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 recognition. Yet 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”

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:

<table>
<thead>
<tr>
<th>Type</th>
<th>Raw frequency</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>to</td>
<td>26375</td>
<td>267.86</td>
</tr>
<tr>
<td>used</td>
<td>644</td>
<td>6.29</td>
</tr>
<tr>
<td>used to</td>
<td>178</td>
<td>1.74</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th>Type</th>
<th>Raw frequency</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>to</td>
<td>24619</td>
<td>207.01</td>
</tr>
<tr>
<td>used</td>
<td>594</td>
<td>4.99</td>
</tr>
<tr>
<td>used to</td>
<td>134</td>
<td>1.12</td>
</tr>
</tbody>
</table>

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,
spesaid. This soil used to be like that
ard. <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


Brown University Corpus of American English.

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


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 be used to advantage when a
Gissing used to ask ~ Has he
affluent society should be used to assist the less
Rolled barley is used to balance grass or
as fresh as it be, though an
you yourself what she be.
man myself though : I be a { G.P. }
But then I be a racing
reading , " wrote Francis Williams , " be a Socialist
done by administrative act be accomplished in
the subject Social Psychology be called Home-making
of the May song be current in
It be fancier , but
At one time " mind * * " be identified with "
of their larger cars be made available
her hair , it never be quite that
This lesson be read only
Sometimes that pleasant Citroen be subject to
Harry of the joint be the barman
There be three separate
I was younger I be what is
Like he be years ago . . .
three feet long but used to being handled , in
of the gold fillets bind up the pŽ³/space>
Miniature cedar trees are block out the
technical school ) should be used to broaden the youngsters '
British sources have been used to calculate the effective
time before I got used to calling them portholes .
I always used to clean my rifle
He used to come every day
He used to come to Pierre's
remember a woman who used to come to see
at Saintes , has been used to complete the drawing
have been or are used to control impurity build
is what bedizened boys used to dance before Mogul
its phrases , especially those used to describe a visit
Kunst was used to describe certain branches
with the conventional equation used to describe fluxes in
unit , can be properly used to describe soils in
a root that is used to describe the herding
as the wave function used to describe the motion.
equation can indeed be used to describe the motion .
however , they may be used to describe the motions
beggars , a term often used to describe the population ,
ferromagnetic spinel is sometimes used to describe those ferrites
method of measurement was used to determine accurately the year group was then used to determine what would his Cambridge days, he used to display a corresponding elaborate dresses than they used to do. Mould many years back used to do. People used to do all their strain, the two being used to draw true stress/young the Royal Navy used to drink it before Heq. that report has been used to estimate the theoretical diametrically opposed contacts were used to facilitate the observation gouge, and the fileq used to finish off. the former crop beingq used to finish off Clara was used to following his lead. The method was used to forecast visibility (as concrete tube sections beingq used to form the sump smoothing plane can beq used to form the taper. Bank years ago weq used to get good hauls, 12 song, told me: Weq used to get up at This solution may beq used to give the contribution those places where weq used to go. much as Cecil Sharpq used to go about in Sheq used to go about the garage, but he wasq used to Grant taking his Iq used to hate Creedy, when for a drink heq used to have his grouse. The Caxtonsq used to have their holidays told me "I alwaysq used to hear a lot Weq used to hear talk about took time to becomeq used to hearing so much household possessions may beq used to help with the Apparently heq used to hide it in they may be fruitfullyq used to His Glory. and these can beq used to illustrate the type overclothe them as theyq used to in the old The term quasi-classical isq used to indicate that their growth equilibrium "paths, areq used to investigate the stability man, if you aren'tq used to it, * * ' he heard You'll getq used to it, adorable baby. that we should getq used to it. I never gotq used to its travel-film colours Two methods can beq used to join the crochet differences between jobs beq used to justify differences in a young man, weq used to keep strictly to to meet people Iq used to know, to see
electric effect can be used to launch ultrasonic waves

Iq_ used to lie awake planning

a counter-irritant almost Iq_ used to listen of nights

Marc Chagalliq_ used to live here and

Then that's why * - " Heq_ used to live in Tangier , "

Thetyq_ used to look * - and some

of an elephant , wasq_ used to make a cake

Some separated lead-210 wasq_ used to make reference standards

crochet lace can beq_ used to make tablecloths , traycloths

provision which was nowq_ used to make the { 0T.E .

ancient Britons , I believe .q_ used to make water hot

as it is nowq_ used to mark a paragraph

Section the term wasq_ used to mean something like

Georgeq_ used to mix 100 stone of

junior to Humbert , whoq_ used to mock him affectionately

You'll have to getq_ used to my bad morning

gauge can now beq_ used to nick in the

three following winters wereq_ used to obtain an independent

Heq_ used to organise film shows

which can then beq_ used to perform an operation .

and devices to beq_ used to perform the various

Iq_ used to play about in

Iq_ used to play rugger , * * ' said

lead carrier solution isq_ used to prepare the reference

how Alexander the Greatq_ used to recline and transact

descriptions can also beq_ used to refer to performances

ratio decidendi } is normallyq_ used to refer to some

it may have beenq_ used to relate Christ's healing

migre * ? 2s , who notoriouslyq_ used to repair to the

she said chattily , Iq_ used to ride a bicycle .

and personality which journalistsq_ used to ridicule , can be

the gate the cockereiq_ used to run to meet

for you fellows , * * ' heq_ used to say . you can

Laughable , theyq_ used to say .

