Corpora and concordancing: Practical Issues

ıntroduction

This article is based on a workshop given at the AFLS and SFS joint Atelier *Les français des corpus* held at the University of the West of England on 16 February 2002. A practical hands-on session does not convert easily into a written paper. The concordancer used was one that I had written for use by myself and my students[1], but the general principles remain valid whatever the software. What I have attempted to do in this article is retain the very practical focus of the session which addressed the topic of corpus construction as well as exploitation[2]. Now as then my target audience is primarily those who are new to concordancing.

1. Finding or building a Corpus

Before you can think about concordancing, you must first obtain a corpus. To be more than just an archive of texts a corpus should be organised according to a set of principles. The first thing you must do is ask yourself what you want your corpus for. Is your focus going to be on written or oral French? If the latter, then there are a few admittedly fairly small corpora available on the web[3]. Individually or combined these may up a reasonable starting point for a concordance based analysis of spoken French, provided they match your criteria regarding speakers (age, sex, etc) or text type (conversation, narrative, formal, etc). You can of course construct your own corpus. This will require you to go out and make some field recordings which you will then have to transcribe and edit according to your purposes. If you intend to perform purely textual analysis, then the task is relatively (!) straightforward even if time-consuming. If you intend to include prosodic features, then you are about to embark on a major enterprise for which you will probably need collaborators.

[1] KWIC-CONCORD - available free on an “as is” basis from the author who can be contacted at jeremy.whistle@northampton.ac.uk.

[2] For a discussion of pedagogic issues, see Whistle, “Concordancing with students using an ‘off-the-Web’corpus” on-line a*t* http://www.hull.ac.uk/cti/eurocall/recall/rvol11no2.pdf. For a commentary see http://www.ict4lt.org./en/en_mod2-4.htm#_Toc481294210

[3] see http://www.uwe.ac.uk/facults/les/staff/kb/corpus.pdf,
http://www.univ-ubs.fr/valoria/antoine/parole_publique,
http://ml.hss.cmu.edu:90/ola/french/

Jeremy Whistle

If, like me, you are interested in written texts, then the task is simpler although at times tedious. If you have a scanner and an OCR[4] package you can scan and convert any number of texts. This gives you much greater control over your source materials but is time-consuming and will require some editing as OCR software is not infallible. Much will depend on the size and clarity of the font and the presence of italics. It is far easier to download from the Web. You will still have to decide what your focus is to be. If you are interested in literature, then, whether your interest is synchronic or diachronic, there are texts a-plenty to be had[5]. If your purpose is a detailed study of one text, for example *Madame Bovary*, or a contrastive study of, say, two texts, for example *Adolphe* and *Le colonel Chabert*, then the problem of the corpus will be quickly resolved. One important limitation is that, for obvious copyright reasons, you will not find any modern texts and so may not be able to derive any valid judgements about modern literary French. If your interest is in journalistic French, then there may not be any ready made corpora but there is a multitude of sites from which you can download articles[6]. If you are interested in technical or business French, you can enter keywords in your search engine and track down appropriate sites[7]. If you are interested in constructing a parallel corpus, you can utilise a site such as that of the French Foreign Ministry[8] which publishes in several languages or a periodical such as *Courrier International*[9] which contains many translations of articles taken from British sources such as the *Guardian*[10] and *New Scientist*[11]. There is increasing interest in learner French[12] and to construct a corpus all you need to do is require your students to hand in all coursework on a floppy. You might feel ethically bound to warn them that their work may be used for research purposes, all the more so if you intend to publish sizable extracts.

[4] Optical Character Recognition. The two programs commonly available are OmniPage Pro and TextBridge. These are often bundled with scanners but in a ‘lite’ form that does not always handle accented characters very well.

