Building a bilingual diachronic corpus of ecology: The long road to completion

Pascaline Dury
CRTT (Research Centre for Terminology and Translation, Lyon, France) and UHA (Université de Haute-Alsace, France)

Introduction: The need for diachronic corpora of scientific English

Over the past few years, there has been a tremendous growth in interest and activity in the area of corpus building and analysis. But most of these corpora are being compiled and made available for various periods of the history of general English[1] only. Corpora of scientific languages are still scarce, and, more precisely, we still need access to corpora covering long and short-term diachronic changes in specialized English. Terminologists, translators and LSP teachers are discovering the benefits of compiling and using computer-based corpora of authentic texts in order to learn more about the meaning and behaviour of terms. We have also witnessed an increasing need for diachronic information in the fields of translation and terminology. The diachronic dimension offers a better insight into a language by recording the emergence, development and demise of terms as they are used.

The diachronic dimension also gives valuable information on the major concepts of a specific field (and on how these concepts have evolved over time), and could be used as a tool to improve the definitions contained in databases accessible to terminologists and translators, for instance. Diachronic corpora, even when they cover a short period of time (i.e. the 20th century for the corpus of ecology described below), can provide examples of certain kinds of structures and term use or contexts characteristic of a certain period.

This paper presents the first steps of building a bilingual, diachronic corpus of ecology, and reviews the practical stages and difficulties in the process of compiling it. It also tries to pick out what the main benefits of such a corpus are for LSP[2] diachronic investigations.

1 The CIBLSP Project: General considerations

1.1 The diachronic corpus of ecology and its main objectives

This corpus focuses on showing the short-term diachronic changes (for the 20th century) in the field of ecology, and in two different languages, English and French. It can therefore be used to spot new terms entering the language of ecology over the last century, and to determine which have disappeared from it. It can also help to identify contexts for new meanings that have been assigned to existing terms, and to pick out prefixes and suffixes which have been added to existing words to create new terms. Last but not least, it can also give an overview of when very specialized terms are being used in less specialized contexts (a process called *de-terminologization*[3]). To meet this end, the corpus has been compiled by sampling both specialized and semi-specialized documents, as will be further explained in Section 2. Then, the corpus will hopefully provide the necessary data for carrying out follow-up studies to the PhD work already done[4] on the diachronic evolution of terms and concepts in the field of ecology. The link that exists between diachronic terminology and translation is indeed very interesting, and my belief is that diachronic knowledge on the evolution of terms and their meaning ensures that the translations produced are adequate as well as conceptually accurate.

1.2 The CIBLSP Project

The building of this bilingual, diachronic corpus in ecology is part of an overall project called *Corpus Informatisés Bilingues de Langues de Spécialités* (CIBLSP), a computer-based bilingual (English and French) LSP corpus. The project started at the University Lumière Lyon 2, at the Research Centre for Terminology and Translation (CRTT) in September 2002. It is carried out under the supervision of Professor Philippe Thoiron and is carried out by six researchers, each building a sub-corpus in a different specialized field of knowledge. Apart from the sub-corpus of ecology mentioned in this paper, the CIBLSP project also includes specialized bilingual sub-corpora in pharmacology, medicine, drugs and vulcanology. This project is built around common compiling criteria and a common methodology of work, i.e. in sampling the documents and in analysing them. Although each sub-corpus is designed according to different aims, the ultimate objective of the overall project is to give a better picture of terminological links across specialized fields[5], and to design better tools for investigating and teaching specialized English. This project being currently underway explains why CIBLSP is still an in-house corpus at the University Lyon 2.

2 The building of the sub-corpus of ecology

We know that there are no hard and fast rules that can be followed to determine the ideal compiling of a corpus, but the paragraphs below present the general criteria and design used in the building of my corpus of ecology, as well as the main decision points which were taken to do so.

