ISSN 1442-438X
CALL-EJ Online, *3*(2), 37-52

Corpus-based computational linguistics: A practical investigation of the procedures involved in the selection, study and exploitation of a relevant corpus

Sean Romano Maddalena (srm@gol.com)
University of Ashiya, Japan

Abstract

This paper charts a corpus analysis research investigation that was conducted in response to a classroom question. The linguistic features under investigation are "used to" and "be used to"; two grammatical forms whose constructional similarity often causes problems for beginner-level students. This intentionally limited study outlines, by way of a step-by-step approach, the practical procedures involved in the assimilation and manipulation of computer-generated data. It is hoped that novice investigators may gain some valuable insight as to what even simplistic inquiries can bring for themselves as linguistic theorists, and to their learners embarking on a greater understanding of language meaning and usage.

A Brief History of Corpus Linguistics

Studies of language can be divided into two main areas: studies of structure and studies of use. Corpus analysis (CA) focuses on the second of these, studying actual language used in naturally occurring texts. Ever since Firth (1957) stated that "You shall know a word by the company it keeps", it has been a practice in linguistics to classify words not only based on their meanings but also based on their co-occurrence with other words. However, in a purely practical sense, it is only in recent times that machines have given us the ability to identify these relationships in a meaningful and significant way.

From the simple listing of words in the Middle Ages by hand to the earliest corpus-based analyses of literary styles, through to the first modern electronically readable corpus, the Brown University Corpus of American English, (and its close cousins the Lancaster-Oslo/Bergen corpus and the Kolhapur Corpus), the computer-aided analysis of vast amounts of authentic data has come a long way in a very short time. Almost half a century ago Firth (1957: 31) made the following prophetic statement: "The use of machines in the linguistic analysis is now established". John Sinclair (1991: 1) describes the evolution through the last three decades in the following way: " Thirty years ago when this research started it was considered impossible to process texts of several million words in length. Twenty years ago it was considered marginally possible but lunatic. Ten years ago it was considered quite possible but still lunatic. Today it is very popular". This popularity has led to an increased understanding of the relationship of meaning to form as formal patterns, previously undetected, have come to light. Sinclair states again, "At the very least, the quality of linguistic evidence is going to be improved out of all recognitionq¥Êt is my belief that a new understanding of the nature and structure of language will shortly be available as a result of the examination by the computer of large collections of texts"

(1991b: 489). Stubbs (1996) concurs, "computer-assisted analysis of texts and corpora can provide a new understanding of form-meaning relations".

It should be noted that CA involves far more than using computers for the simple counting and quantifying of linguistic features into sets of statistics. Though this may be seen as the first step in a two-stage process, it is the subsequent, qualitative analysis that provides the more revealing evidence "to propose functional interpretations explaining why the patterns exist" (Biber, Conrad & Reppen, 1998: 9). As a practical investigation, however, this paper focuses primarily on the procedures involved in obtaining and manipulating the data required to create a corpus, and while it does present some insight into possible pedagogic considerations and offer tentative conclusions based on corpus generated evidence, its scope is intentionally, limited.

Choosing a Corpus

Source, size, and selection

In response to a recent classroom inquiry, the linguistic features under investigation are "used to" and "be used to"; two grammatical forms whose constructional similarity often causes problems for beginner-level students. For this investigation, I chose to use two established corpora, the Lancaster-Oslo/Bergen Corpus (LOB), of British English established by Geoffrey Leech and Jan Svartvik, and its American counterpart, the Brown University Corpus of American English (Brown), running parallel investigations under different methodological conditions. The two corpora are very similar in design: each taken from a total of some five hundred texts across a wide range of registers, a combined total of approximately two million words.

Size is a prime concern for successful corpus-based lexicographic research. As Biber et al. warn: "To study the meaning and use of words, we need a very large corpus — a 1-million word corpus will not provide sufficient data for many words to allow for meaningful generalizations" (1998: 30). However, with more common words in a text of this size, frequencies are generally considered to be quite reliable. At a million or so words each, I was hoping that my choice of general-purpose corpora would provide enough evidence to sufficiently highlight linguistic elements for possible future pedagogic exploitation.

Methodology

As primarily a practical research study, I chose to conduct this investigation employing a number of differing methods. In the first instance, I examined the LOB corpus using a CD-ROM provided by the International Computer Archive of Modern English (ICAME), running the analysis through a software application, the Aston Text Analyser (ATA), supplied by Aston University. I also used part of the LOB corpus to examine the practical problems one might encounter in the creation of a pedagogic corpus, established corpora not always being readily available for investigation and exploitation.