[5] see for example http://abu.cnam.fr/ and http://un2sg4.unige.ch/athena/html/fran_fr.html, http://gallica.bnf.fr/

[6] see http://permanent.nouvelobs.com/, http://www.monde-diplomatique.fr/, http://www.lexpansion.com/pages/default.asp?pid=4, http://www.regards.fr/archives/

[7] for an example of business French go to http://www.thebodyshop.ca/home.asp?Lang=FR&CName=Home

[8] see http://www.france.diplomatie.gouv.fr/

[9] see http://www.courrierinternational.com/actual/accueil.asp

[10] see http://www.guardian.co.uk/

[11] see http://www.newscientist.com/

[12] Two of the atelier contributions, by Annie Lewis and Sylviane Granger, were on the topic of learner corpora. For more information on the latter and the FRIDA project, see http://www.fltr.ucl.be/fltr/germ/etan/cecl/Cecl-Projects/Frida/gateway.htm

My own corpus was constructed to support my language teaching. The bulk of this takes place within a combined or joint honours degree. My students therefore learn French for academic purposes. They are required to read texts in a formal register on a wide variety of subjects. They are also required to write French that is accurate within accepted norms (i.e. acceptable to examiners or potential employers). To meet these objectives, I settled on the French Foreign Ministry site and especially its two publications, *France*[13] and *Label France*[14]. The former is essentially informative and text book in style. The latter is more varied and, though generally informative, does contain interviews and occasionally more contentious material. *Label France* has two great advantages, its focus, which is essentially on France, and the variety of topics. The result is quite a rich vocabulary covering the whole of what used to be called *civilisation française*. Another incidental advantage is that *Label* is published every quarter and so the corpus is continually growing. As at July 2002 it amounted to 663,613 words or tokens and 40,318 types (non-lemmatised). If the *France* files are added, the total is just over 750,000 words.

Once one has identified a possible source corpus and resolved any copyright issues, there are still a few practical difficulties. Which format is your corpus in at source? Which format do you intend to store it in ? One unavoidable problem is the accented characters, those letters that are not to be found on American keyboards and which were missing from the original 7-bit ASCII (or MS-DOS) character set. Certain characters were added in later versions but these still lacked most upper case accented letters and 'oe ligature'. You may still find an old corpus of ASCII encoded texts lurking in some obscure corner of your institution's computer. The problem was largely resolved with the appearance of the ANSI[15] (or Windows) code set although it is only relatively recently that codes were allocated for 'oe ligature'. ANSI has now, at least in theory, been overtaken by UNICODE which can encode up to 65,000+ characters, as opposed to 256 in ASCII and ANSII. In parallel the HTML set has been evolving, at first aligning itself on ANSII but increasingly, it appears, on UNICODE. If we take the case of 'oe ligature', this was missing in ASCII and early versions of ANSII but later

[13] see http://www.france.diplomatie.gouv.fr/venir/voicilafrance/fr/index.html

[14] see http://www.france.diplomatie.gouv.fr/label_france/

[15] many English language texts use the term ASCII indiscriminately to refer to both the MS-DOS (ASCII) and Windows (ANSI) code sets because most English language applications do not make use of the extra characters in the latter

appeared unacknowledged as code 156 (Alt+0156 on normal Windows keyboards). If this seems confusing, the situation was worse with HTML. As with ANSI, in the first versions 'oe ligature' was missing, although some texts used the standard ANSI codes instead, then the set was extended to include &oelig; and &#339;, by which time the alternative &#156; had come into widespread use. The result is that HTML now has four possible codes for a single character. A search for a word like *cœur* on an ASCII encoded corpus will require the search item to be entered as 'coeur' whereas an ANSI encoded corpus will require two items, 'coeur' and 'cœur', to be entered and an HTML encoded corpus will require 'coeur', 'cœur', 'c&oelig;ur', 'c&#156;ur' and 'c&#339;ur' to be entered, or at least allowed for by the search engine. It is slightly better for the other characters insofar as there is only one form in ASCII and ANSI and no more than three in HTML.

The advantage of using an HTML corpus is that texts can be used in their 'raw' form as downloaded from the Web. Many of the major corpus building initiatives are based on the TEI[16] guidelines which are essentially a variant of HTML (SGML or XML). All the characters are within the original 7-bit set and so there should be no problems with transmission and reading to screen of texts, providing you are using up-to-date software. If you are constructing or working on a tagged[17] corpus, this will almost certainly require using one of the HTML family coding schemes. The disadvantage of HTML is the sheer mass of code needed to store information about the text as well as encode characters and formatting information. All of this information has to be filtered out, firstly in order to enable the search and then, once a hit has been scored, to present enough text to contextualise the search item. As a rough average, about half of an HTML text is made up of codes of one sort or another.