2.1 General criteria and overall design of the corpus

2.1.1 Texts taken from the field of ecology

Texts pertaining to the field of ecology make up the bulk of the corpus. But I decided to narrow down the sampling of these documents to a restricted number of ecological subfields, for fear that encompassing the whole domain of ecology, which is very vast, would produce a high rate of various and thus not so conclusive results. Therefore, the collection of texts has been frozen to the following subjects: terrestrial ecosystems, ecological successions, niches, habitats and guilds, species communities and their interactions (especially predation and parasitism). Texts relating to aquatic ecosystems and the ecology of waters are not included, and the political aspect of environmentalism has been left out as well.

2.1.2 A comparable corpus

As mentioned in the first part, this corpus is a bilingual comparable[6] corpus. Such a corpus offers a double perspective on diachronic variations in the field of ecology, and enables us to draw comparisons between the two languages. For example, has the English language of ecology evolved at the same pace as the French language of ecology? Do these two languages show the same diachronic variations, or are there noticeable differences? Has one of these two languages borrowed more from the other?

2.1.3 Period chosen for the corpus

The period chosen for the building of the corpus covers the 20th century, in both languages. The oldest document included in the English corpus dates back to 1903, and I have sampled texts published from then until now. The delimitation of the period studied in the corpus has been governed by the scientific field chosen. Ecology, as we know it nowadays, really started with the founding work of the German zoologist Haeckel (1866) and his coinage of the German term *oekologie*[7]. Of course, we can find ecological or rather "proto-ecological" concepts in various works of the 17th and 18th century, like works by the naturalists Buffon and Linné for instance, but the constitution of ecology as an independent

domain (from biology, botany and zoology) only goes back to the end of the 19th century.

Ecology can thus be considered as a relatively recent field of knowledge compared to other traditional domains like medicine, chemistry or even biology, which have existed for centuries. Ecology, because it appears so comparatively "modern", therefore offers a real interest for studies in diachronic variations. Firstly, it largely remains an unexplored territory in terms of diachrony, scholarly interest in diachronic changes turning more often to older, more "established" fields like medicine, physics, or biology. Secondly, because most of the first "proper" ecological documents in English (that is to say using the term *oecology* in their title or at least in their content) were published around 1900 (often being translated from German books published at the end of the 19th century), it is also possible to spot the very first use of terms and then study the evolution of these terms and their meanings over the years. We can then observe which, among these newly created terms, were borrowed from other neighbouring domains and took a new meaning in ecology, and which were really coined by ecologists. Since the field of ecology has increasingly captured the interest of the general public over the years, it is also interesting to study the significant changes in meaning which appear in time when terms are used by non-specialists[8].

Once the overall design of the corpus had been mapped, the next task was to carefully identify the different statuses of that corpus and collect suitable documents for inclusion.

2.2 Principal decision points

2.2.1 Text status

All the documents selected are written, fully published texts[9] taken from magazines and books. All the texts compiled are specialized or semi-specialized articles and book chapters, thus enabling me to study the first steps of the process of de-terminologization, as specialized terms migrate to the semi-specialized language before being taken up in general language. I therefore consider that semi-specialized texts represent a middle-ground, or a "transition zone" between expert communication and the language used by laymen. Of course, I will then consider expanding the corpus over time to the inclusion of unspecialized documents in order to observe the process of de-terminologization to its end-point. I also decided to sample a large range of various articles and books written by many different authors, in order to neutralize as far as possible the effects of sampling bias and the stylistic idiosyncracies (which do also occur in LSP) of one particular author.

2.2.2 Language status

The two languages used in the corpus are French and English. As far as the English corpus is concerned, I tried essentially to collect articles and books which had been written by native speakers, but in some cases (when an article discusses an important ecological concept or uses an interesting terminology), I also included articles written by non-native speakers, but only if these articles had been reviewed by an editorial board first. I did not keep translated texts, except in two cases: the books by Schimper (1903), and Warming (1909), are translated from German and Danish respectively, because these documents are cornerstone works in ecology and therefore could not be put aside. The same sampling criteria will prevail for the compiling of the French corpus.

