74
Molécule	0.2	2.11	3	0.97

Three main comments may be made from this table. First, *acide* may be either a noun or an adjective. Secondly, except for *formation* which seems to be very general in meaning, the other seven words seem to be closely linked to exobiology. Finally, half of the eight most frequent words in the corpus come from geology. Astronomy and chemistry provide only one term each. These comments should be treated with precaution, however, because they concern only eight words. And, what is particularly important to note is that, within specialized corpora, although terms are nouns rather than verbs, they are mainly compound nouns.

When *acide* is used as a noun, its most frequent adjectival collocate is *aminé*. It is in the chemistry corpus that the range of possible adjectives is the most varied, with more than 30 different ones. When *acide* is used as an adjective, in three of the sub-corpora (geology, chemistry and biology) the noun that most frequently precedes it is *ph*. The other most frequent nouns are: in astronomy, *fonction* and *nuage*; in geology, *roche* and *ocean*; in chemistry, *milieu* and *environnement*; and in biology, *fonction* and *rivière*.

Two corpora present specific uses of *acide*: in astronomy, *adsorption des acides* (acid adsorption); and in chemistry, *ces acides racémisent moins facilement* (these acids racemize less easily).

3.2.2 Semantic results (using distributional clues)

Three kinds of dimensions have been identified and studied. The first two concern semantic phenomena strictly speaking:

- Synonymy: synonymy concerns cases in which one concept may correspond to two or more terms.

- Polysemy: polysemy corresponds to cases in which one term is associated to two or more concepts which are related by part of their meaning.

These two phenomena, synonymy and polysemy, may be observed either in the same domain or in two different domains. What is important to note is that, while they are often considered as critical problems when language is used only as a vehicle for providing information, these phenomena are very interesting to observe when a field is under construction. They are signs of the creative power of language and their fine-grained analysis is one of the best ways to spot creativity within the disciplines.

- Borrowing of a term ([18]): this generally concerns cases in which terms originate from another language. In our study, however, the borrowing concerns rather terms which originate from another discipline implicated in exobiology. In such cases, it may be difficult to ascertain whether the term retains its original meaning or whether it adopts a different meaning related to the new discipline. If we consider that a discipline is based on a system, that is to say a structured representation, it is likely that the use of such terms will be adapted to the new discipline, that is to say, to the new point of view. The meaning will therefore be modified and will become more general or more precise.

All these phenomena show that meaning may change, particularly when several communities (specifically, scientific communities) are in contact and decide to collaborate. As we have said, such phenomena are highlighted by analysing contexts in which a form (term) appears, that is to say, by categorising the distribution of a term. But, in some cases, the writers are aware of such difficulties and express them. This point leads to the definition of the second dimension.

The second dimension concerns the fact that speakers are aware or are unaware of semantic phenomena.

It is possible to identify whether experts are aware of certain semantic characteristics because there are linguistic patterns (so-called metalinguistic patterns) that may be used to signal this awareness (for example, “as said in...”, “it is not the same meaning in”, etc.). In all cases, linguists can collect and estimate these phenomena and submit them to experts in the disciplines concerned in order to initiate discussion about the concepts in question.

The third dimension concerns the fact that semantic phenomena may or may not lead to conflicts between experts. Language may be a “place of power” so the control of uses and definitions of terms may constitute an important issue for scientists. Some contexts show that the definition of certain concepts is fiercely defended.

Taking these three dimensions into account gives us twelve possibilities concerning the way terms work in a multidisciplinary domain. Table 3 summarizes these possibilities.

Table 3. The twelve categories of lexical phenomena in the corpus.

	Aware	Unaware	Controversial	Uncontroversial
Polysemy	x		x	
	x			x
		x	x	
		x		x
Synonymy	x		x	
	x			x
		x	x	
		x		x
Borrowing	x		x	

	x			x
		x	x	
		x		x

3.3 Examples

This section analyses some corpus examples and interprets them in the light of the above categorisation.

Polysemy probably without speaker awareness.
The three extracts below contains the term *prebiotic* (used as an adjective in all three cases).

