Davies, M., Rayson, R., Hunston, S. & Danielsson, P. (eds.) (2007). *Proceedings of the Corpus Linguistics Conference: Corpus Linguistics 2007*. Birmingham: University of Birmingham.

If-Conditionals as Modal Colligations: A Corpus-Based Investigation

Costas Gabrielatos[1]

1. Introduction

> A conditional never involves factuality, or more accurately … [it] never expresses the factuality of either of its constituent propositions. That one or other of the propositions is true may be known independently of the conditional, for instance from the rest of the verbal context or from other sources, but this does not alter the crucial fact that the condition itself does not express this actuality (Comrie, 1986: 89).

The weak claim motivating this study is that *if*-conditionals[2] are strong modality attractors, due to the conditional (i.e. modal) meaning of *if*, with modality appearing in the *if*-clause, the main clause, or both. The strong claim is that *if*-conditionals can be regarded as *modal colligations*. The weak claim can be supported if it is shown that *if*-conditionals contain modality in a significantly higher than average frequency. Before examining the conditions under which the strong claim can be supported we need to turn our attention to the notions of *modality*, *collocation, colligation* and *semantic preference*, which inform the notion of *modal colligation* introduced in this paper.

2. Theoretical background

2.1 Modality and its formal realisations

Modality is "concerned with the speaker's attitude towards the factuality or actualisation of the situation expressed by the rest of the clause" (Huddleston and Pullum, 2002: 173). There are several frameworks grouping modal notions in different ways;[3] however, accounts of modality seem to converge on modality expressing attitude towards actuality, factuality, likelihood, probability, ability, potentiality and desirability (the latter including volition, obligation and permission). Modality can be expressed through a variety of formal means, such as modal auxiliaries (e.g. *may, ought to*), catenative verbs (e.g. *need, want*), adverbs (e.g. *possibly, probably*), the imperative, the past tense (in some contexts, e.g. conditionals), as well as constructions involving lexical verbs (e.g. *it appears that …*),

[1] Department of Linguistics and English Language, Lancaster University
e-mail: c.gabrielatos@lancaster.ac.uk

[2] The terms 'conditional' and '*if*-conditional' are preferred to 'conditional sentence', because not all conditional constructions are full sentences, or are within the same sentence, or conform to the structure of a sentence (see Gabrielatos, 2005). However, for ease of reference, '*if*-sentence' will be used to refer to what the TEI guidelines term *s-unit* (Sperberg-McQueen and Burnard, 2007), that is, a unit containing the word *if*, which is delimited by a sentence-boundary marker (e.g. full-stop, question mark, exclamation mark). See also section 3.2.

[3] See Biber *et al.* (1999), Bybee *et al.* (1994), Coates (1983), Hoye (1997), Halliday (1994), Huddleston and Pullum (2002), Lyons (1977), Palmer (1990, 2003), Quirk *et al.* (1985).

adjectives (e.g. *it is likely/possible that ...; it is imperative that ...*), or nouns (e.g. *there is a chance/likelihood that ...; we have an obligation to ...*). Given the multifaceted formal realisation of modality the terms *modal* and *modal expression* will be used to refer to any word, multi-word expression, or grammatical category expressing modal meaning.

2.2 The notions of *semantic colligation* and *modal colligation*

The notion of colligation is closely related to that of collocation (Firth, 1951/57: 195-196). Collocation is a relation between words, and has been defined in various ways (see Partington, 1998: 15-16). This study adopts what Partington (ibid.: 16) describes as the "statistical" definition, that is "the relationship a lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey, 1991: 6-7). The term *colligation* was introduced by Firth (1968: 181) in order to distinguish lexical interrelations from those holding between grammatical categories:

> The statement of meaning at the grammatical level is in terms of word and sentence classes or of similar categories and of the interrelation of those categories in colligations. Grammatical relations should not be regarded as relations between words as such – between *watched* and *him* in 'I watched him' – but between a personal pronoun, first person singular nominative, the past tense.

However, the currently preferred definition of colligation is less restricted and encompasses the statistically calculated co-occurrence of lexis and grammatical categories (Stubbs, 2002: 65), or more simply, "the grammatical company a word keeps" (Hoey, 1997: 8). In this study, the term ´colligation´ is used more in the sense of Firth's (1968: 181) definition, in that it refers to the co-occurrence of categories, though not only grammatical ones. Irrespective of the definition adopted, colligation, like collocation, is a probabilistic relation. *Semantic preference* is the attraction "between a lemma or word form and a set of semantically related words" (Stubbs, 2002: 65).

The relation described here by the term *modal colligation* is better understood as a hybrid between colligation and semantic preference, and in more general terms it could be termed *semantic colligation*. That is, it is the mutual attraction holding between a sentence class, conditional sentences in general, and *if*-conditionals in particular, and "a set of semantically related words" (ibid.), or, more generally, a semantic category, that is, modality. In that light, the strong claim will gain support if it is also shown that *if*-conditionals attract modality significantly more frequently than non-conditional sentences with *if*.

3. Methodology

3.1 Overview

The corpora used in this study (see section 3.2 below) were analysed in a number of ways. First, the manually annotated Sample was examined for the frequency of modalisation in the *if*-clause and main clause, in order to establish the *modal load* of the *if*-sentences in the Sample (i.e. their degree of modalisation). The next step was to

carry out a number of keyword analyses, automatic and manual, in order to establish the extent to which words or grammatical categories expressing modality are statistically significantly more frequent in the sample of *if*-conditionals and the sub-corpus of the written BNC containing all *if*-sentences.

In some respects, the methodology used in this paper, or, at least, the focus of the study, shows some similarities to *collostructional analysis* (Gries and Stefanowitsch, 2004; Stefanowitsch and Gries, 2003). Collostructional analysis takes a construction as its starting point and examines the lexemes that are attracted or repelled by particular slots in the construction. The degree of attraction/repellence is calculated according to whether a lexeme occurs more/less frequently than expected in a particular slot of a construction. However, this study is not concerned with individual words, nor with the particular slot they occupy within *if*-conditionals, but with words or grammatical categories with modal meaning taken collectively as indicators of modalisation - although it is also of interest to compare the degree of modalisation in the *if*-clause and main clause (see section 4).

3.2 Corpora used in the study

The sample of *if*-conditionals used in this paper was derived as follows. From the query of the word *if* in the written BNC (Aston and Burnard, 1998), which returned 205,275 matches, a random 1,000 instances were selected using the 'thin' function of BNCweb.[4] The resulting sample was analysed manually, and instances of *if* without conditional meaning were removed (see table 1 below).

