THE ENGLISH MODAL AUXILIARY *MUST*: A CORPUS-BASED SYNTACTIC-SEMANTIC ACCOUNT

Leonardo Juliano RECSKI
Universidade Federal de Santa Catarina

RESUMO

Este artigo descreve um método para o estudo do verbo modal "must", que compara estruturas frasais complexas e usos deôntico e epistêmico em diferentes tipos de textos. Sugere-se que a distribuição de freqüência desses padrões gramaticais e semânticos em diferentes textos indique que tipos de ênfases pedagógicas possam facilitar o aprendizado de inglês como uma língua estrangeira.

ABSTRACT

In this article I describe a method to the study of the modal auxiliary verb "must" which compares complex verb-phrase structures and root/epistemic uses across different text types. It is suggested that the frequency distribution of syntactic and semantic patterns can indicate what kinds of pedagogical emphases are likely to facilitate EFL learning.

PALAVRAS-CHAVE

Deôntico, Epistêmico, Abordagem de corpus, Estruturas frasais complexas

KEY-WORDS

Root, Epistemic, Corpus-based approach, Complex verb-phrase structures

Introduction

Many articles and even books have been written on English modal auxiliary verbs. The complexity of meanings expressed by modal verbs has presented a challenge for linguists who have approached modals both in terms of semantic theory (Boyd and Thorn 1969; Halliday 1970; Marino 1973; Huddleston 1976; Johannessen 1976; Bolinger 1989; Groefsema 1992; Klinge 1993; Papafragou 1998) and in terms of descriptive grammar (for research based on corpus data, see Palmer 1979, 1986, 1990; Francis and Kucera 1982; Coates 1983; Collins 1991; Mindt 1997). The main difference between the former approaches in relation to the latter is that the former have relied on the linguists' intuitions whereas the latter on careful investigation of an extensive set of written and spoken texts.

It will not be claimed here that the study of an adequate corpus may provide definite answers to all of the linguists` questions. However, evidence derived from corpora provides an important check on intuitions and prevent linguists from relying on their own invented examples, and therefore from arriving at an unrealistic or biased view of the subject. Moreover, the use of corpus data means that all findings can be quantified, that computer programs can be run to help establish associations between semantics and syntactic categories, and that better qualitative results may be achieved.

A major reason for compiling linguistic corpora is to provide the basis for accurate and reliable descriptions of how languages are structured and used across different text types. In the past four decades, many linguists have opened up new and important grounds in terms of grammatical structure and semantic uses of modal verbs, yet no full-length account of modal verbs that uses corpus-based data and combines these two perspectives across different text types has been put forward.

In this paper, I shall be concerned with the modal auxiliary *must*. More specifically, I intend to provide a corpus-based syntactic-semantic account of *must* using a corpus of one million words of contemporary English which contains five different text types: Academic Writing, Science, Fiction, Business, and Spoken language.

Different analysts have provided insightful information on the semantic range covered by *must*. The problem with most accounts which only make use of semantic theories and are based on introspectively derived examples is that they imply a mental map of semantic terrain covered by *must*, with only brief acknowledgement from the analysts that they are aware of other dimensions being involved - for example, the proportion of root vs. epistemic uses of *must* across different text types and how these are distributed within complex verb-phrase structures (e.g. *must* + *be* + *past participle*).

In other words, accounts of *must* very rarely make explicit what sort of language they are talking about. Is *must* being considered in relation to written or spoken language - formal, informal, science, business, fiction? The bias in most analyses has been to the formal end of the spectrum.

In view of what has been stated in the above paragraphs the present study addresses the following research questions:

1. How is *must* distributed across the different text types?

2. How are root and epistemic uses of *must* distributed across these different text types?

3. How are root and epistemic uses of *must* distributed in relation to different complex verb-phrase structures (e.g., *must* + *infinitive, must* + *be* + *past participle, must* + *have* + *present participle,* etc)?

4. How might the differences in terms of frequency and usage of *must* across different text types suggest what kinds of pedagogical emphases are likely to facilitate language learning?

A corpus-based study on the modal auxiliary *must* may be used to inform how the distribution of its grammatical patterns and semantic roles may be related to different text types, and provide insightful information for the teaching and learning of its usage across different contexts.

1. The data

The data used in the present study is a corpus of one million words of written and spoken English that is divided as indicated in Table 1:

WRITTEN	TOKENS
Academic Writing	193.902
Business	199.586
Science	208.822
Fiction	201.367
SPOKEN	TOKENS
Workplace interactions	43.620
News Reportage	50.838
Press Briefings	52.828
Talk Shows	54.467

Table 1: Structure of the corpus

As seen in Table 1, the written portion or the corpus accounts for roughly 800.000 tokens (75% of the corpus) while the spoken portion accounts for only 200.000 tokens (25% of the corpus). The major reason for such discrepancy is accessibility. It is much easier to obtain written material than spoken material. I shall now give a brief description of each component of the corpus.

a) Academic Writing: a sample of the Louvain Corpus of English Essay Writing (LOCNESS). It comprises argumentative essays produced by British and American undergraduate students of different undergraduate courses;
b) Business: a sample of business sections from the *New York Times* collected between May and October of 1992.
c) Science: a sample of the *New Scientist*, which is a popular scientific magazine, collected between January and May of 1992.
d) Fiction: comprises a portion of about 20.000 words of the following books:

