*John M. Kirk**

Using VARBRUL for Studying Modal Auxiliary Verbs?

The question to be addressed by this paper is whether the statistical package VARBRUL, much loved by sociolinguists in North America for the measurement of effect on variation of different internal and external variables, can be used for the study of modal auxiliary verbs in English. My overall conclusion is both YES and NO! Despite Thibault's (1991) pioneering study of *devoir* in French, not all of VARBRUL can be used for studying modal auxiliary verbs in English. And the reason is this: VARBRUL operates on the basis of 'variables'; for modal verbs, the variable would be such semantic concepts as 'obligation', 'possibility', 'prediction', 'permission', etc. of which individual modal verbs would be the exponents. The crucial assumption is that each of these exponents would be semantically 'equivalent' or 'identical'. And this is the assumption that I don't think can be made.

* *John M. Kirk*
School of English
The Queen's University of Belfast
Belfast BT7 1NN Northern Ireland

Figure 1 (adapted from Leech 1989) shows a list of possible variables in the column entitled 'meanings', and alongside each one is an indication of the likely exponents. So the assumption is that for the concept 'Obligation', the exponents (i.e. the variants for the purposes of VARBRUL) would be *must, should, ought (to), have to, have got to,* and *need (to)* or for the concept of '(future) Prediction' the exponents would be *will, would* and *shall,* as well as *BE going to* and the present simple and the present progressive tense-aspect configurations. The use of VARBRUL assumes that these are equivalent, free choices, and VARBRUL enables us to identify the internal, linguistic environment - as well as the external, social, textual, historical or other external context - in which each of these variant choices are made. Using VARBRUL, we could choose whatever features or factors we feel might be causing or influencing the choice of one of these variants over another, and VARBRUL ends up telling us whether we are right or not, and to what degree. So such internal environmental factors might be: polarity (whether the verb is negated or not); clause type (whether main or subordinate, etc.); subject number and person; type of following lexical verb (whether stative or dynamic); form of following lexical verb (whether marked for aspect or voice); existence of modally harmonic adverbials; and the basic modal meaning (whether it is root or epistemic). And such external factors might be: region; medium (spoken or written); type of text or register within spoken and written medium; sex, age, and other social characteristics of speaker (where relevant). And we could code these in using simple single-symbol mnemonics (cf. Kirk 1994b). VARBRUL would then check your encodings for all these factors and tell you which ones are significant and therefore influencing the variation. So it's a good deal: you tell it what you think, and it'll tell you whether you're right, and by how much. Categorical findings of 0% or 100% are known as 'knockouts', and are knocked out as there is, obviously, no variation. VARBRUL is really only useful where there is variation.

So how does VARBRUL do it? It works on the basis of probabilities and the principle of 'maximum likelihood' for estimating the occurrence of the factors specified. It looks at all the factors together and calculates their probabilities, and then it considers the data in terms of each individual factor in comparison with all the others to see which factors are genuinely significant. It throws up mean overall figures for each

factor, and individual factor scores higher than the mean are then considered ‘significant’. This is known as multi-dimensional, regressional analysis - ‘multi-dimensional’ because all the factors are taken into account; ‘regressional’ because it deals with each factor in turn, step by step. One of the ways it expresses its results is through scattergrams, as in Figure 2, reproduced simply for illustrative purposes [the actual content of these scattergrams is rubbish!]. What does a scattergram tell us? Scores close to the diagonal show that actual occurrences

are close to the calculated probabilities, so that, for the data represented by the squares, there are no significant factors; for scores way off the diagonal, there are significant factors, because the actual behaviour is at odds with the probabilistic behaviour. The on-line version of VARBRUL allows you to click on any square, identify the data underlying it, and discover what the significant factor actually is. (Cf. Kirk 1994a)

So, for instance, Tottie (1988) was able to do this for comparing her two types of negation: *no*-negation and *not*-negation, but without using the scattergram facility. VARBRUL worked for her, because semantic

equivalence could be assumed. Likewise, Nevalainen (1991) analysed the behaviour of a single item (*only* or *but)* in different contexts at different periods. Although the functions changed, semantic equivalence could always be assumed. Likewise, too, in the many phonological studies, for which VARBRUL was originally devised.

For modal verbs, however, I fear that semantic equivalence is too big a price too pay for any usefulness from VARBRUL scattergrams which might follow. Evidence against semantic equivalence is set out in the following examples:

==========

MUST = SHOULD = OUGHT (TO): ROOT MEANINGS ('OBLIGATION')

1a I must stay in and write letters this evening.
1b I should stay in and write letters this evening.

2a *I must stay in - but I'm going out.
2b I should stay in - but I'm going out.