For quick searches which do not require all the ancillary information that a scheme such as TEI provides (the information can in fact be stored in parallel files), it may well be preferable to convert the text files from HTML to ANSI using a browser such as Internet Explorer[18]. Once you have downloaded the files,

[16] see http://www.tei-c.org/

[17] A tagged corpus is one in which every word has a label or tag attached indicating which part of speech the word is as well as related grammatical information (number, tense, person, gender, etc). A parsed corpus goes further providing syntactic information as well. The latter can only be read and analysed with specialist software.

[18] To cope with the latest codes for oe ligature you will need version 6. You can also read in an HTML document into MS-Word and then save it as a text document. The latest version of KWIC-CONCORD

all you have to do is double-click on each file from within Windows Explorer (the file manager, not the browser) and, provided Internet Explorer is the default reader for HTML files, your file will open in an Explorer window and all you have to do is click on ‘File’ in the top left-hand corner, select first ‘Save as’ and then the ‘text’ option and your file will be converted and saved. In corpus linguistics, as in religion, things are not quite that simple. Your new text file will need a certain cleaning up before it is ready for use[19]. You may well have a few relicts of hypertext codes such as ‘retour’ and ‘sommaire’ remaining and a few missing upper-case letters at the beginning of paragraphs where a drop capital in the source text has been represented by a graphical image rather than a text character. If you do this every time you download a few files, life is much easier. Doing it on several hundred is decidedly tedious.

2. Analysing the Corpus

Once you have a corpus, the real fun begins. You can perform word frequency counts. These will provide quantitative information about lexis. They can be specially useful in course design. They will tell you what are the most frequent words in the language as a whole or in the specific text or texts that your course is based on[20]. If you focus on specific categories such as adjectives, a word count will provide qualitative information as well. However, there are limitations. One obvious one is elision. *Le* and *la* will appear separately and as *l*. The hyphen is also a problem. If you ignore it, *grand-mère* will appear as two separate terms, *grand* and *mère*, as will *peut-être*. If you include it, all the inverted forms such as *est-elle* and *a-t-il* will be returned as a single form. You could pre-edit all the texts to resolve the problems. Alternatively, you can simply perform two counts, one recognising the hyphen as a word break and one not, and combine the two. The major and as yet intractable problem is lemmatisation[21] of the results. This is more acute in French, given for example the number of separate forms a verb or adjective can take compared with English. However, it will be almost impossible to lemmatise the results unless you are working on a tagged corpus, and as yet

produces multiple text files as a by-product of a wordcount on HTML files while WordSmith Tools allows you to create a conversion file – to do this you will need to know all the codes for HTML and ANSI and something about the way the file is organised.

[19] you can use Notepad or Wordpad to do this or Word if you remember to select ‘text’ in the ‘save as type’ option

[20] For an example applied to Italian see Loredana Polezzi, ‘Concordancing and the Teaching of *ab initio* Italian Language for Specific Purposes’, *ReCALL*, , 1993, vol.9, pp.14-19

[21] lemmatisation is the process of grouping all inflected forms of a word under the head word, as in a dictionary (plurals with singulars, feminines with masculines, tenses with infinitives)

none of these exists in a completely satisfactory form although there are projects under way[22]. Without parts of speech (POS) tags, it is impossible to distinguish between *le* article and *le* direct object pronoun, *porte* noun and *porte* verb, *fait* verb (third person singular or past participle?) and *fait* noun. To find (or rather to add) this information, you will need to perform a concordance search on each and from the context decide which category the word belongs to. This is feasible in the case of *porte* (83 occurrences in my *Label* corpus), less so in the case of *le* (14,405 occurrences – 67,980 if one includes elided, feminine and plural forms).

The primary function of a concordancer is to perform concordance searches. With the latter you begin to see language as it is used. The concordances will provide you with the basic data on which you can formulate and prove, or disprove, hypotheses. Perhaps the easiest questions to answer relate to lexis and especially the comparative frequency of related items and often a concordance search is much quicker than a full word-count. A search on *marketing* and *mercatique* produced nineteen examples of the former and one of the latter[23]. The figure for *mercatique* has remained stable while that for marketing has doubled as the corpus has doubled in size. As always when quoting figures one must take care. At what point does one decide that the frequency information is valid? Can one simply state that, for example, *marketing* is nineteen times more frequent than *mercatique* ? Only if one adds the rider ; "in the test corpus…" and provides information about its size and the texts that it comprises. Alternatively, one could state that "in the test corpus *mercatique* has a frequency of 1 per 663,613 words whereas *marketing* has a frequency of 19 per 663,613[24] words. It only requires the addition of one more file of, say, 300 words, one of which is *mercatique*, for the first figure to double[25].