(1) *Une atmosphère constituée d'azote moléculaire, de méthane et de vapeur d'eau constitue « la meilleure atmosphère prébiotique » i.e. le mélange gazeux le plus favorable à la synthèse des briques du vivant. (astronomie)*[b]

(2) *Une des difficultés de la reconstitution de l'atmosphère prébiotique (avant l'apparition de la vie) réside dans l'impossibilité actuelle de dater les débuts de la vie sur Terre. (astronomie)*[c]

(3) *La chimie prébiotique est une chimie organique en solution aqueuse, dans des conditions plausible de l'environnement primitif terrestre, conduisant à des composés d'intérêt biologique. (chimie)*[d]

A fine-grained analysis shows that *prebiotic* is not used with exactly the same meaning in the three examples. Note that the first two come from astronomy and the third from chemistry.
The three examples contain what are considered as definitional patterns, which are more or less complete. In (3), we encounter what is considered as an Aristotelian definition (in The Metaphysics): [Definiendum = definiens + specificities] (where the definiens is generally a generic term for the definiendum). In discourse, the copula = may expressed by *is a* [19]. So, from (3), we can presume that *organic chemistry* is a generic term for *prebiotic chemistry*.
In (2), there is also a definitional pattern (*before the beginning of life*) but which is less explicit because the copula is not as clearly expressed as in (3). What can help to understand that the parenthesis concerns *prebiotic* is the etymological elements *pre-* and *biotic*, from which we can infer the link with *before* and *life*.
In (1), two kinds of pattern draw the linguist's attention. First, of course, *i.e* which announces a kind of paraphrase, but also the quotation marks around *la meilleure atmosphère prébiotique*. These quotation marks may be understood in two ways. They probably mean that this term has been borrowed from another speaker (within or outside the discipline) but that it is not completely taken on board by the writer of the extract. But they also indicate that the noun phrase following *i.e* directly concerns the term in quotation marks and that they are equivalent.

Synonymy without speaker awareness.
The case of *exoplanète* (exoplanet) vs *planète extrasolaire* (extrasolar planet).

[b] An atmosphere composed of molecular nitrogen, methane and water vapour constitutes "the best prebiotic atmosphere", i.e. the most favourable mixture of gases for the synthesis of the building blocks of life.

[c] One of the difficulties in reconstituting the prebiotic atmosphere (before the appearance of life) resides in our current inability to date the beginnings of life on Earth.

[d] Prebiotic chemistry is organic chemistry in aqueous solution, in the plausible conditions of the primitive terrestrial environment, leading to compounds of biological interest.

First of all, as can be seen in Table 4, astronomy is the only discipline to use the two terms *exoplanète* (exoplanet) and *planète extrasolaire* (extrasolar planet).

We can assume, without the need for contextual information, that these two terms are synonyms, using an etymological clue: *exo-* and *extra-*, in Greek for the former and in Latin for the latter, mean "outside of". The term *planète extrasolaire* is more precise because it says that the planet in question is outside our solar system. Its equivalence is confirmed by the two examples below in which these terms appear in very similar contexts.

(4) *Signatures spectroscopiques de vie sur les exoplanètes*[e]**.** (astronomy***)***.
(5) *Chercher la vie sur les planètes extrasolaires par la détection de raies d'oxygène dans leur spectre*[f]*.(astronomy).*

Table 4. Distribution of *exoplanète* and *planète extra-solaire.*

	Astronomy	Geology	Chemistry	Biology
Exoplanète	19	0	3	0
Planète extrasolaire	12	1	0	3

Borrowing with speaker awareness.
Examples (6)-(8) below contain the adjective *inert.*

(6) *Les gaz rares des planètes (Ne, Ar, Kr, et Xe) sont chimiquement inertes.*[g] (astronomy)
(7) *Cependant, si cet appauvrissement est particulièrement marqué dans le cas des gaz rares, qui sont chimiquement inertes et donc peu retenus dans les silicates et le métal…*[h] (geology)
(8) *Une fois récupérés sous atmosphère inerte, ces produits sont soumis à différentes analyses afin de déterminer leurs propriétés optiques, leur solubilité dans différents solvants, leur structure moléculaire…*[i] (chemistry)

In examples (6) and (7), *inert* is preceded by *chemically.* We can surmise that *chemically* has to be interpreted as: from a chemical point of view. So, writers (both in astronomy and geology) are aware that when they speak about the "inertia" of noble gases, they are adopting a chemical point of view and they do not call this point into question (see the use of *donc* – therefore - in example (7)). In (8), written by a chemistry expert, the same adjective is not preceded by *chemically*.