3a You must come to dinner with us! (polite invitation)
3b You should come to dinner with us! (rude)

4 The section on MUST in the *COBUILD English Usage* is excellent.
You ..... read it.
want to
should
ought to
must
have to
-'ve got to

MUST = HAVE (GOT) TO: ROOT MEANINGS ('OBLIGATION')

HAVE (GOT) TO = somebody else or some external circumstances decided action is necessary; MUST = speaker has decided action is necessary

4a I have (got) to get a new passport.
4b ?I must get a new passport.

5a Why do you have to? / Why have you got to?
5b ?Why must you?

6a Do you always have to start work at 8.00 a.m.?
6b Must you always start work at 8.00 a.m.?

7a People who qualify must apply within six months.
7b *People who qualify have to apply within six months.

7c This firedoor must be kept unlocked during working hours.
7d I wonder why that door has to be kept unlocked.

(= ‘What is the point of the rule?’)

8a In my opinion, children must be treated firmly.

8b At some schools, children have to obey all sorts of silly rules.
(= ‘I do not agree with the rules.’)

9a I have no secretarial assistance and have to do everything for myself.

9b *I have no secretarial assistance and must do everything for myself.
(recurrent activity)

10a I’ve got to report to the office as soon as I get back.

10b I must report to the office as soon as I get back.
(particular instance)

MUST = HAVE (GOT) TO: EPISTEMIC MEANING (‘LOGICAL DEDUCTION’)

11a There must be some mistake.

11b There has to be some mistake.

11c There’s got to be some mistake.

12a You must be Susan’s husband.

12b ?You have to be Susan’s husband.

12c ?You’ve got to be Susan’s husband.

13a You must be getting old!

13b ?You have to be getting old!

13c ?You’ve got to be getting old!

14a You must be mad to do that. (‘you do that, and I conclude you’re mad’)

14b You have to be mad to do that.

14c You’ve got to be mad to do that.
(‘being mad is a necessary pre-condition for doing that’)

FUTURE PREDICTION or TIME REFERENCE: = EPISTEMIC WILL

15 Nobody knows what the future * HOLD for us
holds
is holding
is going to hold
will hold
shall hold

16 My flight LEAVE in half an hour.
leaves
is leaving
is going to leave
will leave
shall leave

17a He goes to London tomorrow (fact)

17b He will go to London tomorrow. (prediction)

REQUEST

18	I can't carry all this by myself.	*	help me? will would can could

EPISTEMIC POSSIBILITY

19	Bill	*	work for the BBC may might could perhaps works ... maybe works ...

It's possible that Bill works for the BBC.
It's quite likely that Bill works for the BBC.

CAN and MAY

20a The road can be blocked. (deduction from theoretical considerations)
20b The road may be blocked. (simple guess work)

21a Friends can betray you. (general case)
21b Your friends may betray you. (specific instance)

22a On Saturday night, we can have a party. (theoretical possibility)
22b On Saturday night, we may have a party. (Specific or real possibility)

23a Oil exploration can be very costly. (factual prediction: conjecture)
23b Oil exploration is very costly. (assertion)
23c Oil exploration may be very costly. (theoretical prediction: inference)

MIGHT AND COULD are 'perfect alternatives'

24a There could be trouble at the match. (theoretical possibility: inference)
24b There might be trouble at the match. factual possibility: conjecture)

25a I could play if my cold was better.
25b I might play if my cold was better.

25a He will be watching the football match.
25b He must be watching in football match.

26 John isn't here. He * be at the library.

==========

These examples show how certain modal concepts can indeed be expressed by different modal verbs, without difference in meaning; but they also show how the same verbs, in different contexts, expressing the

same concept, do express a difference in meaning. The explanations are often given in pragmatic terms. These examples further show that there are numerous cases where individual modal verbs are substitutable and seemingly synonymous, and also that there are other cases where substitution is not possible without a difference in semantic interpretation.

Besides, modal verbs are notorious for the fuzziness or indeterminacy of meaning, and some scholars have left examples simply uncategorised as between a sense (a) and a sense (b) (cf. Coates 1983 and Collins 1991a-b). Corpus work generates large numbers of instances of modal verbs and we are faced with the task of interpreting them. Each occurrence has to be taken as it stands - they cannot be altered. VARBRUL cannot cope with distinctions such as:

(1) My mother is very ill. I have to return home immediately after I've given my paper. In fact, I must return home. I couldn't forgive myself otherwise.

Here, not only do circumstances necessitate my early return, but I require it of myself. As exponents of 'obligation', all the encoding would be the same for *must* as for *have to*.

Here lies the difference between English and Thibault's study of *devoir* in Montréal French. Thibault proceeds from the assumption that the alternatives to *devoir* are indeed semantically equivalent and therefore, it would seem, entirely substitutable. Thus, I take it that the following utterance would be tautologous:

(2) Je suis supposé à y être à trois heures. En effet, je dois y être.