AUTHOR	BOOK
Charles Dickens	A Christmas Carol
Bram Stoker	Dracula
F. Scott Fitzgerald	This side of paradise
Theodore Dreiser	Sister Carrie
Virginia Woolf	Night and Day
Eleanor H. Porter	Just David
Washington Irving	The Legend Of Sleepy Hollow
Conan Doyle	The Hound of the Baskervilles
Jules Verne	20.000 Leagues Under the Sea
Mark Twain	The Private History of a Campaign That Failed

Table 2: Structure of the Fiction subcorpus

e) Workplace interactions: a sample of spoken American English which consists mainly of academic discussions such as faculty council meetings and committee meetings related to testing.
f) News reportage: a sample of transcripts from CNN's news programs such as *Insight, CNN World Report,* and *World News*. Aired between 2000 and 2002.
g) Press briefings: a sample of transcripts of White House press conferences, which are almost exclusively question-and-answer sessions. Aired between April 2000 and February 2002.
h) Talk Shows: a sample of CNN's interviews and debates programs like *Crossfire, Larry King Live, Late Edition*, and *The Point*. Aired between January and August of 2001.

The extent to which a corpus like the one I have compiled for this study can ever be considered to represent language in general is a matter of some contention. In practice, whether a finite sample of a language like this could ever represent the vast amount of language produced in even a single day is always likely to be, in the final analysis, an act of faith. Nevertheless, generalizations are an essential part of science, and even the great dictionaries and grammars of

English are all generalizations about language in this sense. Thus, it is hoped that this *home-tailored* corpus may serve my main goal: to offer a test bed to the investigation of modal usage across different text types and derive generalizations from it.

2. A brief description of *MUST*

Central modal auxiliary verbs like *must* have the following characteristics (Swan 1998:341):

(a) take negation directly (*mustn't*)
(b) take inversion without do (*must I?*)
(c) code (*John must study and so must Bill*)
(d) emphasis (*Ann MUST solve the problem*)
(e) no -s forms for third person singular (**musts*)
(f) no non-finite form (**to must, *musting,*)
(g) no co-occurrence (**must can*)

Must has two main meanings, epistemic (necessity - alternatively referred to in its different nuances as inference, certainty, conclusion, and so on), and root (obligation, compulsion, requirement and so on). An example of each extracted from the present corpus is given below:

Epistemic	**(1)**	*"You must be very proud of your family, Mss. Hibery."* (Fiction)
Root	**(2)**	*But the important thing is that the monarchy must continue.* (Spoken)

The epistemic meaning indicates the speaker's conviction or assumption about the truth of the proposition expressed, with *must* expressing a greater degree of certainty than *should* and *ought to*. According to Collins (1991:146) the epistemic category is syntactically distinctive: "negation affects the proposition rather than the modality, past tense forms are rare, and it co-occurs with the perfect and progressive aspects". By contrast with epistemic meaning, root meaning is somewhat an indeterminate category. Palmer (1990) recognizes, and attempts to handle the range of root meanings by creating two sub-categories: *deontic*, where the speaker is generally the source of the obligation, and *dynamic*, where the speaker is not the source of the obligation. Nonetheless, Palmer (1990:91) admits that "there is no clear dividing line between the two meanings", but claims that the distinction facilitates description of the relationship between *must* (which he claims may be either deontic or dynamic) and *have (got) to* (which he claims may be only dynamic). In fact, Coates (1983) asserts that the reason she has treated root *must* as one category is due to the uncertainty of the presence or absence of the speakers' involvement in the utterance.

When expressing the speakers' convictions (which may range from strong to weak), epistemic *must* is often found in structures like *must + have + past participle and must + be + past participle*. Some examples from the present corpus follow:

(3) *Because they are similar in widely different organisms, biologists believe they must have arisen very early in evolution. But …* (Science)
(4) *A: Then at noon he'll go to the restaurant Filomena's with Chancellor Kohl. They discuss…*
B: He loves that place, doesn't he … Kohl.
A: He … I think both of them. It must be a sumo wrestling hangout. (Spoken)

In (3) the grounds upon which the deduction is made are specified ('Because …'). In (4) the epistemic nature of must is reinforced by the hedge *I think*.

In many instances epistemic *must* refers to states or activities in the present, as in (3) and (4). Epistemic *must* exhibits a strong tendency to occur with the perfect aspect, as in (6) below, with stative verbs, as in (3) and (4) above, and with the progressive aspect, as in (5).

(5) *The Panel must be hoping that this will be an end to the matter.* (Business)
(6) *Physicians, nurses, and others are often witnesses to death. People who go into these fields must have had to deal with this issue in their training.* (Academic)

With respect to root *must*, the data in the present corpus corroborates the 100 per cent association of root *must* with negation found in Coates's (1983) study. The association of first and second person subjects and root meaning is also very strong in the present data, especially in the spoken portion of the corpus because of the number of occurrences of *I must say*, and *I must admit* which are used as elements of discourse orientation where the speaker imposes the obligation on himself and by doing so actually performs the act (*I must admit/say = I do admit/say*) as exemplified in (7) and (8).