Meaning of *if*	**Examples**	**Freq.**
if = whether	*In the days that followed, Nigel kicked himself for not untying the bunch of flowers and looking to see* if *there was a card inside*. [AC3 2215]	57
as if = as though	*In contrast to the outside, the area was softly carpeted, softly lit, as* if *illness and death had to be cushioned away, made to look as if they didn't exist*. [BPD 200]	55
if = although	*In this context he formulates his now familiar,* if *still empirically untested, distinction between "restricted" and "extended" professionality*. [FAM 360]	13
even if = even though	*And she also reminded herself that even* if *it had included a kiss it didn't mean a thing -- especially as it had come from a man who was living in the outback to get away from women*. [HHB 14622]	12
if only = wish	*Licence Revoked was originally going to be the title for this last Bond film, and* if *only they had*. [ACN 464]	4
if = TERM	*We need a law for relating* IF *and ALT*. [G3N 120]	2
if = that	*As the students have not been introduced to prepayment at this stage in any detail it is better* if *the former assumption is made*. [HW9 548]	1
Idiomatic (if = that)	*I shall have her for that, you see* if *I don't!* [CH4 1077]	1
Creole	*Well if you want* if *you want exchange it den well you afi chat it*. [HXY 1098]	1
if = TYPO	*But* if *was the sequence of images and ideas that was most entrancing*. [B74 1692]	1

Table 1: Non-conditional uses of *if*.

[4] This interface to the BNC was developed by Hans-Martin Lehmann, Sebastian Hoffmann and Peter Schneider. See also, http://es-otto.unizh.ch/bncweb2/manual/bncwebman-home.htm.

This created a sample of 853 *if*-conditionals (Sample, 25,666 words), which was then manually annotated for form (tense/aspect marking and modal expression), meaning (type of modality and modal notion), and type of conditional (i.e. the semantic or pragmatic relation holding between the *if*-clause and main clause).[5] This study also used a number of other corpora, used in their annotated and unannotated versions: a sub-corpus of the written BNC containing all sentences including the word *if* (*If*-BNCw, 5.4 million words),[6] the written BNC (BNCw), the written BNC Sampler (BNCSw), FLOB (Hundt *et al.*, 1998).

3.3 Keyword analysis

The main methodology that will be used to examine whether *if*-conditionals attract modality significantly more frequently than non-conditional sentences will be *keyword analysis* (e.g. Scott, 1997). The frequencies of the words in the corpus under investigation (the *study corpus*) are compared to those in a relevant representative corpus (the *reference corpus*) in order to identify the words that are significantly more, or less, frequent in the study corpus in comparison with the reference corpus. These words are termed *positive* and *negative keywords* respectively.[7] The analysis, and the calculation of statistical significance, also take into account the sizes of the study and reference corpora. Keyword analysis was carried out using the *WordSmith* corpus software (Scott, 1998); however, the frequencies of the central modals (treated as a group) were compared manually, using Paul Rayson's online log-likelihood calculator[8] (see also Rayson and Garside, 2000). The statistical significance of the differences in frequency of any keywords (termed *keyness*) was measured by the log-likelihood statistic, with the significance level set at $p \leq 0.01$ ($LL \geq 6.63$). As the Sample is small, and because the focus is on modal expressions taken collectively, the minimum frequency for the consideration of a word in the keyword analysis was set to one occurrence. In order for results to be comparable, this minimum was used for all study corpora.

It must be reiterated that establishing the keyness of individual modals was a means to an end. The aim was to determine the extent to which modals, as a group, are significantly more frequent in the study corpora, which, in turn, would be a strong indication that the study corpora contain a significantly higher proportion of modality. In other words, the aim is not to establish lexicogrammatical preference, but semantic preference. The hypothesis is that the two study corpora (Sample, *if*-BNCw) will show a much higher proportion of positive than negative modal keywords, not only in absolute terms, but also, and more importantly, in terms of the proportion of modal keywords among all keywords in each comparison.

[5] For typologies of conditionals, see Athanasiadou & Dirven (1997), Quirk *et al.* (1985: 1088-1089), Sweetser (1990: 116-117).

[6] I am grateful to Sebastian Hoffman (Lancaster University) for cleaning and removing the duplicate sentences from the sub-corpus. The annotation of *if*-BNCw was carried out using the web interface *Wmatrix* (Rayson, 2003, 2007).

[7] Unless stated otherwise, the terms 'keyword' and 'keyness' will refer to positive keywords/keyness.

[8] http://ucrel.lancs.ac.uk/llwizard.html

4. Analysis and discussion

4.1 Modalisation of *if*-conditionals in the Sample

Let us start the examination of these claims, by looking at the modalisation of the *if*-conditionals in the manually annotated Sample, or rather, the modal marking of the *if*-clauses and main clauses (see also Gabrielatos, 2003). As table 2 shows, about one-third of *if*-clauses are modalised.

Category	Freq.	*% (n=853)*
Unmodalised	570	*66.8%*
Modalised	280	*32.8%*
Elliptical (non-inferable)	3	*0.3%*
Total	853	*100%*

Table 2: *If*-clause modalisation

To be more precise, one-third of the *if*-clauses in the sample have additional modality marking, as the presence of conditional *if* has already marked the clause for hypotheticality. This, in itself, is a clear indicator of the attraction to modality that *if*-conditionals exert. It is also interesting to note that in some cases modality is marked more than once: 1 percent of all *if*-clauses in the sample, and 3.2 percent of the modalised ones, have two or three modal markers.

Modalisation is much more frequent in the main clause of *if*-conditionals, with almost three-quarters of the main sentences being marked for modality (table 3). Also, multiple modality marking is significantly more prominent in main clauses, with 7.3 percent containing two or three modal expressions.

Category	Freq.	*% (n=853)*
Modalised	607	*71.1%*
Unmodalised	230	*27.0%*
Elliptical (non-inferable)	16	*1.9%*
Total	853	*100%*

Table 3: Main clause modalisation

In the light of the above, a rough calculation can provide an overall picture of the average modal load of *if*-conditionals. The sample contains 853 *if*-conditionals, which can be seen as consisting of two clauses each. In total, more than half of these clauses (887) are modalised, and about 4 percent of them contain two or more modal expressions. The picture that emerges, then, is that, on average, for each *if*-conditional in the sample corresponds at least one modal expression (discounting *if* itself). This seems to indicate that *if*-conditionals do attract modality. However, this is only a rough indicator, as not all *if*-conditionals consist of two clauses (e.g. Frazier, 2003). More importantly, the high frequency of modalisation of the *if*-conditionals in the Sample does not necessarily entail that they attract modality with a higher than average frequency, or that this attraction is statistically significant.

4.2 Keyword analysis

Ideally, in order to test whether modals are found in *if*-conditionals significantly more frequently than average, we need to compare their frequency in the Sample and a corpus of non-conditional sentences. However, compiling such a corpus may be unnecessary. If the comparison of the Sample to a representative corpus of British English (used as the reference corpus) reveals that the study corpus is significantly richer in modals, then the presence of conditional sentences in the reference corpus will only strengthen the validity of the results. On the basis of the number of words in *if*-BNCw (5.4 million), and judging from the makeup of the initial sample, in which *if* has conditional meaning in some 85 percent of the cases, it is estimated that around 6 percent of the words in BNCw are in *if*-conditionals, and this is expected to also be the case with BNCSw.