2.2.3 Period division

I organized the building of the corpus around ten sub-periods of ten years each, the first sub-period, which is 1900–1910, the second 1910–1920 etc. up to the last sub-period, 1990–2002. In order to achieve as much balance and diachronic "representativeness" as possible, each sub-period contains approximatively 75,000 words, one third (around 25,000 words) being collected in books, one-third in semi-specialized documents (mainly articles), and the last third stemming from specialized texts (also mainly articles). Table 1 shows the exact word count for each sub-period. This corpus has been devised as an open-ended corpus and I have structured it in a way (by selecting sub-periods of ten years) that makes improvement and supplementation easy and uncomplicated. Revised corpus versions should not, of course, be introduced every year. Five-year intervals might be appropriate and realistic as far as diachronic changes are concerned.

Table 1: Diachronic corpus of ecology: word count for each sub-period

Subperiods	**Type of document**	**Number of words/ document**
1990–2000…	Specialized	29,064 (5 articles)
	Semi-specialized	25,342 (9 articles)
	Books	28,728 (3 books, 7 chapters)
Total		**83,134**

Subperiods	Type of document	Number of words/ document
1980–1990	Specialized	24,532 (4 articles)
	Semi-specialized	26,256 (8 articles)
	Books	28,031 (1 book, 6 chapters)
Total		**78,719**

1970–1980	Specialized	24,820 (4 articles)
	Semi-specialized	24,871 (6 articles)
	Books	29,123 (2 books, 4 chapters)
Total		**78,814**

1960–1970	Specialized	25,679 (2 articles)
	Semi-specialized	26,270 (13 articles)
	Books	26,852 (3 books, 6 chapters)
Total		**78,801**

1950–1960	Specialized	26,148 (7articles)
	Semi-specialized	25,780 (16 articles)
	Books	24,898 (1 book, 6 chapters)
Total		**76,826**

1940–1950	Specialized	26,787 (5 articles)
	Semi-specialized	12,806 (7 articles)
	Books	29,440 (1 book, 4 chapters)
Total		**69,033**

1930–1940	Specialized	29,416 (3 articles)
	Semi-pecialized	18,686 (11 articles)
	Books	24,820 (4 books, 7 chapters)
Total		**72,922**

Subperiods	Type of document	Number of words/ document
1920–1930	Specialized	30,459 (5 articles)
	Semi-specialized	1,573 (1 article)
	Books	26,851 (2 books, 5 chapters)
Total		**58,883**

1910–1920	Specialized	25,360 (8 articles)
	Semi-specialized	0
	Books	15,682 (2 books, 7 chapters)
Total		**41,042**

1900–1910	Specialized	0
	Semi-specialized	1,316 (2 articles)
	Books	39,500 (2 books, 12 chapters)
Total		**40,816**

Total word count		**678,990**

2.2.4 Length status

The total word count is at present a little under 700,000 words for the English corpus, and as I hope to reach the same approximate number in French, the final corpus will be around 1.5 million words.

3 Difficulties and first considerations

3.1 Difficulties

3.1.1 Availability of old scientific material

One of the most obvious and first difficulties in building a diachronic corpus is linked to the availability of old scientific material. It seems that, though most universities are convinced that it is useful to keep old material written in the general language, they do not always see the point of storing old scientific docu-

ments full of obsolete theories and concepts that no scientist uses any more (and also because many of them lack the adequate storage facilities and space to do so). Therefore, I had to select the appropriate university offering extensive storage of old ecological material in the English language, and which was ready to cooperate. The good match was found with the University of Ulster, Northern Ireland, and my English corpus was essentially built making extensive use of its archives.

Another problem is linked to the absence of specialized articles for the first years of the corpus. The first specialized periodicals in English properly dedicated to the domain of ecology appeared after 1910. The first volume of *The Journal of Ecology*, the oldest available periodical in ecology, appeared in 1913. It was then followed by the publication of the first volume of *Ecology* in 1920, thus making it impossible for me to collect any specialized data taken from articles for the sub-period 1900–1910.