(7) *Wolf, I must say I'm darkly suspicious of anything that Senator Kennedy embraces and that we adopt.* (Spoken)
(8) *My feeling, I must admit, without being part of this polo world myself, is that it's a long tradition [...]* (Spoken)

One of the most common uses of root *must* is related to the speaker's point of view. In affirmative sentences, we can use *must* to say what is necessary and to give strong advice or orders to ourselves or other people. This is especially common in British English; in American English *have to* is generally preferred, particularly in speech (cf.: Swan 1998). Some examples from the present corpus follow:

(9) *He told the BBC that after the details began, he cabled Sharon telling him, you must stop the slaughter, this situation is absolutely appalling.* (Spoken)
(10) *You do not necessarily have to be a complete nihilist, but you must be aware that it is you that is making the choice of what action to take.* (Spoken)
(11) *Ecological bricks must not be an excuse for an ecological annexation, or greens will be exacerbating rather than defusing conflict.* (Science)

In questions, we can ask what the hearer thinks it is necessary or the reason for something:

(12) *Must I go instead into the streets to save the homeless man and rescue the battered women?* (Academic)
(13) *Why must you always leave things spread about on deck?* (Fiction)

I have provided only a brisk overview of semantic and grammatical patterns associated with *must*. Nevertheless, while reviewing the literature and having access to a corpus it was possible to test theories. Being a source of language in use, corpora allow for the investigation of theories through factual examples, the analyst starting from a hypothesis based on the literature and using the corpus to test it. More importantly, though, is the fact that corpora may *require* from researchers that they formulate their own theories, especially when they are confronted with examples they fail to find in any theory.

3. Computer tools

The following section describes the software, tools and utilities used in this research paper.

a) TOSCA-ICLE tokenizer, lemmatizer and tagger:

The program suite for corpus tagging was developed by the TOSCA team at Katholieke Universiteit Nijmegen, The Netherlands. The main tagging tools consist of several programs that run automatically when a corpus is being tagged:

1. The **tokenizer**, which decides what should be considered a word and what not, recognizes punctuation and puts sentence boundaries in the texts;
2. The **lemmatizer**, which provides lemmas for every item in the corpus (except for markup). The lemmatizer looks up each

word in its built-in lexicon database; if the word does not exist there, the lemmatizer tries to derive the base form from the given word according to its pre-programmed rules (cf.: de Hann & van Halteren 1997). See Table 1.3 below for examples:

Word	**Lemma**	**Word**	**Lemma**
students	student	taught	teach
were	be	saw	see
?	&quest;	saw	saw

Table 3: Items and their lemmas

3. The **tagger**, which attributes every token with a part-of-speech tag, including wordclass membership and additional feature information (the word's morphological or syntactic-semantic characteristics). The tagset is based on Quirk et al. (1985).

The success of tagging claimed by the designers of the software is 95%, i.e. about 95% of words fully correspond with their tags. The remaining 5% represent words with inappropriate tags. As a means to illustrate how the tagger works, I have selected a sentence of an essay of the LOCNESS corpus:

Original format

```
Cheating has become a major question of value to the present
student; unfortunately, the consequences that should stop stu-
dents from cheating are unsuccessful.
```

The analyst can then decide how he/she wants the sentence to be tagged. There are several options available. Following is an example of the raw sentence above tagged including words, lemmas, full tags, punctuation, and sentence boundaries:

```
<sent1>
Cheating_cheating_N(sing)has_have_VB(aux,perf,pres)
become_become_VB(lex,cop,edp)a_a_ART(indef)major_major_ADJ
(ge,pos)question_question_N(sing) of_of_PREP(ge)value_value_N
```

```
(sing)to_to_PREP(ge)the_the_ART(def) present_present_ADJ
(ge,pos)student_student_N(sing);_&semi;_PUNC(scolon)
unfortunately_unfortunately_ADV(ge,pos),_&comma;_PUNC(comma)
the_the_ART(def)consequences_consequence_N(plu)that_that_PRON
(rel)should_shall_VB(aux,modal,past)stop_stop_VB(lex,montr,
infin)students_student_N(plu)from_from_PREP(ge)
cheating_cheat_VB(lex,montr,ingp)are_be_VB(lex,cop,pres)
unsuccessful_?unsuccessful?_ADJ(ge,pos)._&period;_PUNC(per)
<sent2>
```

Alternatively, the linguist may wish to tag the sentence including only lemmas and tags and omitting words. The output is shown below:

```
<sent1>
cheating_N(sing)have_VB(aux,perf,pres)become_VB(lex,cop,edp)
a_ART(indef)major_ADJ(ge,pos)question_N(sing)of_PREP(ge)
value_N(sing)to_PREP(ge)the_ART(def) present_ADJ(ge,pos)
student_N(sing) &semi;_PUNC(scolon)unfortunately_ADV(ge,pos)
&comma;_PUNC(comma)the_ART(def)consequence_N(plu) that_PRON
(rel)shall_VB(aux,modal,past)stop_VB(lex,montr,infin)
student_N(plu)from_PREP(ge)cheat_VB(lex,montr,ingp)be_VB
(lex,cop,pres)?unsuccessful?_ADJ(ge,pos)&period;_PUNC(per)
<sent2>
```