These examples show how the analysis of contexts may be partly systematized. Such studies can be time-consuming and they have to be confirmed or discussed by domain experts. But they may be very useful in order to draw a picture of the situation at a given moment in a multidisciplinary field.

[e] Spectroscopic signatures of life on exoplanets.
[f] Searching for life on extrasolar planets via the detection of oxygen bands in their spectrum.
[g] The noble gases of the planets (Ne, Ar, Kr, and Xe) are chemically inert.
[h] However, while noble gases, which are chemically inert and therefore not captured by silicates and metal, are particularly impoverished…
[i] Once these products have been recovered in an inert atmosphere, they are subjected to various analyses in order to determine their optical properties, their solubility in different solvents, their molecular structure

4 Conclusion

This chapter has focused on the construction of exobiology. Linguistics may play a role in this construction by analysing texts from the four disciplines involved in exobiology. By interpreting clues provided by corpus linguistics tools, it is possible to show how the meaning emerges, which is a sign that the discipline itself is emerging. Twelve categories of linguistic phenomena have been identified. They take into account semantic phenomena such as polysemy, synonymy and borrowing but also the awareness (or lack of awareness) of the writers and their possible consequences in the relationship between terms in different disciplines (controversial or not). Real examples of the realization of these phenomena in the corpus of texts have been detailed. Some contexts, named linguistic patterns, can be directly interpreted as marking the awareness or the potential controversy. In most cases, however, it is necessary to analyse in detail what is called the distribution of a word, that is to say, all the contexts in which it appears.

These results are just propositions and they need to be discussed by exobiologists in order that they may build stable definitions, with a clear perception of their scope and adequacy.

References

1. T. Kuhn, *The Structure of Scientific Revolutions.* **University of Chicago Press** (1962)

2. A. Rey, *Essays on Terminology*, Amsterdam / Philadelphia, **John Benjamins Publishing Company** (1995)

3. G. Fourez, *La construction des sciences,* Bruxelles, De Boeck Université, (2002)

4. S. Atkins, J. Clear, N. Ostler, Corpus Design Criteria, Literary and Linguistic Computing, **7 (1)**, 1-16, (1992)

5. T. Mc Enery, A. Wilson. *Corpus Linguistics*. Edinburgh: **Edinburgh University Press**, (2004), (1st edition: 1996)

6. E. Tognini-Bonelli, *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins Publishing Company (2001)

7. Z. Harris, *A Theory of Language and Information: A Mathematical Approach*, Oxford University Press, (1991)

8. J. R. Firth, *Papers in Linguistics (1934-1951),* Oxford University Press, (1957)

9. M. A. K. Halliday, J.R. Martin, *Writing Science: Literacy and Discursive Power*. London, The Falmer Press (1993)

10. C. Gledhill, *Collocations in Science Writing,* Tübingen, Gunter Narr, 7-20, (2000)

11. J. Pearson, *Terms in Context*, Amsterdam and Philadelphia, **John Benjamins** (1998)

12. I. Stengers, J. Schlanger, *Les concepts scientifiques*, Paris, Gallimard, (1991)

13. G. Myers, *Writing Biology, texts in the Social Construction of Scientific Knowledge*. Madison, Wisconsin: **The University of Wisconsin Press** (1990)

14. F. Rastier, M. Cavazza, A. Abeillé, *Semantics for Descriptions*, Chicago, **Chicago University Press** (2002)

15. I. Prigogine, I. Stengers, *Order out of Chaos: Man's New Dialogue with Nature*. Toronto, New York, London, Sydney: **Bantam Books** (1984)

16. R. Temmerman, Questioning the univocity ideal. The difference between socio-cognitive Terminology and traditional Terminology, *Hermes* **18**, 51-90 (1997)

17. A. Condamines, N. Dehaut, Mise en œuvre des méthodes de la linguistique de corpus pour étudier les termes en situation d'innovation disciplinaire : le cas de l'exobiologie. *META*, **56(2)**, 266-283. (2011)

18. M. T. Cabré, *Terminology : theory, methods and applications*, Amsterdam and Philadelphia, **John Benjamins,** (1999)

19. A. Condamines, Corpus Analysis and Conceptual Relation Patterns, *Terminology*, 8-1 141-162 (2002)