The comparison of the Sample with BNCSw[9] yielded 873 positive keywords, of which 27 were modal, and 135 negative keywords, none of which had a modal sense. The modal keywords include all central modals,[10] four marginal auxiliaries (Quirk *et al.*, 1985: 236-237) (*be (un)able to, need, want*), one or more forms of six lexical verbs expressing modal notions (*comply*, *doubted, feel, know/knew, required, think/thinks*), an adjective (*necessary*), two adverbs (*probably*, *hopefully*), two nouns (*evidence, obligation*), and (*be) liable (to)* (see appendix 1). The proportion of positive modal keywords was 3.09 percent, whereas the comparison returned no negative modal keywords. Therefore, it can be argued that the difference in frequency would be more marked, and more modal expressions would be key in the Sample, if the reference corpus did not include conditionals.

However, we also need to consider that BNCSw does not contain full texts, but text samples. It has been argued (Berber-Sardinha, 2000: 12) that "the number of keywords seems to vary considerably as a function of the size of the texts (Mike Scott, personal communication). Shorter texts provide less room for repetition, which in turn influences word frequencies". In order to test whether the sampling of text fragments in BNCSw has affected word frequencies, and hence keywords, the Sample was also compared to FLOB, a representative corpus which also consists of text fragments, and is of the same size as BNCSw. It is argued that if text sampling affects word frequencies and keywords unduly, then the comparison with a corpus consisting of different fragments should be expected to bring about a significantly different list of keywords. This is not the case: out of 758 positive keywords, 21 are modal expressions (2.77 percent), and out of 72 negative keywords, only one (*seemed*) has modal meaning (1.39 percent). Again, all the central modals are keywords, as well as a number of marginal auxiliaries, lexical verbs, adjectives, adverbs and nouns with modal meaning (see appendix 2). The single negative modal keyword in the comparison with FLOB can be explained by the small size of the Sample. Also, one negative keyword common to both comparisons (*was*) may be taken as a result of the higher frequency in the Sample of the modal use of *were* instead of *was* (although *was* is also used with modal meaning in the Sample).[11]

---

[9] The BNC sampler is a shorter version of the full BNC, and consists of two one-million-word sub-corpora of written and spoken British English (http://www.natcorp.ox.ac.uk/getting/sampler.html).

[10] Central modals are: *can, could, may, might, must, shall, should, will, would* (Quirk *et al.*, 1985: 137).

[11] One way to examine the validity of this claim is to investigate key bigrams (i.e. two-word clusters irrespective of meaning) for the existence of clusters such as 'proper noun + *were*' or '*I/he/she/it were*', indicating the use of *were* with modal meaning. This technique can also be used to establish the keyness of multi-word modals in general, such as *am/are/is/was/were to*, or *have/has/had to*. The

However, we should also entertain the possibility that the apparent semantic attraction is not a characteristic of *if*-conditionals in general, but of the makeup of the *if*-conditionals in the specific random sample, which might be unusually rich in modal expressions. In order to be able to discount this possibility, the Sample will be compared with *if*-BNCw, in which some 85 percent of *if*-sentences can be expected to have conditional meaning (based on their proportion in the Sample). If the comparison reveals the same key modals, or a comparable number of key modals, this will be deemed a strong indication that the Sample contains an uncharacteristically high proportion of modality marking. If, however, as is predicted, the comparison does not reveal any positive or negative modal keywords, or if the number of negative and positive modal keywords is balanced, then the Sample can be considered representative in terms of modal load, and the results of the keyword analysis so far will stand. Expressed in terms of proportion, the more modal expressions turn out to be key, the more loaded modally the Sample will prove to be in comparison to *if*-BNCw, and, hence, less representative.

The comparison yielded only one positive modal keyword (*shall*) representing 0.19 percent of all positive keywords, and one negative modal keyword (*wants*), representing 2.78 percent of all negative keywords, which indicates that the Sample is representative. However, the hypothesis will be further supported if *if*-BNCw itself is compared to BNCSw and FLOB in order to determine keywords. If, as is hypothesised here, *if*-conditionals attract modality significantly more strongly than non-conditional sentences, and given that larger corpora tend to return a larger number of keywords, then the comparison should yield a larger number, and proportion, of positive modal keywords than that derived when the study corpus was the Sample, and no, or significantly fewer, negative keywords. Also, the positive modal keywords are expected to include the most frequent modals. The hypothesis is borne out: all central, and some marginal, modal auxiliaries, as well as the main lexical verbs, adverbs, adjectives and nouns expressing modal notions are positive keywords in both comparisons (see appendices 3 and 4). The comparisons returned 92 (4.42 percent) and 63 (3.86 percent) positive modal keywords respectively, and only 6 (0.14 percent) and 9 (0.20 percent) negative modal keywords respectively.

As table 4 below shows, all four comparisons consistently returned a significantly higher number of positive modal keywords, both in absolute and relative terms. We also need to keep in mind that, in all comparisons, the positive keywords included the most common modals. Also, the significance of the modal keywords in all comparisons was overall higher than the non-modal ones. In the comparison of the Sample with BNCSw and FLOB about one-quarter of modal keywords was among the top 25 percent of all keywords; in the comparison of *if*-BNCw with BNCSw and FLOB more than half of modal keywords were among the top 25 percent of all keywords.

---

analysis of bigrams showed that *it were* and *are to* are key in the Sample. However, key n-grams should be treated with caution for three reasons. N-grams are merely pairs of adjacent words, not necessarily lexicogrammatical, or even meaningful, units. Also, the frequency of a given n-gram depends on the frequency of all the constituent words, and, therefore, the keyness of a given n-gram may depend on the relative high frequency of words other than the ones relevant to the claim under investigation. Consequently, the statistical calculation of key n-grams is less reliable than that of single words (Rayson, *et al.*, 2004: 4).

Comparison	Modal keywords			
	Positive	**Negative**	***Positive %***	***Negative %***
Sample * BNCSw	27	0	*3.09*	*0*
Sample * FLOB	21	1	*2.77*	*1.39*
if-BNCw * BNCSw	93	6	*4.47*	*0.14*
if-BNCw * FLOB	63	9	*3.92*	*0.20*

Table 4: Modal keywords returned in the comparisons

At this point, we must consider the following question: Is the attraction of modality a feature of conditional constructions, or simply the presence of the word *if*? In order to address this question we need to compare non-conditional with conditional *if*-sentences, as the absence of positive and/or the presence of negative modal keywords will support the claim that it is the *if*-conditionals, rather than the word *if* itself, that attract modality. The comparison of the Sample with the non-conditional *if*-sentences in the sample yields one positive modal keyword, the central modal *may*, representing 8.33 percent of all positive keywords (the highest proportion in all comparisons). There were no negative modal keywords, although the total number of negative keywords was almost six times that of the positive ones. All the above indicate that the Sample and the non-conditional *if*-sentences do not differ substantially in terms of their modal load. This can be interpreted as suggesting that it may be useful to re-assess the utility of separating *if*-constructions into 'conditional' and 'non-conditional'.[12] An alternative course of action can be the examination of the modal load of different types of conditionals. Informal observations during the annotation and analysis indicate that the type of conditionals termed *content* (Sweetser, 1990: 116-117), or *direct* (Quirk *et al.*, 1985: 1088-1089), as exemplified by (1), attract modality more frequently than the type of conditionals termed *indirect* (Quirk *et al.*, 1985: 1088-1089), *speech act* (Sweetser, 1990: 118-121), or *pragmatic* (Athanasiadou and Dirven, 1997), as exemplified by (2).