3.1.2 The scanning of the material

As mentioned by all linguists who embark on a corpus-based activity, corpus compiling is time and energy consuming. But as far as data capture is concerned, diachronic corpora are even more challenging to build. Indeed, to be properly analysed, the selected material of a corpus has to be converted into machine-readable form. Electronic documents are now increasingly available in nearly every scientific domain, and in this respect the Web proves to be a very helpful instrument, which highly facilitates data sampling. Even if this holds true for recent material, it is very difficult to find old scientific documents already converted in electronic form, and they are simply non-existent in the domain of ecology. This means that most of the material selected for the corpus had to be scanned using Optical Character Recognition software (OCR)[10] before being converted into Ascii character files. But for the capture of old and sometimes degraded printed material, keyboarding, though labour-intensive, was the only solution. However, as OCR software is not foolproof, I had to carefully proofread the scanned material. It seems that despite the best intentions, it is very difficult to spot all the mistakes when having to go through large amounts of text requiring this type of mechanical accuracy. As I found several errors after my first round of proofreading, I had to embark on a second-round of proofreading in order to polish up the files.

3.1.3 Copyright settlements

Another pragmatic constraint is related to copyright settlements. Though the old material included in the corpus is not subjected to copyright anymore, I still have to tackle the copyright issue for the more recent documents. This is not always a straightforward process, since copyright holders do not always understand what a corpus is and how it is used, and fear that granting the copyrights for a book or a magazine somehow means that less people will buy it.

3.2 What is left to be done

The second proofreading of the English corpus is completed and I am now contacting the publishers about copyright settlements. This corpus still needs to be edited, a list of lexical items needs to be extracted from it and validated by experts, and a linguistic analysis of the diachronic variations appearing in the corpus still needs to be produced. Then, the whole process of compilation, validation and analysis will have to be carried out again for the French corpus. Last but not least, once all the sub-corpora of the CIBLSP project are completed and fully analysed, we will be able to draw comparisons between them, and study the terminological links which appear between the domains involved.

4 Concluding remarks: The long road to completion

The completion of the corpus may appear as a slow and fastidious process to the reader, and I am aware that, though I have been working on its compilation for over a year, and have already reached a word count of 700,000 words for the English corpus, I am just about to start to analyse it. But I am also aware that the initial stages of designing the corpus carefully, and making the right sampling decisions are essential to ensure a final analysis and results of a high-quality level. Because the corpus is intended as a bilingual, comparable corpus, I also have to make sure in the initial stages of its design that the documents sampled in French and in English are relevant, adequate and "comparable" enough, in order to ensure that the similarities as well as the distinctive features of the two languages as far as ecology is concerned appear clearly to us in the end. We are therefore convinced that nothing good can come out of a corpus which has been hastily compiled, and which does not follow the golden rule of "choice, but not chance".

As I mentioned earlier in the article, this corpus of ecology can prove to be a valuable tool to observe the migration of terms and concepts from specialized communication to the general language over time, and I will complete it to meet this aim in due course. I therefore keep it in mind that a diachronic corpus must

stay open-ended and must be completed and improved regularly with new documents. In so doing, I adopt the method of successive approximations (as described by Atkins, Clear and Ostler 1992) which involves identifying the strength and weaknesses of a corpus and in the light of this, enhancing it by deleting or adding adequate new material. The road to completion may indeed be quite long before reaching the final point.

Acknowledgments

I should like to acknowledge gratefully the help I have received from the University of Ulster for the compilation of the English corpus; help from Lynette Logan and Jonathan Hyndman at the campus library of Coleraine; and help from Alicia Black and Jim Fitzsimons from the campus library of Jordanstown. I would also like to thank Oliver Hetherington, from the School of the Built Environment, who accepted to host the project, and who made sure I worked in optimal conditions during my four month stay in Belfast.