As a reflection of recent advances in Internet technology, I was also interested in conducting a limited parallel study, making use of an on-line version of the Brown corpus,

a free but time-restricted service provided by the University of Pennsylvania's Linguistic Data Consortium, (LDC). Details of distribution and copyright restrictions of both texts are included, (Appendix C).

It should be noted here that although the Brown corpus is also supplied on the ICAME CD-ROM, I chose not to access it in the traditional way preferring instead to examine the benefits and shortcomings of locating and accessing corpora via the alternative, and increasingly popular, on-line method.

Equipment Used

The study was conducted with the aid of a generic desktop personal computer running the Windows operating system. Software support was provided by the WinATA Mark 2 text analyser, a word processor, MS-Word 97, and an Optical Character Recognition (OCR) program, Caere Omni-Page Pro 9.0 used in conjunction with a flatbed scanner.

Data Input: Scanning and OCR

Equipment and procedure

In some instances, teachers and researchers may not have access to established corpora due to resource limitations. In other cases, most notably for investigations in English for Specific Purposes (ESP), it might be necessary to manually create a specific pedagogic corpus. In creating such a corpus for use in CA, one possible means of inputting data is to scan text directly into a computer using a suitable combination of hardware and software. To explore the limitations of such a procedure, I used a Microtek ScanMaker X6 scanner, a low budget flatbed model, together with Caere Omni-Page Pro 9.0 OCR software, which was supplied as part of the scanner package.

For the limited purposes of this exercise, I first selected a section of some five hundred words from my LOB corpus, cut and pasted them into a new document, and saved this as a separate text file. This was then printed onto a standard sheet of A4 paper and then scanned directly into the computer. Almost flawless text conversion is testimony to the development of OCR software in recent times. A few years ago a similar exercise may well have resulted in a bout of severe frustration, even when scanning a simple page of text. These days, more advanced programs such as Omni-Page Pro offer much greater speed, reliability, and flexibility, especially when integrated into established word processing applications such as Word and Word Perfect. Carefully scanned pages of text assimilated in this way can form the basis for a ‘personal’ pedagogic corpus, to be subsequently examined by a suitable text analysis program.

Some Points to Note

There are two significant considerations that can affect the quality of the final output from the scanning procedure. Firstly, and most importantly, is the quality and condition of the document that one wishes to scan. I was using a printed black text on a clean sheet of white plain paper. Highly colored, glossy, marked, or even creased papers have all been known to cause problems with OCR software. The second consideration relates to the

complexity of the document. As my inquiry revealed, regular text is not a problem for this kind of application. However, when one mixes text, graphics, and tables, more time needs to be spent in the setup process before attempting the conversion. I also found in this exercise that the software occasionally flagged correct words simply because they were not in the dictionary it was using.

LOB and ATA

Installation

Installation of the ATA software suite is via CD-ROM. It is important to note during the installation process that for the software to function correctly, all files must be extracted into the same location and not into separate folders. The correct installation creates two executable programs; ataIndex and ataInsight which must be run separately, one after the other. The first of these, as the name suggests, creates and indexes the corpus. In the case of LOB, this entails specifying the correct path for the location of the text to be indexed and titling the project appropriately. When the indexing has been completed, it is then necessary to run the second application, ataInsight. This opens an 'Open ATA project' window in which the now indexed LOB text can be found. On selecting 'OK', the program starts its analysis of the chosen project.

Frequency and filter

My investigation is to specifically look for occurrences of "used to" within the corpus. To do this, it is first necessary to locate "used" from the 'Word Frequency List' which opens automatically on the left side of the screen. Selecting this entry, (with 'Collocations' checked in the right-button mouse menu) creates a list of contexts in a right-hand window; some 181 entries in total.

Next, it is desirable to refine a little further using the collocation 'Filter' option, reducing the list to those lines containing my chosen sub-string. Adding "to_" to the filter generates a final list of 178 concordances which contain only my target search string, "used to". By selecting 'Export' from the right-button mouse menu, concordances can then be exported with relative ease from within the application and opened in a word processor, ready for tabulation, (Appendix A). From a total of 1,022,828 tokens, the following frequency list is generated. Relative frequencies are out of 10,000:

Type	Raw frequency	Relative frequency
to	26375	257.86
used	644	6.29
used to	178	1.74

Fig.1 LOB Corpus frequencies for "to", "used" and "used to".