[22] for example the Winbrill project (http://jupiter.inalf.cnrs.fr/cgi-bin/mep.exe?HTML=mep_winbrill.txt) and the CHILDES project (http://cnts.uia.ac.be/childes/) whose CLAN software incorporates a tagging program - neither is easy to use

[23] all the figures given from this point on, unless otherwise stated, are based on a corpus of 527 files taken from *Label France*, issues 19 to 46

[24] or that *mercatique* has a global frequency of 0.0002% or 2 per million words and *marketing* a frequency of 0.0029% or 29 per million words.

[25] This is one good reason for large corpora. One disadvantage of a corpus based on a quarterly publication like *Label France* is that issues tend to be thematic. One issue can radically affect the frequency of a given word. Predictably, the three issues (nos. 29, 31 and 33) dedicated to the world cup in 1998 generated 72% of the occurrences of *football* (92 out of 127). A single issue on the position of women in France (no.37) generated 367 occurrences of the word *femme* out of a total of 855 (43%) while another issue dedicated to the Universal Declaration of the Rights of Man (no.34) generated 418 occurrences of the word *droit* out of a total of 846 (49%).

Although this information could be extracted from a frequency list, this would not work with a compound form such as *bien que*. The list will tell you that there are 697 occurrences of *bien* and 6,685 of *que* and *qu'* combined but it cannot tell you how many of these are examples of the conjunction *bien que*. A search on *bien que* produced fifty examples of the first, compared with two examples of *quoique* and none of *malgré que* or *encore que*. A second level of analysis could then be performed on the constructions that followed the conjunctions: adjective alone (*bien que favorable*), present or past participle (*bien qu'étant, bien qu'enrichie..*) or full tense (*bien qu'elle soit différente*). A frequency list will not throw up collocations or mutual information. The ability to extract collocations can be particularly useful when rejecting forms found in students' work. A good example is **immigration illégale*, as opposed to *immigration clandestine* or **payer hommage* as opposed to *rendre hommage*. This can be especially useful for non native-speaker teachers when marking students' work since their intuitions can otherwise be more easily challenged or rejected by students.

Lexico-semantic information requires a full concordance. One productive area is to highlight English loan-words whose meaning is more restricted in French: *marketing*, as opposed to *commercialisation*, or *management*, as opposed to *gestion*. They can also help clarify the mutually exclusive uses of pairs that correspond to a single term in English: *haut* and *élevé* corresponding to 'high' or *savoir* and *connaître* corresponding to 'know'. Concordances can be used to highlight lexico-grammatical differences between English and French such as permissible constructions after verbs such as *augmenter* and *diminuer* or *résulter*[26]. Among the grammatical questions that can be addressed is that of auxiliary use with verbs such as *paraître* and *apparaître*. A concordance search will furnish convincing evidence that the latter should now be included with the small group of verbs conjugated with *être* while the former is increasingly found with *être* when used in the sense of 'to be published'(see Appendix 1). One can also use concordances to test rules found in grammars[27] such as the statement that *espérer que* is always followed by the indicative. A search produced fourteen occurrences (see Appendix 2), two of which clearly contained a subjunctive:

[26] obviously, the corpus will not of itself throw up equivalents for 'to result in'

[27] for a discussion of the relative frequencies of *si.. et que..* (as opposed to *si..et si*) and the tenses used in the second term, and the degree to which different corpora can give different results, see Raphael Salkie's article: http://www.unl.ac.uk/sals/afls/resguide.htm#Corpus

1. il faut incontestablement espérer que se **mette** en place
   une confédération mondiale

2. on peut espérer que la féminisation du monde du travail **soit**
   synonyme d'une meilleure efficacité et d'une plus grande humanité.

What the concordance cannot do is formulate a rule. My provisional rule, based upon examples seen or heard elsewhere, was that the subjunctive was found when the implication of *espérer* was closer to 'wish' than 'hope'. The second example quoted would seem to challenge this. Another rule, also involving the subjunctive, is that of the mood to be used after *après que*. We all know that the subjunctive is widely used yet grammars still insist on the indicative. What advice should one give one's students? A search on the *Label* corpus provides no examples of the subjunctive, unlike two alternative corpora (Appendix 3). The results do however suggest that we may have to give more place to the past anterior in our teaching.