In her study, Thibault deals with each of the four senses of *devoir* separately. She compares its behaviour in each sense with that of the alternative expression. She then combines the results. These are then correlated with external factors such as sex and age of the informants, and interpreted. These overall results are in turn compared with the overall results of a previous project. The general conclusions are expressed in terms of sense category changes mirroring social changes, in the direction of change from above.

So what's the positive case for VARBRUL? It's its cross-tabulation facility, and the convenience it has for corpus linguists who analyse on the basis of concordance output. I showed last year how to input; now I can show you the output! This facility is not exclusive to it - we've seen

papers at ICAME along similar lines and produced using SPSS or Mini-Tab, for instance. But the cross tabulations are useful, if only because they're a convenient way of getting out what you can so conveniently code in.

Consider the following tables. I coded in data for 2014 tokens of WILL that I found in five corpora. I was concerned with three variables: root and epistemic senses; the status of the following lexical verb as stative or dynamic (you may recall Coates's claim that stative verbs occur 100% with epistemic WILL); and subject person (whether 1, 2 or 3, regardless of number). Table 6 presents each of these variables cross-tabulated in terms of actual numbers of occurrences and percentage distributions. It shows, for instance, that 25 stative verbs followed root senses of WILL, thus challenging Coates's claim. Or corpus 5, the LSC, had very few root WILLs at all - and when you consider the context of weather forecasts and such like (*there will be sunny showers all over the country this afternoon*), it's hardly surprising.

So I'd keep VARBRUL for its convenience for this type of analysis, and I'd be in good company - some veteran VARBRULers, such as Montgomery (1989), who has written an excellent introduction to the package, uses it for no more than this either. For the arguments outlined, however, I'm afraid I cannot recommend its unique statistical capacity to identify significant causes of variation in large collections of data where equivalence in meaning has to be assumed. If the question of semantic equivalence were surmountable, however, then the study of modality and modal verbs could proceed. There would first be a series of concept studies. Then the relevant parts for any particular verb could be extracted and combined for comparison, thus offering possibly refreshing new insights into the behaviour of each individual verb, as Thibault was able to do for *devoir*.

Table 6: Cross Tabulations of three variables in five corpora. The variables are root and epistemic sense; stative and dynamic following lexical verb; and subject person.
The corpora are (1) NITCS; (2) Miller; (3) Byrne; (4) Leuven; (5) LSC.

How else can modal verbs be studied? How else but using frequencies and percentage distributions with which ICAMERs are familiar and possibly, in an European way, more at home. Let me present some examples based on the published findings of others (notably Coates 1983 for Lund and LOB, and Collins 1991a-b for the Australian data) as well as my own work.

Table 5 presents the frequencies per 1,000 words of the nine central modals in 10 corpora. If I were to make one comment, it would be that WILL is confirmed as the most frequent modal in all corpora except one, my NITCS. There, the most frequent modal is WOULD, which can be explained in terms of the data: within an interview situation, anecdotes, recollections and reminiscences about the past and about how life used to be - the very function of WOULD. So the frequency is a consequence of the data, not the regionality. We could not legitimately say that in Irish English WOULD replaces WILL as the most frequent modal!

Table 5: Frequencies per 1,000 words of the nine central modal verbs in 10 corpora

Table 6 presents three sets of figures for the distribution of root and epistemic senses of WILL in 12 corpora: the actual numbers of occurrence, the frequency per 1,000 words, and the percentage distribution of each. If I were to make one comment, it is that in all cases the epistemic sense predominates, ranging from 53.0% in Byrne and 53.1% in Aus.W. to 91.0% in LSC.

Table 6: Root and Epistemic WILL in 12 corpora: occurences, frequencies and percentages

Table 7 reflects this multi-textual, multi-national set of corpora in terms of the ascending frequency of root meanings of WILL - from formal spoken and written texts, to informal spoken and written, to the problematic category of dramatic texts.

Table 7: Root and Epistemic Will in 12 corpora in ascending order of root frequency

Table 8 presents the percentage distributions for MUST in eight corpora. Here two patterns emerge: root/obligation MUST predominates in LOB, Lund and the Aus.W. corpus; epistemic/deductive MUST predominates in Australian and Northern Irish speech and in Scottish dramatic texts, no doubt because the latter make far greater use of HAVE (GOT) TO for the root/obligation sense. At any rate, between Lund and NITCS, there is quite a swing between the sense distribution of MUST.

Table 8: Root and Epistemic MUST in 8 corpora: percentages

Table 9 presents the occurrences of each of the five senses of SHOULD and their percentage distributions in six corpora, one of them the MILLER corpus of conversations among Edinburgh undergraduates. Its 82.2% root SHOULD distribution compares favourably with the spoken Australian 90.4%, and NITCS is in line with this at 59.3%, compared with Lund's 42.0%.