Observations

Presentation, an important consideration not merely for aesthetic purposes, also demands a practical working knowledge of basic word processing operations. Ideally, for beginner-level students, concordances are presented in a clear and easy to read tabular format, sorted alphabetically to enable the swift identification of collocation patterns, (Appendix A and Appendix B).

Brown Corpus

As mentioned above, the Brown corpus is accessed through the University of Pennsylvania's LDC internet site. It offers a selection of corpora for real-time analyses through access, as a 'guest user' is restricted to twenty days. On acceptance of the user terms and conditions, one is invited to enter the relevant search criteria in a series of selectable fields.

An initial search returns a tagged frequency list and generates concordances for the identified search pattern. The complete list of Brown concordances is provided in their processed form, (Appendix B).

From a total of 1,189,209 tokens, the following frequency list is generated. Once again, relative frequencies are calculated out of 10,000:

Type	Raw frequency	Relative frequency
to	24619	207.01
used	594	4.99
used to	134	1.12

Fig.2 Brown Corpus frequencies for "to", "used" and "used to".

Observations

Established corpora are often the culmination of a great deal of time, effort, and, most significantly, money. Such investment is jealously guarded and may not, therefore, be made generally available without due considerations of costs. In some cases, this may prove to be prohibitive to the less fortuitous researcher. In this light, it can be seen that the ability to access a large on-line corpus in real-time is extremely useful for those unable to avail themselves of the more traditional resources, and also appealing to those who lack the practical wherewithal necessary for the successful exploitation of a complicated text analysis program. Such corpora also offer the added benefit of speed; a list of concordances can be generated in a matter of seconds. However, at this early stage of development, the on-line corpus does not yet offer the flexibility or power of a dedicated software package, such as ATA, to sort or to filter, as need dictates.

Analysis

The majority of the concordances in LOB are taken up with “used to” employed to describe past situations and events. There is a visible tendency within the list to collocate with the verb “to be” and also with other common verbs:

- as fresh as it used to be, though an
- you herself what she used to be.
- But then I used to be a racing
- reading ," wrote Francis Williams,” used to be a Socialist

The corpus provides twenty-eight instances of “be used to” meaning to be “accustomed to”. The propensity is for the item to collocate with a noun or a verb, notably the gerund. Of the total number, only eleven occur with the gerund, which is the collocate most commonly highlighted in beginner-level textbooks. Textbooks also tend to focus on the gerund occurring after the target form:

- time before I got used to calling them portholes.
- Clara was used to following his lead
- seemed to have been used to seeing couples engaged

whereas LOB offers examples of the gerund occupying a position before the target form:

- a bit of getting used to
- She took time getting used to the indoor lavatories

And a single instance of a noun coming between the two:

- garage, but he was used to Grant taking his

A further significant observation is that more than half of these concordances demonstrate collocations with the verb “get”:

- You'll have to get used to my bad morning
- heavy, but one got used to this

Though not the focus of this particular exercise, the list also provides some examples of the target form performing a third linguistic function, the passive voice:

- descriptions can also be used to refer to performances
- ratio decidendi}is normally used to refer to some
- beggars, a term often used to describe the population,
- ferromagnetic spinel is sometimes used to describe those ferrites

With Brown, as with LOB above, “used to” describing past events tends to collocate with the verb “to be” and other common verbs:

- eem high, but they used to be even higher,
- spe said, This soil used to be like that
- ard roll. <s> This used to be part of

Also present, as noted in LOB, are instances of "used to" employed in the passive voice:

- ma. The method used to scan the eye
- I rand, IOCSIXG, is used to specify the second

The Brown corpus offers twelve examples of "used to" meaning to be "accustomed to"; less than half of the total number present in LOB. Of these, only five collocate with the gerund:

- ke a little getting used to — not because it
- ur people have been used to accepting things as
- that must have been used to booming, `` and th
- he governor was not used to having his integrit
- jealous. <s> He's, used to me bringing home

and only two of the twelve co-occur with the verb "get":

- ke a little getting used to — not because it
- little time to get, used to. After a

Possible Pedagogic Applications

In the classroom, concordances produced through the analysis of a suitable corpus can provide valuable data for the testing of existing grammatical models and practical material for the production of cloze exercises. A closer examination can also reveal patterns and constructions that may not be covered in prescribed textbooks.