The concordancer will also allow for stylistic analysis of literary texts. A wordcount performed on two early 19th century novels, *Adolphe* and *Le colonel Chabert*, reveals that the former contains 4,874 types for 31,239 tokens (a ratio of 1:64) while the latter has 4,748 types for 23,923 tokens (a ratio of 1:5). While the two are not exactly comparable in length, it is clear that the latter has a richer vocabulary. One might expect this considering the greater importance of the physical world, and even of the grotesque, for Balzac. It will not surprise those who know the novels that, in *Adolphe, je* is second in frequency, surplanting *le*, with *j'* in 17th position, adding up to 3.7837% of all types whereas, in *Chabert*, it comes in in 11th position, with *j* in 40th position, together making up a mere 1.6135% of types. Given the importance of description in Balzac, one might expect to find more examples of *était* in *Chabert* (45th position) but in fact *Adolphe* (33rd position) contains 30% more examples, once adjustment is made for length. One might expect to find more adjectives in the top 350 words of *Chabert* but all one finds are *vieux* (77th in list), *pauvre* (105th), *malheureux* (226th), *cher* (236th), *haute* (284th), *belle* (311th), *heureux* (331st) and *riche* (347th). *Heureux* and *heureuse* appear in *Adolphe* (in 292nd and 350th positions respectively, as does *malheureux* (269th). More significant are *libre* (250th) , *triste* (307th) and *bizarre* (313rd). Further searching becomes time-consuming but would be much quicker with a tagged corpus. As far as nouns are concerned, the two most striking are *cœur* (64th in *Adolphe* and 190th in *Chabert*) and *amour* (67th in *Chabert* and 308th in *Adolphe*). On the other hand you will search in vain in *Adolphe* for *francs* and *argent* (142nd and 152nd respectively in *Chabert*), both of which link thematically with *pauvre* and *riche*, but the former does have *fortune* in 201st position (118th in the latter). None of these findings is earth shattering but they do confirm intuitive readings of the works. More revealing is a search on *Ellénore* in *Adolphe* and the words (nouns, adjectives, verbs) in the immediate environment. She is in fact more often than not the object or beneficiary of verb constructions and when she is the subject it is frequently of verbs like *être* and *paraître*. Very often the name is in final position, followed by a colon or semi-colon as well as a fullstop. This could be construed as stylistic confirmation of her role as victim.

3. Complications arising from searching in French

There are however a number of technical problems that are more acute in French than English. The examples of *bien que* quoted above include two forms of *que*, with and without elision. Entering *que* will miss all forms of *qu'*. Entering *qu* may well fail to find both. You can simply enter both forms but this will require two searches. Entering a 'mask', using the multiple character wildcard * (*qu*),* will find both but will also find unwanted forms like *bien quand*, *bien quel* and *bien quelques*[28]. The size of the problem will depend on the precise search term and the corpus[29]. An alphabetical sort of the results will often allow you to extract the forms you want but this does require extra work. The verb *espérer* presents a different problem, the alternation of stem vowel between *é* and *è*. The vowel problem can be resolved by using the single character wildcard ? representing any single character. This still leaves the problem of endings, of which there are 33 in all (including past historic and imperfect subjunctive[30]). The simplest solution would be to enter 33 different forms but this would be time consuming, prone to error and slow, necessitating 33 consecutive searches. The most practical solution is to use the wildcard * and enter *esp?r**. Because of the particular letter pattern of *espérer* the only other word that fits this pattern is *espérance* and this can, depending on the software, be dealt with by entering it as a word to exclude. If

[28] to solve this problem, KWIC-CONCORD uses the slash to indicate possible elision, e.g. *qu/e*

[29] a similar search performed on the reflexive pronoun *se* produced 41562 examples of *s**, 5029 examples *s/e* but only 4872 examples of *s/e* if linked to *s'il* as word to exclude.