(1) If the material is not topical and is delayed for some time before being sent out the date should be changed. [EX6 714]

(2) Out of the corner of his eye he saw Hammond start forward. "But you promised ..." Spatz interrupted Hammond, his face hard. "I promised nothing, if you recall." [GUG 121]

4.3 Relative frequency of central modals in the Sample

Overall, the results of the automatic keyword analyses carried out so far seem to indicate that *if*-conditionals do attract modality more than average, and that the higher frequency of the most common modal words/expressions is statistically significant. However, it has to be borne in mind that keyword analysis only provides a general picture. Also, since the analysis was carried out on untagged corpora, some distortion may have been caused by the lack of differentiation between the modal auxiliaries *may, might*, *must* and *will*, and their respective homographic nouns. Therefore, there are grounds for carrying out a manual analysis, akin to a keyword analysis, in which the frequencies of the central modals, as established by the manual annotation and

[12] For example, Athanasiadou and Dirven (1996: 613) argue that the type of relationship exemplified in (2) is not conditional.

analysis of the Sample, is compared to the frequency of central modals in an annotated reference corpus (BNCw), and the statistical significance is calculated. In this comparison, some form of lexical abstraction, akin to lemmatisation, has been performed. That is, the frequencies of the contracted forms (*'d, 'll*), and the contracted or idiosyncratic negative forms (e.g. *can't, cannot, shan't, won't*) of central modals have been pooled together with their full forms. As will become evident below, this comparison reveals some issues which pose interesting questions for the quantitative aspect of corpus linguistics research.

The frequency of the central modals in the Sample was compared with their frequency in BNCw. Table 5 shows the actual and normalised frequencies (per million word) of the central modals in the two corpora, the difference in frequency (as a positive or negative percent value), and its statistical significance.[13] The central modals are given in descending order of their statistical significance.

	Sample		BNCw			
Modal	**Freq.**	**Freq./mil.**	**Freq.**	**Freq./mil.**	**Diff. %**	**LL**
would	152	6030.07	232738	2666.43	*126.2%*	**78.46**
might	33	1309.16	50757	581.51	*125.1%*	**16.87**
will	113	4482.88	271838	3114.40	*43.9%*	**13.35**
can	85	3372.08	194664	2230.22	*51.2%*	**12.73**
must	32	1269.49	63840	731.40	*73.6%*	**8.16**
may	47	1864.56	107805	1235.10	*51.0%*	**6.99**
should	41	1626.53	97043	1111.80	*46.3%*	5.25
could	51	2023.25	139997	1603.91	*26.1%*	2.56
shall	8	317.37	17426	199.65	*59.0%*	1.48

Table 5: Keyword comparison of central modals in the Sample and BNCw.

Although all central modals have a higher relative frequency in the Sample (ranging from 26 percent to 126 percent), the results of the comparison differ from those of the automated keyword analyses in two interrelated respects. Table 5 above shows that fewer central modals are significantly more frequent in the sample: six compared to nine. Furthermore, those with a statistically significant higher frequency show a lower average strength of significance. Obviously, this begs for an explanation, which will be sought in the nature of keyword analysis and the log-likelihood statistic.

Both values are sensitive not only to the relative frequency of the word in focus, but also to its actual frequency in the study and reference corpora, and, consequently, to the size of the two corpora. For example, although *should* and *will* have almost the same relative frequency difference in the Sample and BNCw, +46.3 percent and +43.9 percent respectively (see table 5 above), the log likelihood score of *should* is only 5.25 (below the 99 percent confidence level), whereas that of *will* is 13.35 (well above the 99.9 percent confidence level). The reason behind this difference in significance levels is the marked difference in the raw frequency of *should* and *will* in the sample (41 and 113 respectively). The sensitivity to the actual frequency of a word is demonstrated more clearly by the case of *shall*, the relative frequency difference of which (+59 percent, i.e. higher than either *should* or *will*) shows a log-likelihood score of merely 1.48. This is because its very low raw frequency in the sample (8 tokens) makes it highly probable that its frequency, and any differences in relative frequency, are due to chance. A further example is the case of *should*, the relative frequency of

[13] Values of 6.6 or higher (corresponding to a significance level of $p \leq 0.01$) are shown in bold.

which is only marginally below the 99 percent significance level (see table 5 above). An increase in actual frequency of a mere two tokens (from 41 to 43) would bring up the log-likelihood value from 5.25 to 6.87, that is, it would take it across the threshold of statistical significance set in this study. The sensitivity to the actual corpus size is more clearly shown if we derive the log-likelihood values not on the basis of actual frequencies and corpus sizes, but using the normalised frequencies (per million words), that is, if the comparison is performed as if both the study and reference corpora were 1 million words each. For example, the log-likelihood value for *shall* calculated using the normalised frequencies is 27.04 - a dramatic difference from the 1.48 score derived on the basis of the actual frequencies and corpus sizes.

In the light of this, it seems reasonable to carry out the comparison using reference corpora that are closer to the size of the study corpus, namely BNCSw and FLOB, which were also used in the first group of comparisons (i.e. the automatic keyword analyses). Interestingly enough, as tables 6 and 7 demonstrate, both comparisons yield one keyword fewer than before, five instead of six. In fact, in the comparison with BNCSw, *could* shows a slightly, although not statistically significant, lower frequency in the Sample (-16.3 percent, LL=1.67).

	Sample		**BNCS**			
Modal	**Freq.**	**Freq./mil.**	**Freq.**	**Freq./mil.**	**Diff. %**	**LL**
would	152	6030.07	2615	2416.64	*149.5%*	**92.75**
Might	33	1309.16	471	435.23	*200.1%*	**27.63**
May	47	1864.56	1022	944.40	*97.4%*	**17.04**
Can	85	3372.08	2264	2092.09	*61.2%*	**16.18**
Will	113	4482.88	3546	3276.75	*36.8%*	**9.77**
Must	32	1269.49	863	797.47	*59.2%*	5.80
should	41	1626.53	1388	1282.61	*26.8%*	2.09
Could	51	2023.25	2615	2416.44	*-16.3%*	1.67
Shall	8	317.37	224	206.99	*53.3%*	1.24

Table 6: Keyword comparison of central modals in the Sample and the BNCSw.