The initial intent of this study was to examine the differences in usage between "used to" and "be used to". My learners do not have a significant problem with the former but do express confusion when attempting to differentiate it from the latter. My institution's current choice of text-only instructs in the use of "be used to" co-occurring with the gerund and, consequently, my students have only been exposed to this construction in their English classes. However, the majority of these concordances in Brown and LOB occur with no gerund at all, a point worthy of highlighting in the classroom. Though different in meaning, the number of cases of "get used to" provided by the corpora, most prominently LOB, may be seen as noteworthy and also deserving of my students' attention, as this particular construction is not covered in the students' textbook at all. A practical pedagogic approach to both of these issues would be to expose my students to the corpus-generated data as part of a series of carefully coordinated lessons. Through the insights I have gained in the course of this particular study, my eventual aim would be to bring CA directly into the classroom, possibly as part of the

school's regular computer studies classes, and allow my students to join the investigation as part of a hands-on practical exercise.

However, to add a note of caution, as my small investigation reveals, there are significant differences in both frequency and usage to be found even across two very 'similar' corpora. It is important therefore to make only tentative inferences regarding grammatical rules or patterns of use and to acknowledge the limitations of dealing with such a small sample of data. A future piece of research conducted on a much larger text might allow for some more definite conclusions to be made.

A further possible pedagogic option, requiring an extension of this study, would be to heed the advice of Willis & Willis (1996) and Peacock (1997: 152) to produce a set of authentic materials: "materials which are used in genuine communication in the real world" (Wong, Kwok & Choi, 1995: 318), taken from a spoken, rather than written, corpus and to investigate specifically any increased signs of motivation with my less-conscientious learners.

It is perhaps a fitting conclusion to note that in the course of writing this paper a further development in the evolution of computational linguistics and the internet is reported: ICAME is now the latest in a growing number of institutions offering on-line access to all of its corpora, in this case to registered users of its commercially available CD-ROM. It seems likely that such innovations, offering increased levels of accessibility to an ever-growing body of linguistic data, will continue into the foreseeable future.

References

Biber, D., Conrad, S., & Reppen, R. (1998).*CORPUS LINGUISTICS: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Brown University Corpus of American English.

University of Pennsylvania, Linguistic Data Consortium: http://www.ldc.upenn.edu/

Firth, J. R. (1957). A synopsis of linguistic theory. Studies in linguistic analysis. Oxford: Oxford University Press.

Lancaster-Oslo/Bergen corpus (1961). International Computer Archive of Modern English. Bergen, Norway.

Peacock, M. (1997). The effect of authentic materials on the motivation of EFL learners. *ELTJ*, 51(2), 144-156.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stubbs, M. (1996). *Text and Corpus Analysis*. London: Blackwell.

Willis, J. & Willis, D. (1996). Consciousness-raising activities. In D. Willis & J. Willis, (Eds.), *Challenge and Change in Language Teaching*. London: Heinemann.

Wong, V., Kwok, P. & Choi, N. (1995). The use of authentic materials at tertiary level. *ELTJ*, 49(4), 318-322.

Appendix A

a bit of gettingq_	**used toq_ .**
plane can only beq_	used toq_ a limited extent
of the man habituallyq_	**used toq_ a shoulder-holster .**

such computers can beq_	used toq_ advantage when a
Gissingq_	used toq_ ask ~ * ' Has he
affluent society should beq_	used toq_ assist the less
Rolled barley isq_	used toq_ balance grass or
as fresh as itq_	used toq_ be , though an
you herself what sheq_	used toq_ be .
man myself though : Iq_	used toq_ be a { 0G.P . }
But then Iq_	used toq_ be a racing
reading , " wrote Francis Williams , "q_	used toq_ be a Socialist
done by administrative actq_	used toq_ be accomplished in
the subject Social Psychologyq_	used toq_ be called Home-making
of the May songq_	used toq_ be current in
Itq_	used toq_ be fancier , but
At one time " mind * * "q_	used toq_ be identified with "
of their larger carsq_	used toq_ be made available
her hair , it neverq_	used toq_ be quite that
This lessonq_	used toq_ be read only
Sometimes that pleasant Citroenq_	used toq_ be subject to
Harry of the jointq_	used toq_ be the barman
Thereq_	used toq_ be three separate
I was younger Iq_	used toq_ be what is
Like heq_	used toq_ be years ago . . .
three feet long butq_	**used toq_ being handled , in**
of the gold filletsq_	used toq_ bind up the pŽ º/span>
Miniature cedar trees areq_	used toq_ block out the
technical school) should beq_	used toq_ broaden the youngsters '
British sources have beenq_	used toq_ calculate the effective
time before I gotq_	**used toq_ calling them portholes .**
I alwaysq_	used toq_ clean my rifle
Heq_	used toq_ come every day
Heq_	used toq_ come to Pierre's
remember a woman whoq_	used toq_ come to see
at Saintes , has beenq_	used toq_ complete the drawing
have been or areq_	used toq_ control impurity build
is what bedizened boysq_	used toq_ dance before Mogul
its phrases , especially thoseq_	used toq_ describe a visit
Kunst wasq_	used toq_ describe certain branches
with the conventional equationsq_	used toq_ describe fluxes in
unit , can be properlyq_	used toq_ describe soils in
a root that isq_	used toq_ describe the herding
as the wave functionsq_	used toq_ describe the motion
equation can indeed beq_	used toq_ describe the motion .
however , they may beq_	used toq_ describe the motions
beggars , a term oftenq_	used toq_ describe the population ,
ferromagnetic spinel is sometimesq_	used toq_ describe those ferrites