[30] obviously some of these are very unlikely to occur in a modern corpus.

you do the same with *mener*, by entering it as *m?n**, you will end up with 5,535 finds of which only 165 are appropriate. They can be grouped by an alphabetical sort and extracted relatively easily. A search performed on all forms of *connaître* using the mask *conn** will also throw up a number of unwanted forms most of which will be *connaissance* followed by derivatives of *connecter* and *connoter* plus a few oddballs like *Connery* and *Connaught*. A similar search on *savoir* would require the mask *s** which would find every word in the corpus beginning with s, leaving a nigh on impossible task of extraction. The solution depends on your software. If your concordancer allows alternative endings, you could enter *savoir* as *s+u,ais,ait,avons,……*[31] until all forms are included but the search would be slow as every word beginning with the letter s would have to be checked. Here one has to resort to a trade-off between time and ease of processing results. One compromise solution is to enter the following masks:

sai+s,t
sav+oir,ons,ez,ent,ais,ait,ions,iez,aient
sach+ons,ez,e,es,ions,iez,ent
saur+ai,as,a,ons,ez,ont,ais,ait,ions,iez,aient
su/t

A literary corpus would require expansion of the last line. In the end, the best solution may well be to enter the infinitive form *savoir*[32] since all you need is a representative rather than an exhaustive selection. Once you have eliminated all the nominal forms, this will probably leave you enough examples on which to base hypotheses.

4. Concordance Software

The concordancer used for the workshop was one I had created for my own use, KWIC-CONCORD (see note 1), and with which I was familiar. My main reasons for writing it were to handle the problems caused by accented characters, complex inflections and elision. Most of the programs available were written for use with English and these problems had not been high among their creators' priorities. I also wanted to be able to search any number of texts of any length, without having to pre-edit them, and to have no restrictions on the number of hits recorded. I also wanted to be able to manipulate the results on screen and to copy,

[31] The plus sign is the symbol used in KWIC-CONCORD to identify alternative endings. Different concordancers use different symbols.

[32] the same holds for *connaître*.

save or print them. The program was conceived essentially as a support for teaching rather than a sophisticated research tool. It is a ‘streaming’ concordancer that operates ‘on the fly’, reading text in line by line and saving to screen and file, and does not rely on building a massive database. It can therefore be made available for students to use without administrators worrying about demands on disk space or where institutional policy is to delete non-essential data files. The most sophisticated program from the research point of view must be *WordSmith Tools*[33]. This is highly flexible as a concordancer and, because it constructs complex databases, allows highly sophisticated statistical manipulation of the results. The author however recognises that it may be too complicated for use with undergraduates. There are several other widely used concordancers. The simplest is *Concordancer for Windows*[34] which was written for Windows 3.1 but allows a variety of search patterns to be entered and also collocation sorts. *MonoConc*[35] and *Concordance*[36] both provide very good collocation information. The latter will also produce a very fast total concordance of a corpus but includes numbers and punctuation symbols which distort the results. It also does not like long lines of text. All produce fast concordances and frequency lists but do not handle phenomena like elision and inflections very well.

Conclusion

The usefulness, not to say necessity, of concordancing in lexicography and text linguistics is now almost universally recognised. Searching and analysing a large corpus is extremely time consuming, all the more so if it is part of a project to create new reference or teaching materials. However, it is difficult to believe that anyone would now formulate a lexical or grammatical rule without first checking it against an appropriate corpus. Intuitions are often unreliable and liable to be overtaken by changes, welcome or otherwise. It is especially difficult to follow the evolution of a language when one does not live and work in a country where it is spoken. If you are not a native-speaker your exposure to the language will have been less and there will inevitably be gaps in your knowledge. Your students will know even less than you do and be more prone to L1 interference. Quick concordance searches not only provide authentic examples of use for

[33] see the author’s home page http://www.liv.ac.uk/~ms2928/index.htm

[34] this is freeware – for more information contact martinek@top.cz or msiegrist@hrz1.hrz.th-darmstadt.de

[35] to download a demonstration version go to http://www.ruf.rice.edu/~barlow/mono.html full price $79

[36] a trial version is available at http://www.rjcw.freeserve.co.uk/ full price $89 + $10 or £55

incorporation into teaching materials but enable the testing of individual intuitions as well as the reliability of reference and teaching materials. Remember though that the concordancer is only a tool. It cannot ask questions nor give answers. It will, however, help you to find with remarkable speed the data which may (or may not!) enable you to formulate the answers you seek.