Heq_ used to say : ^ Have whatever

Of Kitchener heq_ used to say with humorous

reminiscent of what weq_ used to see pŽ®St .

seemed to have beenq_ used to seeing couples engaged

embarrassment if she isq_ used to seeing her mother

that force should beq_ used to settle this problem .

the May carol heq_ used to sing , with his

me the one sheq_ used to sing in Kimbolton

a shaped rubber isq_ used to smooth the hollow

was young schoolboy I used to sneak off to

Sheq_ used to solve all the
the clinical weekends he used to spend with her.

applied, and every means used to stop the train.
in contrasting tones were used to strengthen garments at
model which may be used to study both the
He used to stump round the
possibility of power being used to supplement hand tools.
Iq used to take the small
and colleague, Campbell Dixon, used to tell of a
The straight-edge can be used to test the straightness
is bought, can be used to the best advantage.
at (B). A mallet is
become (1) tired, or (2) more used to the disturbance.
Soho, to get me used to the food, he
might as well get used to the idea.
they very quickly get used to the idea of
She took time getting used to the indoor lavatories
They're used to the snatch racket.
that most people get used to them.
Jane was used to these sudden exigencies
or chieftain to get used to these trimmings because
to tinsel compliments, we used to think him unworldly.
in an Embassy - I used to think it was heavy, but one got used to this.
You are not yet used to this sort of
decorative kale are conveniently used to tone in with
horses; they had been used to trains since they
The brush contacts were used to trigger off a
He often used to try to imagine
His friends used to try to persuade
friend, William James, who used to urge that the
in London that Jones used to use in the
slaves * - everything he was used to using while he
a literary province I used to visit fairly often;
She used to walk straight to
He used to walk to the
page, would have been used to weigh bales of
They could be used to weigh several sacks
its simplest form it was used to work in the
they are a team used to working together, they
like that she had used to write to me.

Appendix B

ke a little getting used to -- not because it
iling teasing as he used to. <p> <s> Husky
from it that she had used to. <p> <s> You
little time to get used to. <s> After a
ur people have been used to accepting things as new melody is used to accompany his narrative.
repetitious The logical scheme used to accomplish the form.
brain cells inquiries were used to adjust compilation tent.
questions. I'm used to all three, but herese one hebephrenic man. 
age seven-iron shot he used to approach the green.
s> They could be used to approach the green. 
platform and can be used to automatically hold

citiz--uglier than you used to be, and you
ss glorious than it used to be, it is
istered here as it used to be, with unleavened
or less than itq used to be used to be ?? <p> <s>
eem high, but theyq used to be even higher " ,q,q,q,
spe said, This soilq used to be like that
ard roll. Quality. Thiq,q used to be part of
as e Catskills, whichq,q used to be the summer
that must have beenq used to calculate a tempera
feeli ransports that wereq used to carry Communist ageq,q,q,q.
the mails were thenq used to carry it out " ,<q,q,q,q,
tional codes can beq,q used to challenge and countq,q,q,q.
and Margaret recall,q,q used to characterize her asq,q,q,q.
of materials can beq,q used to construct a satisfaq,q,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,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.
eso 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 spreadsqu_q_q_q_ |
| ten 2 B filter wasq_q_ | used to filter off residualq_gq_ |
| 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_ |
| 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 |
| als as the cave-menq_q_ | used to have in the |

he governor was notq_q_ | used to having his integrit |

and had already becomeq_ | used to Hesperus ' snappinq_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 rently , 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_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 purposes -- 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_ |

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 upsteamq_q_q_ |

ratus will also beq_q_ | used to measure transitionq_q_q_q_q_e |

s steel screws wereq_q_ | used to minimize corrosioneq_q_q_e |

The DA statement isq_q_ | used to name and define |
The DC statement is used to name and enter sample; bio-assay methods used to obtain them.

tient of mine, who used to often seclude herself yesterday. You used to paint in them,

dly state funds were used to pay for the as a child I used to play ".

he corner where you used to play when you very summer. He used to play with the

out "surpluses had been used to provide a private ce ed aluminum plate, used to provide the drying

asq_Miss Giles always used to refer to her most of what we used to regard as the

tic up there, she used to say, with the of myewish intellectuals used to say. <p> se by instinct, he used to say: such places

The party that won used to say something about ma. The method used to scan the eye

S statement must be used to select the major stem. DIOCS is used to select the major b". It is used to separate two or

s> The symbol is used to separate two or foam and can be used to slit continuous she

me rand, IOCSIXF, used to specify the first I rand, IOCSIXG, used to specify the second

upstream frequency starter was used to start the arc. <
erb garden was also used to stop bleeding, and a lock, which is used to store cumulative reg

Throu was constructed and used to study transition prq was a trick they used to try and conceal

San Juan, but I 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.