method of measurement wasq_	used toq_ determine accurately the
year group was thenq_	used toq_ determine what would
his Cambridge days , heq_	used toq_ display a corresponding
elaborate dresses than theyq_	used toq_ do .
Mould many years backq_	used toq_ do .
Peopleq_	used toq_ do all their
strain , the two beingq_	used toq_ draw true stress /
young the Royal Navyq_	used toq_ drink it before
Heq_	used toq_ drink the cheap ,
that report has beenq_	used toq_ estimate the theoretical
diametrically opposed contacts wereq_	used toq_ facilitate the observation
gouge , and the fileq_	used toq_ finish off .
the former crop beingq_	used toq_ finish off the
Clara wasq_	**used toq_ following his lead ,**
The method wasq_	used toq_ forecast visibility (as
concrete tube sections beingq_	used toq_ form the sump
smoothing plane can beq_	used toq_ form the taper .
Bank years ago weq_	used toq_ get good hauls , 12
song , told me : Weq_	used toq_ get up at
This solution may beq_	used toq_ give the contribution
those places where weq_	used toq_ go .
much as Cecil Sharpq_	used toq_ go about in
Sheq_	used toq_ go about the
garage , but he wasq_	**used toq_ Grant taking his**
Iq_	used toq_ hate Creedy , when
for a drink heq_	used toq_ have his grouse .
The Caxtonsq_	used toq_ have their holidays
told me " I alwaysq_	used toq_ hear a lot
Weq_	used toq_ hear talk about
took time to becomeq_	**used toq_ hearing so much**
household possessions may beq_	used toq_ help with the
Apparently heq_	used toq_ hide it in
they may be fruitfullyq_	used toq_ His Glory .
and these can beq_	used toq_ illustrate the type
overclothe them as theyq_	used toq_ in the old
The term quasi-classical isq_	used toq_ indicate that their
growth equilibrium " paths , areq_	used toq_ investigate the stability
man , if you aren'tq_	**used toq_ it , * * ' he heard**
You'll getq_	**used toq_ it , adorable baby .**
that we should getq_	**used toq_ it .**
I never gotq_	**used toq_ its travel-film colours**
Two methods can beq_	used toq_ join the crochet
differences between jobs beq_	used toq_ justify differences in
a young man , weq_	used toq_ keep strictly to
to meet people Iq_	used toq_ know , to see

electric effect can beq_	used toq_ launch ultrasonic waves
Iq_	used toq_ lie awake planning
a counter-irritant almost Iq_	used toq_ listen of nights
Marc Chagallq_	used toq_ live here and
Then that's why * - " " Heq_	used toq_ live in Tangier , "
Theyq_	used toq_ look * - and some
of an elephant , wasq_	used toq_ make a cake
Some separated lead-210 wasq_	used toq_ make reference standards
crochet lace can beq_	used toq_ make tablecloths , traycloths
provision which was nowq_	used toq_ make the { 0T.E .
ancient Britons , I believe ,q_	used toq_ make water hot
as it is nowq_	used toq_ mark a paragraph
Section the term wasq_	used toq_ mean something like
Georgeq_	used toq_ mix 100 stone of
junior to Humbert , whoq_	used toq_ mock him affectionately
You'll have to getq_	**used toq_ my bad morning**
gauge can now beq_	used toq_ nick in the
three following winters wereq_	used toq_ obtain an independent
Heq_	used toq_ organise film shows
which can then beq_	used toq_ perform an operation .
and devices to beq_	used toq_ perform the various
Iq_	used toq_ play about in
Iq_	used toq_ play rugger , * * ' said
lead carrier solution isq_	used toq_ prepare the reference
how Alexander the Greatq_	used toq_ recline and transact
descriptions can also beq_	used toq_ refer to performances
ratio decidendi } is normallyq_	used toq_ refer to some
it may have beenq_	used toq_ relate Christ's healing
migre * ? 2s , who notoriouslyq_	used toq_ repair to the
she said chattily , Iq_	used toq_ ride a bicycle .
and personality which journalistsq_	used toq_ ridicule , can be
the gate the cockerelq_	used toq_ run to meet
for you fellows , * * ' heq_	used toq_ say , you can
Laughable , theyq_	used toq_ say .
Heq_	used toq_ say : ^ Have whatever
Of Kitchener heq_	used toq_ say with humorous
reminiscent of what weq_	used toq_ see pŽ ®St .
seemed to have beenq_	**used toq_ seeing couples engaged**
embarrassment if she isq_	**used toq_ seeing her mother**
that force should beq_	used toq_ settle this problem .
the May carol heq_	used toq_ sing , with his
me the one sheq_	used toq_ sing in Kimbolton
a shaped rubber isq_	used toq_ smooth the hollow
was young schoolboy I	used toq_ sneak off to
Sheq_	used toq_ solve all the