Let's suppose now that the linguist is only interested in the overall frequency of different word classes in this sentence. This can be easily achieved by omitting words, lemmas, and sentence boundaries. Thus, the original raw sentence would look like this:

```
N VB VB ART ADJ N PREP N PREP ART ADJ N PUNC ADV PUNC ART N
PRON VB VB N PREP VB VB ADJ PUNC
```

As shown above, the output files can contain the sentence in any combination of words, lemmas and tags, the possible formats being:

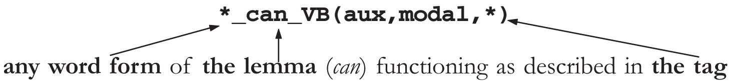
WORD_LEMMA_TAG	WORD_TAG
WORD_LEMMA	LEMMA_TAG
LEMMA	TAG

b) Concord and WordList (as part of the WordSmith Tools)

Wordsmith Tools is a relatively small, but undoubtedly useful, piece of software running on a personal computer. The programs in *WordSmith Tools* can handle virtually unlimited amounts of text. They can read text from CD-ROMs, so giving access to corpora containing many millions of words. The main advantage of *Wordsmith Tools* is that it displays the output directly on the screen. The output can also be saved as a file and printed out. *Wordsmith Tools* can be used not only on plain English texts, but also on texts in other languages, and on English texts with grammatical encoding. The functions of the *WordSmith Tools* include frequency listing, alphabetical listing, keyword in context (KWIC) analysis, further searching on both sides of the keywords, and closer investigation of the target items in larger contexts.

Concord can be used to search a collection of texts and display all the instances of a chosen word alongside its context.

Let's suppose that we want to retrieve all the instances of the lemma *can*. By having a corpus tagged in the WORD_LEMMA_TAG format this can be easily done using *Concord* and a search string as follows:



```
*_can_VB(aux,modal,*)
```

any word form of **the lemma** (*can*) functioning as described in **the tag**

This search will produce results with the original forms *can, can't, cannot, could, couldn't* functioning as auxiliary modal verbs.

WordList is an extremely useful tool when it comes to counting overall frequencies of chosen items in a corpus. When run on a raw text, *WordList* will generate alphabetical and frequency lists of all words appearing in the corpus, with their total and relative frequencies. Moreover, it can also make lists of n-word combinations.

When using a tagged corpus, the work with *WordList* requires a completely different approach than *Concord*. While in concordancing it is not a problem to search for various combinations of words, lemmas, and tags in the WORD_LEMMA_TAG format, *WordList* requires a differently designed format corpus.

As we are searching a tagged corpus, we need *WordList* to understand all the WORD_LEMMA_TAG units as single entities, that is, as single words. First we must inform *WordList* that the words are going to be long - or it would disregard them. This can be done in the menu *Settings / Text characteristics / Word length*, which is set 1 to 50. Also we must tell *WordList* that the punctuation marks, which appear within these characters, need to be treated as normal letters, and not as word separators. This can be done in *WordList* menu *Settings / Text characteristics / Handling / Characters within words*. In the empty field, we have to type the following characters:

_ , ; . ( ) / ? & ! - > < " '

After the above procedures, it is possible to retrieve all modal auxiliary frequencies in a matter of seconds using a corpus in the LEMMA_TAG format. The output of such a search is seen in Figure 1 where the frequencies were count for lemmas, i.e. the numbers in the *can* lines, for example, stand for the occurrences of *can, can't, cannot, could*, and *couldn't* and their respective frequency percentages are related to total number of words of this subcorpus (percentages under 0.01 are not shown by default).

```
N  Word                                 Freq.     %  Lemmas
1  CAN_VB(AUX,MODAL,PAST)                  52  0,08
2  CAN_VB(AUX,MODAL,PAST,NEG)              13  0,02
3  CAN_VB(AUX,MODAL,PRES)                 117  0,19
4  CAN_VB(AUX,MODAL,PRES,NEG)              38  0,06
5  MAY_VB(AUX,MODAL,PAST)                  20  0,03
6  MAY_VB(AUX,MODAL,PRES)                  26  0,04
7  SHALL_VB(AUX,MODAL,PAST)                68  0,11
8  SHALL_VB(AUX,MODAL,PAST,NEG)             9  0,01
9  SHALL_VB(AUX,MODAL,PRES)                 2
10 WILL_VB(AUX,MODAL,PAST)                151  0,25
11 WILL_VB(AUX,MODAL,PAST,ENCL)            12  0,02
12 WILL_VB(AUX,MODAL,PAST,NEG)             15  0,02
13 WILL_VB(AUX,MODAL,PRES)                161  0,26
14 WILL_VB(AUX,MODAL,PRES,ENCL)            71  0,12

15 WILL_VB(AUX,MODAL,PRES,NEG)              5
```