Notes

1. For instance the ARCHER corpus (cf. Biber *et al.* 1994), which covers the period 1650–1990 for the general English language, and the CONCE corpus (cf. Kytö *et al.* 2000), covering the 19th century general English, but also the Helsinki corpus, covering various registers for the years 850–1710 (ICAME, The Norwegian Computing Centre for the Humanities).
2. LSP = Language for Special Purposes.
3. See Meyer and Mackintosh (2000: 111) . "…when a term transcends the boundaries of expert language and starts to be used by the general public – a process we call de-terminologization."
4. See Dury (1999) and (2003).
5. See Arlin, Depierre, Dury, Josselin, Lervad and Rougemont (2004). For instance, the team will hopefully be able to show that specialized fields of knowledge are not hermetically closed to each other. On the contrary, terms and concepts can be seen as "mobile entities" which can be borrowed and used in different fields, thus proving that inter-domain lexical and conceptual sharing exists.
6. Comparable corpora consist of sets of texts in different languages that are not translations of each other. I use the word "comparable" to indicate that the texts in the different languages have been selected because they have some characteristics or features in common.

7. Ernst Haeckel (1866 : 286) created the term oekologie on the Greek formant oikos: "By oekologie we mean the body of knowledge concerning the economy of nature – the investigation of the total relations of the animal both to its organic and its inorganic environment".
8. For example, terms like ecosystem, biosphere, greenhouse effect are now widely used in general language because the reality they designate for experts is also of great interest to the general population.
9. That is to say texts which were printed in multiple copies for distribution and are copyright registered. All the articles included in the corpus are unabridged. Only books are not included in full; only the relevant chapters corresponding to my initial choice of subfields have been kept. Including full books would have also challenged the overall balance of word counts and the distribution of words between specialized and semi-specialized documents, and books.
10. I used HP Precision scan Pro 2.0 software to scan the corpus material.

References

Arlin, Nathalie, Amélie Depierre, Pascaline Dury, Amélie Josselin, Suzanne Lervad and Claire Rougemont. 2004. Projet Ciblsp, corpus informatisés bilingues de langues de spécialité. *Aisl* 1: 1–12.

Atkins, Sue, Jeremy Clear and Nicholas Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7(1): 1–16.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 7(4): 243–257.

Biber, Douglas, Edward Finegan and Dwight Atkinson. 1994. ARCHER and its challenges: Compiling and exploring a representative corpus of historical English registers. In U. Freis, G. Tottie and P. Schneider (eds.). *Creating and using English language corpora*, 1–13. Amsterdam-Atlanta, GA: Rodopi.

Bowker, Lynne and Jennifer Pearson. 2002. *Working with specialized language. A practical guide to using corpora*. London and NewYork: Routledge.

Bramwell, Anna. 1989. *Ecology in the 20th century: A history.* New Haven: Yale University Press.

Dury, Pascaline. 1999. Etude comparative et diachronique des concepts ECOSYSTEM et ECOSYSTEME. *Meta* 44 (3): 484–500.

Dury, Pascaline. 2003. *Etude comparative et diachronique de dix dénominations fondamentales du domaine de l'écologie en anglais et en français*. Villeneuve d'Ascq: Presses Universitaires du Septentrion.

Haeckel, Ernst. 1866. *Generelle Morphologie der Organismen: Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte Descendenz-Theorie.* Two volumes. Berlin: G. Reimer.

Kytö, Merja, Juhani Rudanko and Erik Smitterberg. 2000. Building a bridge between the present and the past: A corpus of 19th-century English. *ICAME Journal* 24: 85–97.

Meyer, Ingrid and Kristen Mackintosh. 2000. When terms move into our everyday lives: An overview of de-terminologization. *Terminology* 6 (1): 111–118.

Rissanen, Matti. 1989. Three problems connected with the use of diachronic corpora. *ICAME Journal* 13: 16–19.