	Sample		**FLOB**			
Modal	**Freq.**	**Freq./mil.**	**Freq.**	**Freq./mil.**	**Diff. %**	**LL**
would	152	6030.07	2719	2664.06	*126.3%*	**76.09**
Will	113	4482.88	2603	2550.40	*75.8%*	**29.16**
Can	85	3372.08	1997	1956.65	*72.3%*	**20.55**
might	33	1309.16	642	629.02	*108.1%*	**13.64**
May	47	1864.56	1102	1079.73	*72.7%*	**11.44**
Must	32	1269.49	815	798.53	*59.0%*	5.76
should	41	1626.53	1148	1124.80	*44.6%*	4.82
Shall	8	317.37	197	193.02	*64.4%*	1.64
could	51	2023.25	1771	1735.21	*16.6%*	1.11

Table 7: Keyword comparison of central modals in the Sample and FLOB.

The similar results of the comparisons with BNCw, BNCSw and FLOB seem to indicate that the relative sizes of the study corpus and the reference corpora do not explain the lower number and statistical significance of key central modals yielded by the comparison of abstracted central modals, as contrasted with the automatic keyword analysis on the basis of word-forms. However, the relevance of the size of

the sample corpus should not be ruled out. The justification for further investigation along these lines comes from a comparison of the discrepancies between the actual frequencies in the Sample of some of the central modal verbs in the two types of comparison (i.e. word-forms and abstracted verbs). For example, *may* and *might* have a frequency of 72 and 47 respectively in the first group of comparisons (see appendices 1 and 2), but only 47 and 33 in the second group (see tables 6 and 7 above). It should also be noted that this difference cannot be accounted for by the exclusion of the nouns *May, must, might* and *will* in the second group of comparisons, as they were too infrequent in the Sample to affect results.[14] Rather, the discrepancy seems to be due to the fact that *if*-conditionals in natural language are not always the neat pairs of a conditional and main clause in direct sequence presented in logic and language teaching materials, as examples (3)-(6) from the Sample demonstrate.

(3) Yes, I come from Lochaber, and the Lochaber people, if they were here, would be at one with the people of Breadalbane. [A0N 706]

(4) If the leg is cured while it is still attached, it is technically a gammon -- hence the confusion caused by the term "gammon ham". [ABB 217]

(5) As an academic critic and university teacher specializing in modern literature and literary theory, I spend much or my time these days reading books and articles that I can barely understand and that cause my wife (a graduate with a good honours degree in English language and literature) to utter loud cries of pain and nausea if her eye happens to fall on them. [A1A 208]

(6) Why should the fact that D was engaged on causing damage to property at the time (even damage to D's own property) make his conduct into an offence punishable with life imprisonment when, if D were engaged on some other activity, it would not be punishable as such and would only amount to manslaughter if a death happened to be caused? [ACJ 627]

The examples above show that *if*-conditionals may be embedded within, or be co-ordinated with, other clauses, which creates the frequency discrepancies identified between the automated and manual comparisons. At the same time, these examples indicate the less than straightforward nature of delimiting *if*-conditionals in a corpus (Gabrielatos, 2005). Let us take examples (3) and (4). In order to only calculate the words in the *if*-conditional, we would need to remove the introductory clause "*Yes, I come from Lochaber*" and the conjunction *and* in (3), and the clause '*hence the confusion caused by the term "gammon ham"*' in (4), as, by providing additional information, they are not part of the conditional *per se*. Examples (3) and (4) also provide an indication of the possible overestimation of the sample size: the 'inessential' elements account for 27.3 percent of (3) and 37.5 percent of (4). Examples (5) and (6) indicate the difficulty involved in extracting embedded *if*-conditionals. However, even when isolating the *if*-conditional is simple, as in (3) and (4), it is inadvisable, as it unavoidably projects the analyst's intuitions, and is tantamount to tampering with the evidence (ibid.). Therefore, by virtue of the corpus-based methodology used here, these elements were included in the token count in all comparisons. Perhaps, then, the comparisons of modal load should be carried out in terms of the number of s-units in a corpus rather than the number of words (see also Ball, 1994: 299-300).

[14] Only the nouns *May* (3) and *will* (2) were present in the Sample.

At this point, we need to remind ourselves of the claim under investigation. The automatic keyword analysis and the frequency comparison of abstracted central modals in the Sample, on the one hand, and BNCw, BNCSw and FLOB, on the other, unavoidably focused on words. However, the claim is that *if*-conditionals attract modality, rather than each and all modal expressions, at a significantly higher degree than average. Of course, the question of whether specific modal expressions demonstrate various degrees of attraction to conditional contexts is itself highly interesting and clearly worth investigating. However, the claim refers to modalisation (i.e. to modality as a notional category), whatever the nature or make-up of its linguistic realisation. In this light, the results of the keyword analyses can be seen as a conservative estimate of the modal load in the Sample, as they did not include the imperative, or the modal use of the past tense (i.e. when past tense marking denotes remoteness in terms of actuality, factuality or likelihood, rather than past time).

Baker (2004) advises that, in conjunction with a focus on individual keywords, corpus-based research would be wise to also examine the keyness of groups of notionally related low-frequency keywords. In the same vein, it also seems sensible to examine the keyness of such notional groups when the individual words are not key. Ideally, then, in order to test the claim, the frequencies of all modal expressions (lexical and grammatical) in the Sample and reference corpora should be totalled, and then their relative frequencies tested for statistical significance. This has been done manually for the Sample, but it does not seem feasible to carry out automatically on the much larger reference corpora, and will be prohibitively time consuming to carry out manually. For example, the modal use of the past tense of lexical verbs cannot be picked out automatically, although, in the Sample, the past tense accounts for the modalisation of 8 percent of all *if*-clauses, and almost half (46.7 percent) of the modalised *if*-clauses (Gabrielatos, 2003). What is feasible, however, is to carry out such a comparison with the central modals taken collectively. This comparison may not, on its own, be in a position to provide conclusive evidence; it can, however, provide a strong indication of whether *if*-conditionals attract modality, as the group of central modals, although representing just over 12 percent of the modal types in the Sample, accounts for almost 60 percent of the modal tokens.[15] The comparison revealed that the group of central modals is clearly more frequent in the Sample than in any of the reference corpora, and that this difference is significant at a very high level of confidence, with the *p* value being no higher than 10^{-14}. Tables 8, 9 and 10 show the raw and relative frequency of central modals in the Sample and reference corpora, the relative frequency difference and its statistical significance.

Sample		**BNCw**		**Diff. %**	**LL**
Freq.	**Freq./mil.**	**Freq.**	**Freq./mil.**		
562	22,295.39	1,176,108	13,474.40	*65.5%*	121.36

Table 8: Comparison of Sample and BNCw

[15] Similar results are reported in Gabrielatos and McEnery (2005), based on a 1.15 million word corpus of MA dissertations written by native speakers of English.