Figure 1: Modal frequency list in the *Talk Shows* subcorpus

4. Discussion of the results

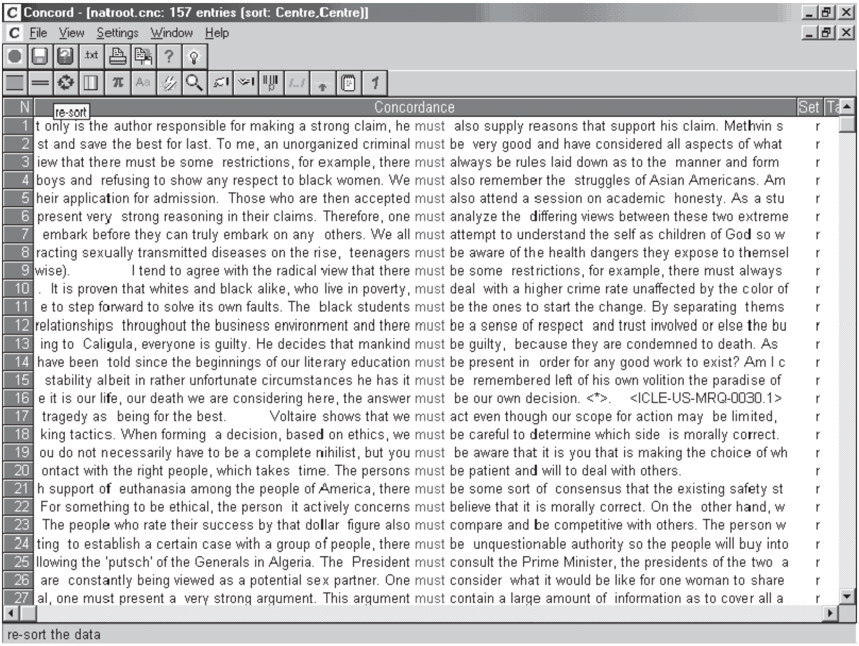
By using *WordList* and the corpus tagged in the LEMMA_TAG format it was possible to quickly map out the frequency distribution of *must* across the different text types in the corpus. Table 4 shows the vast differences which there are across the subcorpuses:

Modal	Academic	Spoken	Business	Fiction	Science
must	175	51	67	208	152

Table 4: Comparison of gross frequencies

The modal *must* displayed similar frequency distributions in the Academic Writing, Fiction and Science subcorpuses but much lower frequencies in the Spoken and Business subcorpuses. This frequency distribution seems to indicate that *must* occurs more frequently in formal text types and less frequently in more informal text types as exemplified by the frequency distribution of *must* tokens in Spoken English. The frequency distribution provided in Table 4 is, nevertheless, inevitably crude: it ignores factors influencing semantic and grammatical choices regarding the modal *must*. For example, *must* may be used in a root or in an epistemic way, but we are not told in what proportions or whether these proportions remain constant if different text types with different degrees of formality are taken into account. In addition to this, we have no information as to how the two meanings *must* expresses can be related to complex verb-phrase structures (*e.g. must + infinitive or must + have + past participle*).

In order to account for these insufficiencies, first it was necessary to map out the frequency distribution of root and epistemic meanings uses of *must* across the different text types. This had to be done manually, analyzing each occurrence of the modal *must* across each text type. The method employed can be seen in Figure 2, where the analyst inserts the desired code (in this example *r* = root) under the column *Set* to the right of each concordance line. After each concordance line has been assigned a code (*r* = root, and *e* = epistemic), the analyst can then re-sort the concordance lines by *Set*, thus, retrieving individual occurrences of root and epistemic uses.



```
 N  Concordance                                                                                                                  Set Ta
 1  t only is the author responsible for making a strong claim, he must  also supply reasons that support his claim. Methvin s     r
 2   st and save the best for last. To me, an unorganized criminal must be  very good and have considered all aspects of what     r
 3  iew that there must be some  restrictions, for example, there must always be rules laid down as to the  manner and form        r
 4  boys and  refusing to show any respect to black women. We must also remember the  struggles of Asian Americans. Am           r
 5  heir application for admission.  Those who are then accepted must also attend a session on academic  honesty. As a stu        r
 6   present very  strong reasoning in their claims. Therefore, one must analyze the  differing views between these two extreme   r
 7    embark before they can truly embark on any  others. We all must attempt to understand the self as children of God so w     r
 8  racting sexually transmitted diseases on the rise,  teenagers must be aware of the health dangers they expose to themsel      r
 9  wise).            I tend to agree with the radical view that there must be some  restrictions, for example, there must always   r
10  .  It is proven that whites and black alike, who live in poverty, must deal  with a higher crime rate unaffected by the color of r
11   e to step forward to solve its own faults. The  black students must be the ones to start the change. By separating  thems    r
12  relationships  throughout the business environment and there must be a sense of respect  and trust involved or else the bu   r
13   ing to  Caligula, everyone is guilty. He decides that mankind must be guilty,  because they are condemned to death. As       r
14  have been  told since the beginnings of our literary education must be present in  order for any good work to exist? Am I c  r
15    stability albeit in rather unfortunate circumstances he has it must be  remembered left of his own volition the paradise of r
16   e it is our life, our death we are considering here, the answer must  be our own decision. <*>.    <ICLE-US-MRQ-0030.1>     r
17    tragedy as  being for the best.          Voltaire shows that we must act even though our scope for action may  be limited,  r
18    king tactics. When forming  a decision, based on ethics, we must be careful to determine which side  is morally correct.   r
19    ou do not necessarily have to be a complete nihilist, but you must  be aware that it is you that is making the choice of wh r
20    ontact with the right people, which takes  time. The persons must be patient and will to deal with others.                r
21   h support of  euthanasia among the people of America, there must be some sort of  consensus that the existing safety st    r
22    For something to be ethical, the person  it actively concerns must believe that it is morally correct. On the  other hand, w r
23    The people who rate their success by that dollar  figure also must compare and be competitive with others. The person w   r
24   ting  to establish a certain case with a group of people, there must be  unquestionable authority so the people will buy into r
25  llowing the 'putsch' of the Generals in Algeria. The  President must consult the Prime Minister, the presidents of the two  a  r
26    are  constantly being viewed as a potential sex partner. One must consider  what it would be like for one woman to share   r
27  al, one must present a  very strong argument. This argument must contain a large amount of  information as to cover all a    r
```