Jeremy Whistle

APPENDIX 1

blèmes d'adaptation sont alors apparus. En effet, le livre est un peu un voyage intéri
floraison de médias libres est apparue sur l'ensemble de la planète. Le deuxième chang
liards de kWh. Ainsi donc sont apparues sur le plan international des possibilités d'i
est aussi à Futuropolis qu'est apparue cette volonté de considérer la BD comme un genr
ces nouvelles générations sont apparus sur les scènes, nombreuses et variées, du théât
nu la réussite. Libération est apparu dans un contexte très particulier. Il y a eu des
litiques ». La dissymétrie est apparue plus tard. D'une part au moment de la ratificat
er son efficacité, il nous est apparu impératif de nous démarquer de l'interprétation
els, d'autre part. Il nous est apparu essentiel de s'attaquer à ce déséquilibre, auque
ique. Cette discrimination est apparue suffisamment intolérable pour que des associati
ne capacité d'intervention est apparu. D'un côté, il y a quand même des difficultés po
rente dernières années, il est apparu que le développement de l'intégration européenne
ais de nouveaux cinéastes sont apparus ces dernières années, qui construisent une œuvr
n et les images sont peu à peu apparus. Les ministres se sont, eux aussi, mis au goût
rancophone. Plus de titres ont paru au cours des cinq dernières années, disent les spé
ton, professeur à Harvard, est paru en 1997 chez Odile Jacob (Paris). 3. Président du
es villes étrangères sont déjà parus et grâce à douze coéditeurs, un réseau internatio
oire des femmes japonaises est paru, s'inspirant de la nôtre. Ainsi, elle a eu un effe

APPENDIX 2

e un "France Info étranger". J'espère simplement que dans dix ans, cette radio existera
t qu'il faut incontestablement espérer que se mette en place une confédération mondiale
nt modestes. Les organisateurs espèrent simplement que la fête débordera du stade et de
table. Cela dit, je continue à espérer que cela n'est pas leur position définitive. Je
u Conseil de l'Europe. Il faut espérer que le système européen pourra garder son effic
rite Duras était quelqu'un qui espérait toujours que la vie serait plus intéressante q
de dialogue peut-être. On peut espérer que la féminisation du monde du travail soit sy
lle façon de vivre ensemble. J'espère que ces célébrations contribueront à donner l'im
tendances. En 1700, qui aurait espéré que la variole serait éradiquée ? En 1800, que l
des lecteurs à long terme. En espérant que, ce faisant, je fabrique en même temps des
nce et en Europe, nous pouvons espérer que la présidence française fera avancer l'Euro
e s'accepter comme mortel et d'espérer que cet enfant va rester après vous. C'est ce q
lettres de noblesse. Lui, qui espère que "le style Starck n'existe pas", se renouvell

APPENDIX 3

Label corpus

prospérité de la petite cité, après que Turgot, intendant du Limousin […] en ordonna[37]
réussi à être porté à l'écran après que Luis Buñuel, Costa-Gavras, […] en eurent rêvé
Mérimée, le père de Carmen, et après que la grande romancière Colette eut choisi sa ré
tte modestie que l'on acquiert après que l'on a beaucoup appris et beaucoup donné à ap
et beaucoup donné à apprendre, après aussi que l'on a connu les honneurs au sein d'une
, qui arrêta net tant de tirs, après que Youri Djorkaeff fut allé réconforter Ronaldo,
ène est soudain. Il se produit après que le disque solaire s'est amenuisé pendant envi
étaux ont été plantés, parfois après que les pépiniéristes eurent appris à cultiver de
ommuniquer avec les Etats-Unis après que toutes les lignes avaient été coupées. C'est *control corpus 1*
majorité", écrit FRANCE-SOIR, après que le Président du groupe UDF […]ait de son prop
d'offres de 0,2 point à 3,35% après que la BUNDESBANK eut ramené son taux de prise en
t le plus gros titre du FIGARO après que l'armée irakienne ait tiré trois missiles sur
manifestants se sont dispersés après que VAZGUEN MANOUKIAN eut annoncé que la commissi
control corpus 2 (oral)
i découverte d'ailleurs ici... après qu'on ait commencé à, à diriger le centre là et e

Jeremy Whistle

University College Northampton

[37] here and elsewhere (indicated by […]) the verb has had to be recovered from the wider context