Sample		BNCSw		Diff. %	LL
Freq.	**Freq./mil.**	**Freq.**	**Freq./mil.**		
562	22,295.39	15,008	13,868.42	*60.8%*	105.87

Table 9: Comparison of Sample and BNCSw

Sample		FLOB		Diff. %	LL
Freq.	**Freq./mil.**	**Freq.**	**Freq./mil.**		
562	22,295.39	12,994	12,731.43	*75.1%*	143.29

Table 10: Comparison of Sample and FLOB

5. Conclusion

The comparisons of the modal load of the *if*-conditionals in the Sample with three balanced representative corpora of British English have provided evidence overwhelmingly supporting the claim that *if*-conditionals, seen as a group, carry a significantly higher modal load than average. The significantly high degree of modalisation, together with the not inconsiderable frequency of multiple modality marking, and the modal nature of *if* itself, also seem to indicate that modality attracts modality.

However, there remain a number of issues to resolve through further research. In terms of lexis-based analysis, a collocational analysis of the word *if* needs to be carried out in order to establish the number, nature and strength of its modal collocates. Also, as the comparison of conditional and non-conditional *if*-sentences did not reveal significant differences in modal load, different types of non-conditional *if*-sentences need to be examined separately (e.g. reported questions), preferably using a larger sample. Similarly, the modal load of *if*-sentences should be compared to that of comparable constructions involving a subordinate and main clause, such as *when*-sentences, particularly given the existence of the expression *if and when*. Irrespective of whether the analysis is carried out in terms of number of words or number of sentences, the correlation between modal load and type of conditional also needs to be investigated. At the same time, the influence of genre and context of use on the attraction of modality in general, and specific modal expressions in particular, should be examined. Finally, as the present analysis only involved written language, a sample of *if*-conditionals from the spoken BNC needs to be examined using the above techniques.

References

Aston, G. and L. Burnard (1998) *The BNC Handbook*. Edinburgh: Edinburgh University Press.

Athanasiadou, A. and R. Dirven. (1996) Typology of *if*-clauses, in E.H. Casad (ed.) *Cognitive Linguistics in the Redwoods: The expansion of a new paradigm in linguistics*, pp. 609-54. Cognitive Linguistics Research, 6. Berlin: Mouton de Gruyter.

Athanasiadou, A. and R. Dirven (1997) Conditionality, Hypotheticality, Counterfactuality, in A. Athanasiadou and R. Dirven (eds) *On Conditionals Again*, pp. 61–96. Amsterdam: John Benjamins.

Baker, P. (2004) 'Querying keywords: Questions of difference, frequency, and sense in keywords analysis'. *Journal of English Linguistics* 32(4), 346-359

Ball, C.N. (1994) 'Automated text analysis: Cautionary tales'. *Literary and Linguistic Computing* 9(4), 265-302.

Berber-Sardinha, T. (2000) Comparing corpora with WordSmith Tools: How large must the reference corpus be?, in Proceedings of the Workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000). 1-8 October 2000, Hong Kong, pp. 7-13. Also available online from http://acl.ldc.upenn.edu/W/W00/W00-0902.pdf

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999) *Longman Grammar of Spoken and Written English.* London: Longman.

Bybee, J., R. Perkins and W. Pagliuca (1994) *The Evolution of Grammar: Tense, aspect, and modality in the languages of the world*. Chicago: The University of Chicago Press.

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

Comrie, B. (1986) Conditionals: A typology, in E.C Traugott, A. Meulen, J.S. Reilly, and C.A. Ferguson (eds) *On Conditionals,* pp. 77-99. Cambridge: Cambridge University Press.

Firth, J.R. (1951/1957) Modes of Meaning, in *Papers in Linguistics 1934-1951,* pp. 190-215. London: Oxford University Press.

Firth, J.R. (1968) A Synopsis of Linguistic Theory, in F.R Palmer. (ed.) *Selected Papers of J.R. Firth 1952-59*, pp. 168-205. London: Longmans.

Frazier, S. (2003). 'A corpus analysis of *would*-clauses without adjacent *if*-clauses'. *TESOL Quarterly* 37(3), 433-466.

Gabrielatos, C. (2003) Conditionals: ELT typology and corpus evidence. Paper presented at the 36th Annual BAAL Meeting, University of Leeds, UK, 4-6 September 2003. Also available online from http://eprints.lancs.ac.uk/140/.

Gabrielatos, C. (2005) Elliptical and discontinuous *if*-conditionals: Co-text, context, inference and intuitions. Paper presented at Corpus Linguistics 2005, University of Birmingham, 15-17 July 2005.

Gabrielatos, C. and T. McEnery (2005) Epistemic Modality in MA Dissertations, in P.A. Fuertes Olivera (ed.), *Lengua y Sociedad: Investigaciones recientes en lingüística aplicada*. Lingüística y Filología no. 61, pp. 311-331. Valladolid: Universidad de Valladolid. Also available online from http://eprints.lancs.ac.uk/102/01/Epistemic_modality_in_MA_dissertations.pdf

Gries, S.Th. & A. Stefanowitsch (2004) 'Extending collostructional analysis: A corpus-based perspective on 'alternations'.' *International Journal of Corpus Linguistics* 9(1), 97-129.

Halliday, M.A.K. (1994, 2nd ed.) *An Introduction to Functional Grammar*. London: Edward Arnold.

Hoey, M. (1991) *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Hoye, L. (1997). *Adverbs and Modality in English*. London: Longman

Huddleston, R. and G.K. Pullum (2002) *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Hundt, M., A. Sand and R. Siemund (1998) *Manual of Information to Accompany the Freiburg-LOB Corpus of British English ('FLOB')*. Freiburg: Englisches Seminar, Albert-Ludwigs-Universität Freiburg.

Lyons, J. (1977). *Semantics*. Cambridge: Cambridge University Press.

Palmer, F.R. (1990, 2nd ed.) *Modality and the English Modals*. Cambridge: Cambridge University Press.

Palmer, F.R. (2003). Modality in English: Theoretical, descriptive and typological issues, in R. Facchinetti, M. Krug and F.R. Palmer (eds.) *Modality in Contemporary English,* pp. 1-17. Berlin: Mouton de Gruyter.

Partington, A. (1998) *Patterns and Meanings: Using corpora for English language research and teaching*. Amsterdam: John Benjamins

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

Rayson, P. (2003) Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. Ph.D. thesis, Lancaster University. Also available online from http://juilland.comp.lancs.ac.uk/computing/users/paul/phd/phd2003.pdf

Rayson, P. (2007) Wmatrix: A web-based corpus processing environment. Computing Department, Lancaster University. Available online from http://www.comp.lancs.ac.uk/ucrel/wmatrix/.

Rayson, P. and R. Garside (2000) Comparing corpora using frequency profiling, in proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics, 1-8 October 2000, Hong Kong, 1 - 6. Also available online from http://www.comp.lancs.ac.uk/computing/users/paul/publications/rg_acl2000.pdf

Rayson P., D. Berridge and B. Francis (2004) Extending the Cochran rule for the comparison of word frequencies between corpora. In Purnelle G., Fairon C., Dister A. (eds.) *Le Poids des Mots: Proceedings of the 7th International Conference on Statistical Analysis of Textual Data, Vol. II (JADT 2004), Louvain-la-Neuve, Belgium, March 10-12, 2004*, Presses Universitaires de Louvain, 926 - 936. Also online: http://www.comp.lancs.ac.uk/computing/users/paul/publications/rbf04_jadt.pdf

Scott, M. (1997) 'PC analysis of key words-and key key words'. *System*, 25, 233-245.