Figure 2: Concordance lines for *must* in its root use re-sorted by *Set*

Text Type	Root	Epistemic
Academic Writing	87	13
Business	58	42
Fiction	65	35
Science	76	24
Spoken	67	33

Table 5: Distribution of root and epistemic uses of *must* (%)

As seen in Table 5, the root meaning (i.e., obligation) appeared in most text types as the core or primary meaning displayed by *must* (the exception being the *Business* subcorpus where root and epistemic meanings were very similar in terms of overall frequencies). In fact, Coates (1983) in her study of modal auxiliary verbs, found that in both the Lancaster-Oslo/Bergen (LOB) corpus and the corpus of the Survey of English Usage, root *must* accounted for the majority of cases regarding this modal. Coates (1983:33) points out that in most root uses of *must* it is possible to distinguish the following features:

1. Main verb is activity verb
2. Speaker is interested in getting subject to perform the action
3. Speaker has authority over the subject

In order to illustrate such features I have extracted a few examples from the present corpus:

(14) "Hush! You must answer their questions, "Katherine whispered, desiring, at all costs, to keep him quiet. (Fiction)
(15) European patent application 455 518 from Polymeters Response International of Winchester typifies the dilemma faced by inventors with new ideas for curbing fraud. If they file a patent application, they must describe the fraud before outlining the cure. (Science)
(16) Walker shows African American that they must find a self-understanding that is void of the standards that were and are imposed on them by the white race. (Native)

In its most common usage, epistemic *must* conveys the speaker's confidence in the truth of what he/she is saying, based on a logical process or deduction from facts known to him/her. Typical examples from the present corpus are:

(17) Our Sun is a latecomer in the Galaxy as a whole so, if other civilizations are common, there must be many that are older than ours, and could undoubtedly master the problems of interstellar travel. (Science)
(18) I think it must take a person with special attitude to accept this ever-present issue. (Academic)

Corpus-based research can also show that there are differences in the use of *must* across the different text types in the kinds of complex verb-phrase structures it occurs. Table 6 lists the major verb structures and the extent to which *must* is used in those structures across the different text types.

Modal structure	Academic	Business	Fiction	Science	Spoken
MUST alone			2	1	4
MUST + infinitive	69	69	74	72	76
MUST + *be* + past participle	27	13	10	18	10
MUST + *be* + present participle	1	15	1	1	2
MUST + *have* + past participle	2	2	10	7	6
MUST + *have* + *been* + past participle	1	1	2	1	2
MUST + *have* + *been* + present participle			1		

Table 6: Use of *must* in various verb-phrase structures (%)

As shown in Table 6, over 68% of the tokens of *must* across all text types occurred in the *must* + *infinitive* structure, which indicates that this is the pivotal or core grammatical pattern for *must*. The second most common verb-phrase structure was *must* + *be* + *past participle* with at least 10% of its tokens. Rather tellingly, from a pedagogical viewpoint, five of the seven possible structures have few or no tokens at all (the exceptions being *must* + *be* + *present participle* in the Business subcorpus and *must* + *have* + *past participle* in the Fiction subcorpus). However, even though Table 6 lists all major verb-phrase structures of *must* we are not told in which proportion root and epistemic uses of this modal are distributed across these structures. This figure is given in Table 7 below:

Modal structure	**Academic**		**Business**		**Fiction**		**Science**		**Spoken**	
	root	epist.	root	epist.	root	epist.	root	epist	root	epist.
MUST alone	0	0	0	0	100	0	0	100	0	100
MUST + infinitive	89	11	89	11	72	28	83	17	70	30
MUST + *be* + past participle	94	6	67	33	95	5	86	14	100	0
MUST + *be* + present participle	100	0	0	100	0	100	0	100	100	0
MUST + *have* + past participle	0	100	0	100	0	100	0	100	0	100
MUST + *have* + *been* + past participle	0	100	0	100	0	100	0	100	0	100
MUST + *have* + *been* + present participle	0	0	0	0	0	100	0	0	0	0

Table 7: Root and epistemic uses of *must* across the verb-phrase structures (%)

Many important conclusions may be drawn from Table 7. As regards *must + infinitive* verb structures (72% of the tokens of *must*), in all of the text types at least 70% of this structure was related to root *must*, which indicates a strong association between infinitives and root meanings for this particular modal. One plausible interpretation for this correlation between root *must + infinitive* may be due to the fact that root modals are normally performative (Palmer 1986:98), that is, the speaker clearly takes the responsibility for imposing the necessity, and since this is usually expressed at the moment of speaking, it requires present tense. The important point here, however, is that the speaker is in a position to lay the obligation, and is thus in a position of some authority.