Table 9: The four senses of SHOULD in 6 corpora: occurences and percentages

One interpretation of the frequencies and distributions in Tables 8 and 9 is that they're all in line except English English, where its social and political prestige may have put it ahead of the regions and colonies, which remain conservative. This, of course, is just one of the topics open for investigation through the International Corpus - just how out of step is the educated speech and writing of England - that one variety

about which we know most? Or how far all these figures are no more than reflections of the data on which they are based, and we are still not ready on the basis of these relatively tiny samples to extrapolate out and induce generalisations about the whole? So, as the COBUILD team have been urging us for years, the more - the better!

Modal verbs are frequent, systematic, but complex; they are polysemeous; they are sometimes synonymous with other modal verbs, and sometimes offer clear semantic contrasts with the same modal verbs; and some verbs in some senses occur with high frequencies of syntactic correlations. We also know that in some regional varieties of spoken English, certain modals do not occur at all: such as the absence in Scottish and Irish English speech of *may, ought* and *shall*, except as formal borrowings from Standard English, used sometimes pretentiously, and that these vernacular patterns influence the writing of their speakers; and we also know that the modal verbs are significant markers of register variation and have been used in the identification of text types.

Modal verbs have received extensive study and been subjected to a wide range of analyses. Two of these approaches interest me in particular: those which are focused on individual verbs, and those which are focused on semantic concepts and their exponents. By using VARBRUL to study the exponents of individual concepts, my intention had been to combine the results into a series of new studies of each individual verb, and I thought I had found a lead in Thibault's study of *devoir*. All the same, there is still much to investigate in the spoken and written data available to me and my Belfast students in electronic form: (undergraduate conversations in Scottish English; interviews and recollections in Northern Irish English; in due course, the Irish spoken component of ICE; in due course, the Scottish and Northern Irish spoken components of the BNC; Scottish dramatic texts; and Scottish biblical texts; electronic editions of Irish writers such as Molloy and Doyle, who realistically represent the vernacular; and, in due course, the written components of ICE and BNC).

Existing frequencies (e.g. Coates 1983 and Hermerén 1986) can be further compared, and wider issues such as the status of Scottish and Northern Irish vernacular Englishes considered in the light of their similarities as well as differences and in their overall 'heteronymy' (i.e. lack of autonomy) from Standard English (cf. Kirk 1987).

References

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

Collins, P. (1991a): 'The Modals of Obligation and Necessity in Australian English'. In: K. Aijmer and B. Altenberg (eds.) *English Corpus Linguistics: Studies in Honour of Jan Svartvik.* London: Longman, pp. 145-165.

Collins, P. (1991b:) 'WILL and SHALL in Australian English'. In: S. Johansson and A.-B. Stenström (eds.) *English Computer Corpora*. Berlin: Mouton de Gruyter, pp. 181-199.

Hermerén, L. (1986): 'Modalities in Spoken and Written English: An Inventory of Forms'. In: G. Tottie and I. Bäcklund (eds.) *English in Speech and Writing: A Symposium* (Studia Anglistica Upsaliensia, vol. 60) pp. 57-91

Kirk, J.M. (1987): 'The Heteronomy of Scots with Standard English'. In: C. Macafee and I. Macleod (eds.) *The Nuttis Schell: Essays on the Scots Language Presented to A.J. Aitken*. Aberdeen: AUP, pp. 166-181.

Kirk, J.M. (1994a): 'VARBRUL'. In: L. Hughes and S. Lee (eds.) *Resources Guide*, THIRD EDITION. Oxford: CTI Centre for Textual Studies, pp. 27-28.

Kirk, J.M. (1994b): 'Concordances or Databases?'. In: U. Fries and G. Tottie (eds.) *Proceedings of the Fourteenth ICAME Conference.* Amsterdam: Rodopi, pp. 107-115.

Leech, G. (1989): *An A-Z of English Grammar and Usage.* London: Arnold.

Montgomery, M. (1989): 'Introduction to Variable Rule Analysis'. In: J*ournal of English Linguistics*, vol. 22, pp. 111-118.

Nevalainen, T. (1991): *BUT, ONLY, JUST: Focusing Adverbial Change in Modern English 1500-1900* (Mémoires de la Société Néophilologique de Helsinki, vol. LI).

Thibault, P. (1991): 'Semantic Overlaps of French Modal Expressions'. In: *Language Variation and Change*, vol. 3, pp. 191-222.

Tottie, G. (1988): '*No*-Negation and *Not*-Negation in Spoken and Written English'. In: M. Kytö, et al. (eds.) *Corpus Linguistics, Hard and Soft*. Amsterdam: Rodopi, pp. 245-265.