Scott, M. (1998). *WordSmith Tools Version 3*. Oxford: Oxford University Press.

Sperberg-McQueen, C.M. and L. Burnard (2007). TEI P5: Guidelines for electronic text encoding and interchange. The Text Encoding Initiative Consortium. Available online from http://www.tei-c.org/P5/Guidelines/AI.html

Stefanowitsch, A. and S.Th. Gries (2003) 'Collostructions: Investigating the interaction of words and constructions'. *International Journal of Corpus Linguistics* 8(2), 209-243.

Stubbs, M. (2001) *Words and Phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.

Sweetser, E.E. (1990) *From Etymology to Pragmatics: Metaphorical and cultural aspects of semantic structure*. Cambridge: Cambridge University Press.

Appendices

Appendix 1: Sample * BNCSw: positive modal keywords

Keyword	Sample		BNCSw		LL	*p*≤
	Freq.	/mil.	Freq.	/mil.		
would	152	5859.7	2,338	2134.3	110.26	0.0000000000
might	47	1811.9	468	427.2	61.40	0.0000000000
can	106	4086.4	2,079	1897.9	47.67	0.0000000000
may	72	2775.6	1,252	1142.9	41.75	0.0000000000
will	132	5088.7	3,107	2836.3	36.52	0.0000000000
should	66	2544.3	1,366	1247.0	26.07	0.0000003269
shall	21	809.6	223	203.6	25.49	0.0000004413
liable	6	231.3	10	9.1	24.50	0.0000007411
want	27	1040.9	379	346.0	22.61	0.0000019777
could	67	2582.9	1,494	1363.8	21.72	0.0000031567
think	27	1040.9	491	448.2	14.29	0.0001564376
know	31	1195.1	627	572.4	12.94	0.0003210141
need	27	1040.9	527	481.1	12.25	0.0004641809
probably	14	539.7	193	176.2	12.05	0.0005188992
imply	4	154.2	14	12.7	11.72	0.0006187793
hopefully	3	115.7	7	6.4	10.71	0.0010656740
able	15	578.3	235	214.5	10.52	0.0011839127
must	41	1580.6	1,018	929.3	9.50	0.0020535023
necessary	14	539.7	224	204.5	9.46	0.0020991049
thinks	5	192.8	34	31.1	9.39	0.0021859028
unable	6	231.3	51	45.6	9.23	0.0023861809
knew	12	462.6	184	168.0	8.72	0.0031392924
comply	3	115.7	11	10.6	8.57	0.0034248075
doubted	2	77.1	3	2.7	8.48	0.0035975266
evidence	10	385.5	142	129.6	8.23	0.0041298065
feel	11	424.1	175	159.8	7.51	0.0061432803
obligation	3	115.7	14	12.7	7.41	0.0064834715
required	12	462.6	211	192.6	6.80	0.0091246543

Appendix 2: Sample * FLOB: positive modal keywords

Positive Keyword	Sample		FLOB		LL	*p*≤
	Freq.	/mil.	Freq.	/mil.		
would	152	5859.7	2,308	2232.8	101.79	0.0000000000
will	132	5088.7	2,284	2209.6	68.72	0.0000000000
can	106	4086.4	1,772	1714.3	59.17	0.0000000000
may	72	2775.6	1,208	1168.7	39.87	0.0000000000
might	47	1811.9	641	620.1	37.56	0.0000000000
should	66	2544.3	1,115	1078.7	36.04	0.0000000001
shall	21	809.6	197	190.6	27.40	0.0000001623
liable	6	231.3	18	17.4	18.42	0.0000177183
must	41	1580.6	803	776.9	16.04	0.0000620163
want	27	1040.9	439	424.7	15.89	0.0000671768
could	67	2582.9	1,569	1517.9	15.54	0.0000808949
need	27	1040.9	464	448.9	14.22	0.0001624178
imply	4	154.2	17	16.4	10.07	0.0015055711
necessary	14	539.7	206	199.3	9.87	0.0016769837
comply	3	115.7	9	8.7	9.21	0.0024075144
required	12	462.6	180	174.1	8.19	0.0042222887
hopefully	3	115.7	12	11.6	7.84	0.0051039042

Positive Keyword	Sample		FLOB		LL	*p*≤
	Freq.	/mil.	Freq.	/mil.		
probably	14	539.7	239	231.2	7.47	0.0062589230
unable	6	231.3	59	57.1	7.42	0.0064358436
willing	4	154.2	26	25.2	7.41	0.0064944336
think	27	1040.9	604	584.3	7.27	0.0070050098

Appendix 3: *If*-BNCw * BNCSw: positive modal keywords

Keyword	*If*-BNCw		BNCSw		LL	*p*≤
	Freq.	/mil	Freq.	/mil.		
would	35,782	6585.49	2,409	2223.11	3,698.5	0.0000000000
can	22,708	4179.29	2,272	2096.68	1,190.7	0.0000000000
will	27,253	5015.77	3,110	2870.02	1,012.6	0.0000000000
could	16,588	3052.93	1,602	1478.38	940.6	0.0000000000
might	6,803	1252.06	469	432.81	679.4	0.0000000000
may	11,832	2177.62	1,256	1159.08	536.1	0.0000000000
want	5,353	985.19	379	349.75	515.9	0.0000000000
ll	5,152	948.20	396	365.44	441.3	0.0000000000
know	5,937	1092.67	628	579.54	271.2	0.0000000000
necessary	2,907	535.02	224	206.72	248.0	0.0000000000
should	10,260	1888.30	1,383	1276.28	206.8	0.0000000000
wish	1,643	302.39	93	85.78	206.0	0.0000000000
able	2,790	513.48	235	216.87	205.9	0.0000000000
need	4,636	853.23	528	487.26	172.7	0.0000000000
think	4,111	756.61	491	453.11	131.9	0.0000000000
possible	3,442	633.48	409	377.44	112.3	0.0000000000
knew	1,929	355.02	184	169.80	112.1	0.0000000000
probably	1,951	359.07	193	178.11	104.7	0.0000000000
claim	1,218	224.17	97	89.52	98.7	0.0000000000
wants	852	156.81	60	55.37	82.7	0.0000000000
shall	2,018	371.40	223	205.79	82.0	0.0000000000
doubt	1,270	233.74	117	107.97	79.3	0.0000000000
chance	1,335	245.70	127	117.20	78.0	0.0000000000
certain	1,648	303.31	174	160.57	75.6	0.0000000000
must	6,788	1249.30	1,026	946.83	73.4	0.0000000000
ought	543	99.94	33	30.45	63.1	0.0000000000
perhaps	2,086	383.92	254	234.40	62.7	0.0000000000
required	1,809	332.94	211	194.72	62.6	0.0000000000
certainly	1,247	229.50	128	118.12	61.4	0.0000000000
evidence	1,319	242.76	142	131.04	57.5	0.0000000000
wishes	497	91.47	31	28.61	56.1	0.0000000000
surely	692	127.36	66	60.91	40.2	0.0000000000
likely	2,563	471.71	365	336.84	39.4	0.0000000000
intention	386	71.04	27	24.92	37.8	0.0000000000
willing	511	94.05	43	39.68	37.8	0.0000000000
actually	1,011	186.07	115	106.13	37.8	0.0000000000
allowed	1,017	187.17	118	108.89	35.8	0.0000000001
reasonably	379	69.75	28	25.84	34.4	0.0000000016
wished	438	80.61	36	33.22	33.7	0.0000000035
obligation	252	46.38	14	12.92	32.2	0.0000000110
unable	540	99.38	51	47.06	32.0	0.0000000122
believe	1,568	288.58	211	194.72	31.8	0.0000000141
unlikely	592	108.95	59	54.45	31.2	0.0000000201
impossible	575	105.83	58	53.52	29.5	0.0000000533
assumption	259	47.67	17	15.69	27.5	0.0000001538