In relation to *must + be + past participle* structures (17% of the tokens of *must*), with the exception of the Business subcorpus, in all of the other text types over 85% of the tokens of *must* in this particular verb structure were related to root meanings. In this particular verb structure, the speaker or writer usually takes responsibility for the imposing of the necessity for future actions (19), or reports what someone/something is required to do (20). The context makes it clear in:

(19) Once it has been established that this being is indeed human and living, the next step is to prove that it is a separate individual entity from the woman carrying it. This must be done in order to refute any argument that pro-choice

activists have stating that this unborn human life is part of its mother's body and, therefore, be subject to whatever decision the woman decides to make in regard to her own body. (Academic)

(20) This particular point has been criticized by a group of legal, biomedical and psychology specialists, who say children conceived from donated eggs or sperm must be allowed to trace the identity of the donor. (Science)

As for the verb structure *must + be + present participle* which accounted for only 2% of the tokens of the modal *must*, the Academic and Spoken subcorpuses (one occurrence each) displayed a 100% association between this verb structure and root meanings. The other three subcorpuses, Business (9 occurrences), Fiction (1 occurrence) and Science (1 occurrence), displayed a totally different picture with 100% association between epistemic meanings and this particular verb structure. The limited number of occurrences for this particular complex verb-phrase structure does not allow us to provide a clear picture as to why the Spoken and Academic subcorpuses were related to root usage whereas the other subcorpuses were related to epistemic usage. Many more occurrences of this particular complex verb-phrase structure across different text types would have to be investigated in order to provide a better picture as to why this structure is used with different meaning across different texts.

(21) I thought at the time that I must be dreaming when I saw them, they threw no shadow on the floor. (Fiction - Epistemic)

(22) Many shareholders must be wondering whether there is any point in holding on any longer. (Business - Epistemic)

(23) He found that the correct equations insisted that black holes must be emitting radiation - that they had a temperature, just like Bekenstein had said. (Science - Epistemic)

(24) He was willing to admit that some things were wrong in the Soviet Union, which was very unusual [...] and he was trying to say to people, look you must be enterprising. (Spoken - Root)

(25) Tolerance is not enough. Tolerance denies understanding. Tolerance ignores the ethic American experience. We must be willing to accept other people's ideas and beliefs about things. (Academic - Root)

Other two verb-phrase structures, *must + have + past participle*, and *must + have + been + past participle*, which accounted for 8% of the tokens of *must*, showed a 100% association with epistemic meaning throughout the subcorpuses, which, in turn, indicates that the reasons for such conclusions are implied from past evidence. In these cases, the propositions are in the past, and this is achieved by using *have* or *have been* before the main verb. Notice that the two examples below refer to past judgments:

(26) A: Kathleen, thanks. Kathleen Koch working the air crash story tonight. B: While we were listening to that, we thought in the calm of the air traffic controller's voice, what it must have been felt up there, when they had lost a plane that day. (Spoken)
(27) If this is so, the magnetic field must have played an important role in the evolution of the Universe. (Science)

The verb-phrase *must + have + been + present participle*, occurred only once in the Fiction subcorpus, which indicates the rarity of such structure (at least in my corpus):

(28) Towards morning I slept and was awakened by the continuous knocking at my door, so I guess I must have been sleeping soundly then. (Fiction)

Finally, *must* very seldom occurred alone. The examples extracted from the corpus indicate that in this particular structure *must* can have both, root and epistemic meanings.

(29) 'Man of the worldly mind!' replied the Ghost, 'do you believe in me or not?' 'I do,' said Scrooge. 'I must.' (Fiction)

(30) A: …I don't know if that is right or fair, but I know it's true and he must too, and imagine carrying that around these days.

6. Concluding remarks

A total of 653 occurrences of the modal auxiliary verb *must* were found in the present corpus. It was found that more formal text types such as Fiction, Academic Writing, and Science, in which the degree of subjectivity of propositions tends to be avoided for the sake of clarity, authority, and credibility, display more tokens of *must* if compared to less formal ones such as Business and Spoken language, which allow for a greater degree of subjective propositions. Such difference is likely to stem from the fact that *must* is primarily used as a root modal, that is, it is used to express obligation and/or necessity (objective), and it is peripherally used epistemically, that is to say, as an inference, certainty or conclusion (subjective).