the clinical weekends heq_	used toq_ spend with her .
applied , and every meansq_	used toq_ stop the train ,
in contrasting tones wereq_	used toq_ strengthen garments at
model which may beq_	used toq_ study both the
Heq_	used toq_ stump round the
possibility of power beingq_	used toq_ supplement hand tools .
Iq_	used toq_ take the small
and colleague , Campbell Dixon ,q_	used toq_ tell of a
The straight-edge can beq_	used toq_ test the straightness
is bought , can beq_	used toq_ the best advantage .
at (B) . A malletq_	used toq_ the chisel is
become (1) tired , or (2) moreq_	**used toq_ the disturbance .**
Soho , to get meq_	**used toq_ the food , he**
might as well getq_	**used toq_ the idea .**
they very quickly getq_	**used toq_ the idea of**
She took time gettingq_	**used toq_ the indoor lavatories**
They'req_	**used toq_ the snatch racket .**
that most people getq_	**used toq_ them .**
Jane wasq_	**used toq_ these sudden exigencies**
or chieftain to getq_	**used toq_ these trimmings because**
to tinsel compliments , weq_	used toq_ think him unworldly ,
in an Embassy * - Iq_	used toq_ think it was
heavy , but one gotq_	**used toq_ this .**
You are not yetq_	**used toq_ this sort of**
decorative kale are convenientlyq_	used toq_ tone in with
horses ; they had beenq_	**used toq_ trains since they**
The brush contacts wereq_	used toq_ trigger off a
He oftenq_	used toq_ try to imagine
His friendsq_	used toq_ try to persuade
friend , William James , whoq_	used toq_ urge that the
in London that Jonesq_	used toq_ use in the
slaves * - everything he wasq_	**used toq_ using while he**
a literary province Iq_	used toq_ visit fairly often ;
Sheq_	used toq_ walk straight to
Heq_	used toq_ walk to the
page , would have beenq_	used toq_ weigh bales of
They could beq_	used toq_ weigh several sacks
its simplest form itq_	used toq_ work in the
they are a teamq_	**used toq_ working together , they**
like that she hadq_	used toq_ write to me .

Appendix B

ke a little gettingq_q_	**used to -- not because it**
iling teasing as heq_q_	used to . <p> <s> `` Huskyq_q_q_q_
from it that sheq_q_	used to . <p> <s> `` You
little time to getq_q_	**used to . <s> After a**