From a pedagogical viewpoint, the distribution of frequencies provided here might be helpful for both EFL teachers and students. It may be used to indicate which meanings of the modal auxiliary *must* are most commonly employed and the relationship between these meanings and the grammatical structures in which they can be found. In this way, the findings of the present rendering indicate that the EFL teacher should start his/her description of *must* by its root meaning indicating that the two most common grammatical structures to appear with this meaning are *must + infinitive* and *must + be + past participle*. At this point, the teacher can also inform his/her students that the more formal the kinds of texts they are writing or uttering the more likely they are to use root *must* (with *infinitive* and *be + past participle*) because this will lend support, authority, and credibility to their propositions. This can be seen in Table 7 where we find that epistemic *must* is seldom used in the *be + past participle* structure. Out of the 72 occurrences of the structure *must + be + past perfect*, 54 occurred in academic writing, 51 occurrences (94%) being related to root uses. An important characteristic of academic writing, specially of argumentative essays, then emerges: writers tend to *urge* for actions to be taken by using *must* + *be* + *past participle* in sentences like *These problems, […], must be resolved, the decision must be made that[…], to enforce these rules athletes must be tested, European Community Law must be abided by, Community legislation must be applied uniformly, a balance must be struck*.

As for epistemic modality, which is an important peripheral use of *must*, the results reveal that subjective propositions (inferences) display a 100% correlation with past tenses - *must have* (Palmer 1986:50; Coates 1983:42). In addition, epistemic *must* may also appear followed by an *infinitive*, *be* + *past participle*, or *be* + *present participle* structures. The modality expressed in these structures is in the present, because the judgments are made in the act of speaking, *must* being in this sense usually performative: *I think you must be very clever, they must be bad-minded, she imagined, black-holes must be emitting radiation*. Altogether, these three structures accounted for 68% of its epistemic use, which indicates the predominance of epistemic *must* referring to a present state, and being commonly used with the verb *be*.

This paper has provided an account of the English modal auxiliary *must*, based on a corpus of spoken and written language. Unlike the majority of previous studies, the present rendering admits to being about more than semantics. With the modal auxiliary *must*, as a point of departure, a framework was formulated to shed light on the interrelation of both semantics and grammatical patterns involving *must*. Personally, I believe that the main benefits of adopting a corpus-based technique to the investigation of the modal verb *must* have been that

a) it has enabled qualitative statements to be made on the distribution of forms and meanings of *must* across a variety of text types;

b) the use of *authentic* data has provided a safeguard against the danger of biases present in studies based on introspectively derived examples;

c) all tokens in the corpus, no matter how unruly, had to be accounted for, which has called for a model capable of handling the range and complexity of meanings the modal *must* may have.

My aim in this paper has been to interpret the data rather than to impose a preconceived system upon it. The framework illustrated here amounts to a re-orientation of research into the English modals. It has been suggested that a syntactic-semantic approach to modal verbs carried out with the overall severity and rigor of corpus-based research may help EFL researchers, teachers, and students achieve a better understanding of the interrelations of the numerous syntactic and semantic patterns modal auxiliary verbs are known to have.

Acknowledgements: I am very grateful to two anonymous referees for their pertinent comments on an earlier draft of this paper. This paper has also benefited greatly from the expertise of Viviane Heberle and Apóstolo Nicolacópulos, whose comments and suggestions were very valuable.

Recebido em abril de 2002. Versão revisada aceita em

References

Bolinger, D. (1989) Extrinsic Possibility and Intrinsic Potentiality: 7 On May and Can + 1. *Journal of Pragmatics* **13**, 1-23.

Boyd, J. and Thorne, J. P. (1969) The Semantics of Modal Verbs. *Journal of Linguistics*, **5**, 57-74.

Coates, J. (1983) *The Semantics of Modal Auxiliaries*, Croom Helm.

Collins, P. (1991) *The Modals of Obligation and necessity in Australian English*, In Johansson & Stenström (1991) 181-200.

de Haan, P. and van Halteren, H. (1997) *The TOSCA-ICLE Tagset – Tagset Manual*. Nijmegen: The TOSCA research Group for Corpus Linguistics.

Francis, W. N. and Kucera, H. (1982) *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

Groefsema, M. (1992) Can, may, must and should: A Relevance theoretic account. *Journal of Linguistics* **24**, 53-79.

Halliday, M. A. K. (1970) Functional Diversity in Language as Seen From a Consideration of Modality and Mood in English. *Foundations of Language,* **4**, 225-42.

Huddleston, R. (1976) Some Theoretical Issues in the Description of the English Verb. *Lingua*, **40**, 331-83.

Johannessen, N-L. (1976) *The English Modal Auxiliaries: A Stratificational Account*, In Coates, J. (1983) *The Semantics of Modal Auxiliaries*, Croom Helm.

Klinge, A. (1993) The English Modal Auxiliaries: from lexical semantics to utterance interpretation. *Journal of Linguistics*, **29**, 315-57.

Marino, M. (1973) A Feature Analysis of the English Modals. *Lingua*, **32**, 309-23.

Mindt, D. (1995) *An Empirical Grammar of the English Verb: Modal Verbs*. Berlin: Cornelsen.

Palmer, F. R. (1979) *Modality and the English Modals*, London and New York: Longman

_____. (1986) *Mood and Modality,* Cambridge University Press.

_____. (1990) *Modality and the English Modals, 2nd Edition*. London: Longman

Papafragou, A. (1997) Inference and word meaning: The case of modal auxiliaries. *Lingua,* **105**, 1-47.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985) *A Comprehensive Grammar of the English Language* London: Longman.

Swan, M. (1998) *Practical English Usage*. Oxford University Press.