ur people have beenq_q_	**used to accepting things as**
a new melody isq_q_	used to accompany his narraq_q_q_
repetitious The logical schemeq_q_	used to accomplish the formq_q_q_q_
residual hese inquiries wereq_q_	used to adjust compilationsq_q_q_ tient
uestions . <s>I 'mq_	**used to all three , but**
herse one hebephrenic manq_q_	used to annoy me , month
ageq_ seven-iron shot heq_q_	used to approach the greenq_q_q_q_
s> They could beq_q_	used to attack a nation '
platform and can beq_q_	used to automatically holdq_q_q_q_ iling
citiz--uglier than youq_q_	used to be , and you
ss glorious than itq_q_	used to be , it is
nistered here as itq_q_	used to be , with unleaveneq_q_q_
or less than itq_q_	used to be ? ? <p> <s>
eem high , but theyq_q_	used to be even higher " ,q_q_q_
spe said , This soilq_q_	used to be like that
ard roll . <s> Thisq_q_	used to be part of
as e Catskills , whichq_q_	used to be the summer
that must have beenq_	**used to booming , `` and th**
ese profiles can beq_q_	used to calculate a temperaq_q_q_
feeli ransports that wereq_q_	used to carry Communist ageq_q_q_q_
the mails were thenq_q_	used to carry it out " . <q_q_q_q_
tional codes can beq_q_	used to challenge and countq_q_q_q_
and d Margaret recall ,q_q_	used to characterize her asq_q_q_ >
les of crystals areq_q_	used to classify and identiq_q_q_
of materials can beq_q_	used to construct a satisfaq_q_q_
cattle of thousand spectatorsq_q_	used to crowd it in
holes and can beq_q_	used to cut exact-size discq_q_q_
the words he hadq_q_	used to defend Cromwell . <q_q_q_ he
grea emical methods wereq_q_	used to demonstrate the renq_q_q_
K factor , a termq_q_	used to denote the rate
s> Mines can beq_q_	used to deny access to
elastic resonance shifts isq_q_	used to derive a general
was a Spanish wordq_q_	used to describe cattle ofq_q_q_q_
s ,sometimes it isq_q_	used to describe felt humanq_q_q_
integritq_ ind words travelersq_q_	used to describe Little Rocq_q_q_
prbody temperature isq_q_	used to describe the radiatq_q_q_
e aircraft could beq_q_	used to destroy other mobilq_q_q_
ese sound waves areq_q_	used to detect submarines ,q_q_q_ ma . <
the the anonymous Womanq_q_	used to do , and he
each time as heq_q_	used to do . <s> When
second aerated lagoons beq_q_	used to eliminate the problq_q_q_
h tiles , marble areq_q_	used to emphasize the feeliq_q_q_
ve operation EQU isq_q_	used to equate symbolic namq_q_q_
d transom which wasq_q_	used to fasten them to

a satisfa lf-unloading wagonsq_q_	used to fill silos spreadsq_q_q_q_
tten 2 B filter wasq_q_	used to filter off residualq_q_q_
er last week ,Iq_q_	used to follow Williams eveq_q_q_
power which can beq_q_	used to frustrate the citizq_q_q_q_ --
atement may also beq_q_	used to generate an RDW
old days when `` weq_q_	used to get the seamen
af A hebephrenic manq_q_	used to give a repetitiousq_q_q_q_q_
was another . <s> Iq_q_	used to go with Watson
mulated that can beq_q_	used to good advantage . <pq_q_q_
eel lonely , and weq_q_	used to hang a sign
aps as the cave-menq_q_	used to have in the
he governor was notq_	**used to having his integrit**
and had already becomeq_	**used to Hesperus ' snappingq_q_q_ he**
eem strange to earsq_q_	**used to hillbilly and jazz**
and he was notq_q_	**used to horseback . <s> Now**
ngs Thorpe, can beq_q_	used to illustrate anotherq_q_q_q_q_ power
vocatio pleading cannot beq_q_	used to impose unnecessaryq_q_q_q_ h
nk together like weq_q_	used to in the old
the progr `` technology " isq_q_	used to include any and
of time is merelyq_q_	used to increase the realisq_q_q_
mobil rrently , marina isq_q_	used to indicate a municipaq_q_q_q_ **
w seldom they did :q_q_	**used to it , probably . <s>**
n tactics have beenq_q_	used to justify like tacticq_q_q_q_
spreads Computers are beingq_q_	used to keep branch inventoq_q_q_ <
the new jail , weq_q_	used to keep prisoners in
ng cover , could beq_q_	used to keep the wastes
the eye . <s> Weq_q_	used to kid him by
ny ? ? <s> He neverq_q_	used to like any hot
cereal aining appliance isq_q_	used to lock them in
c. <s> The Presidentq_q_	used to look at it
by the same methodq_q_	used to look up a
ith , Styka . <s> Iq_q_	used to love this country
he coconut palm areq_q_	used to make candles in
as urposes -- also areq_q_	used to make soaps , detergq_q_q_
of public places thatq_q_	used to make the Jew
zon apabilities must beq_q_	used to maximum advantage tq_q_q_q_ . <
jealous . <s> He 'sq_q_	**used to me bringing home**
count mimesis " is hereq_q_	used to mean the recallingq_q_q_q_q_
if it could beq_q_	used to measure the elasticq_q_q_
s> Sonar can beq_q_	used to measure the thickneq_q_q_
radiat ed thermocouple wasq_q_	used to measure the upstreaq_q_q_
aratus will also beq_q_	used to measure transitionq_q_q_q_ ese
s steel screws wereq_q_	used to minimize corrosionq_q_q_q_ e
The DA statement isq_q_	used to name and define

The DC statement isq_q_	used to name and enter
sample ; e bio-assay methodsq_q_	used to obtain them . <s>
tient of mine , whoq_q_	used to often seclude herseq_q_q_
s> yesterday .<s> Youq_q_	used to paint in them ,
ly state funds wereq_q_	used to pay for the
as a child Iq_q_	used to play " . <s> He
he corner where youq_q_	used to play when you
very summer . <s> Iq_q_	used to play with the
out "surpluses had beenq_q_	used to provide a private
ce forces have beenq_q_	used to provide defense zonq_q_q_
she ed aluminum plate ,q_q_	used to provide the dryingq_q_q_q_
asq_ Miss Giles alwaysq_q_	used to refer to her
most of what weq_q_	used to regard as the
ntic up there , sheq_q_	used to say , with the
of my ewish intellectualsq_q_	used to say . <p> <s>
se by instinct , heq_q_	used to say : such places
The party that wonq_q_	used to say something aboutq_q_q_
ma . <s> The methodq_q_	used to scan the eye
S statement must beq_q_	used to select the major
stem . <s> DIOCS isq_q_	used to select the major
b " . <s> It isq_q_	used to separate two or
s> The symbol isq_q_	used to separate two or
foam and can beq_q_	used to slit continuous sheq_q_q_
me rand , IOCSIXF , isq_q_	used to specify the first
I rand , IOCSIXG , isq_q_	used to specify the secondq_q_q_q_q_
upstrea equency starter wasq_q_	used to start the arc . <
erb garden was alsoq_q_	used to stop bleeding , andq_q_q_
a lock ,which isq_q_	used to store cumulative req_q_q_q_
Throu was constructed andq_q_	used to study transition prq_q_q_
corrosion e been successfullyq_q_	used to suggest ways to
than to an Americanq_q_	**used to summers in New**
pirical data can beq_q_	used to support whatever prq_q_q_
sort of thing thatq_q_	used to take place in
e evening . <s> Sheq_q_	used to tell me , `` When
South nt this opportunityq_q_	used to tell them about
ygous Af cells wereq_q_	used to test each sample ;q_q_q_q_
invento <s> Where Americansq_q_	used to think of a
unt of a machine-familyq_q_	used to this very day
he enemy-Jew can beq_q_	used to transform the ordinq_q_q_
i er , Model 565 , isq_q_	used to transport the boatq_q_q_q_
was a trick theyq_q_	used to try and conceal
San Juan , but Iq_q_	used to work on a

Appendix C

Copyrights and distribution:

LOB Corpus:

The corpus and accompanying manual are available at cost to bona fide researchers through the International Computer Archive of Modern English (ICAME), at the Norwegian Computing Centre for the Humanities, Bergen, Norway.
The following restrictions on the use of the material must be strictly observed:

- No copies of the corpus, or parts of the corpus, are to be distributed under any circumstances without the written permission of ICAME.
- Print-outs of the corpus, or parts thereof, are to be used for bona fide research of a non-profit nature. Holders of copies of the corpus may not reproduce any texts, or parts of texts, for any purpose other than scholarly research without obtaining the written permission of the individual copyright holders, as listed in the manual accompanying the corpus.
- Commercial publishers and other non-academic organizations wishing to make use of part or all of the corpus or a print-out thereof must obtain permission from all the individual copyright holders involved.

Brown Corpus:

The Linguistic Data Consortium grants to you a license to use this data subject to the following understandings, terms, and conditions:

1. Permitted Uses.
- This data may only be used for linguistic research.
- Small excerpts of text or audio data from LDC-Online materials may be displayed to others or published in a scientific or technical context, solely to describe the research and related issues. Statistics and other summaries of LDC-Online materials may also be published in the same context. Except for such publication of small excerpts or statistical summaries in scientific or technical works, neither LDC-Online materials themselves nor access to them may be sold or transferred to others.

2. Access by Individuals.
- To access this data, you must be a staff member, consultant, or individual providing service or researching an organization that is a member of the LDC, and you must agree to this user agreement and its provisions. You must terminate your access when these conditions no longer apply.

3. Copyright.
- Except as specifically permitted above the display, reproduction, transmission, distribution, or publication of these databases is prohibited.
- Violations of the copyright restrictions on the data may result in legal liability.