Keyword	*If*-BNCw		BNCSw		LL	*p*≤
	Freq.	/mil	Freq.	/mil.		
doubted	123	22.64	3	2.77	27.1	0.0000001857
belief	391	71.96	35	32.30	25.8	0.0000003723
knows	640	117.79	72	66.44	24.7	0.0000006670
comply	193	35.52	11	10.15	24.0	0.0000009483
believes	332	61.10	29	26.76	23.0	0.0000016587
intended	502	92.39	53	48.91	23.0	0.0000015948
suggest	588	108.22	66	60.91	22.8	0.0000017556
expect	864	159.01	109	100.59	22.8	0.0000017558
deemed	221	40.67	15	13.84	22.5	0.0000020956
assume	411	75.64	41	37.84	21.6	0.0000033077
possibly	485	89.26	52	47.99	21.4	0.0000037713
potentially	153	28.16	8	7.38	20.7	0.0000052984
seem	982	180.73	131	120.89	20.7	0.0000053735
doubtful	192	35.34	13	12.00	19.6	0.0000094904
obviously	556	102.33	66	60.91	18.2	0.0000199498
assuming	177	32.58	12	11.07	18.0	0.0000215316
obliged	217	39.94	17	15.69	18.0	0.0000216526
thinks	410	75.46	44	40.60	18.0	0.0000217893
compelled	67	12.33	1	0.92	17.5	0.0000282487
unsure	97	17.85	4	3.69	16.0	0.0000646262
desirable	217	39.94	19	17.53	14.9	0.0001112545
prevent	539	103.62	69	63.68	13.5	0.0002422281
presumably	218	40.12	22	20.30	11.2	0.0008325512
permitted	223	40.04	23	21.23	10.8	0.0009916140
necessarily	360	66.26	44	40.60	10.7	0.0010856490
seems	1,279	235.39	202	186.41	10.0	0.0015271711
ability	491	90.37	67	61.83	9.3	0.0022398166
possibility	533	98.10	74	68.29	9.3	0.0022492136
compel	25	4.60	0	0	9.1	0.0025665928
require	680	125.15	100	92.28	8.7	0.0031427261
unwilling	94	17.30	7	6.46	8.4	0.0036839675
expecting	146	26.87	14	12.92	8.4	0.0037795191
authorised	131	23.93	12	11.07	8.3	0.0040314551
granted	384	70.67	51	47.06	8.3	0.0040748352
permission	309	56.87	39	35.99	8.2	0.0042947028
recommended	199	36.62	22	20.30	8.1	0.0044977809
tendency	158	20.08	16	14.77	8.0	0.0046054190
essential	665	122.39	99	91.36	7.9	0.0048998352
plausible	75	13.80	5	4.61	7.8	0.0051898067
chances	318	58.53	41	37.84	7.7	0.0054403744
obligatory	21	3.86	0	0	7.6	0.0057160771
authorized	21	3.86	0	0	7.6	0.0057160771
oblige	36	6.63	1	0.92	7.5	0.0062149893
preferably	82	15.09	6	5.54	7.5	0.0060187350
requires	365	67.18	49	45.49	7.5	0.0062349923
validity	112	20.61	10	9.23	7.4	0.0064102570
potential	633	116.50	96	88.59	6.7	0.0097284419

Appendix 4: *If*-BNCw * FLOB: positive modal keywords

Keyword	*If*-BNCw		FLOB		LL	*p*≤
	Freq.	/mil	Freq.	/mil.		
would	35,782	6585.49	2,308	2261.36	3,433.7	0.0000000000
will	27,253	5015.77	2,284	2237.85	1,734.5	0.0000000000
can	22,708	4179.29	1,772	1736.19	1,640.9	0.0000000000

Keyword	*If*-BNCw		FLOB		LL	*p*≤
	Freq.	/mil	Freq.	/mil.		
could	16,588	3052.93	1,569	1537.30	818.1	0.0000000000
may	11,832	2177.62	1,208	1183.59	481.9	0.0000000000
want	5,353	985.19	439	430.13	353.6	0.0000000000
should	10,260	1888.30	1,115	1092.47	349.3	0.0000000000
might	6,803	1252.06	641	628.05	338.0	0.0000000000
need	4,636	853.23	464	454.62	198.7	0.0000000000
wish	1,643	302.39	95	93.08	179.1	0.0000000000
must	6,788	1249.30	803	786.77	173.5	0.0000000000
possible	3,442	633.48	339	332.15	153.7	0.0000000000
likely	2,563	471.71	240	235.15	129.1	0.0000000000
know	5,937	1092.67	767	751.50	104.8	0.0000000000
able	2,790	513.48	294	288.06	103.9	0.0000000000
shall	2,018	371.40	197	193.02	92.1	0.0000000000
required	1,809	332.94	180	176.36	78.7	0.0000000000
were	20,121	3703.16	3,209	3144.16	77.4	0.0000000000
entitled	660	121.47	40	39.19	68.1	0.0000000000
willing	511	94.05	26	25.47	63.7	0.0000000000
wanted	1,927	354.65	215	210.66	60.3	0.0000000000
prefer	513	94.41	29	28.41	57.3	0.0000000000
chance	1,335	245.70	134	131.29	56.8	0.0000000000
judgment	361	66.44	15	14.70	53.6	0.0000000000
sure	1,851	340.67	218	213.60	48.1	0.0000000000
probably	1,951	359.07	239	234.17	43.5	0.0000000000
desired	270	49.69	11	10.78	40.7	0.0000000000
doubt	1,270	233.74	148	145.01	34.3	0.0000000019
require	680	125.15	65	63.69	32.6	0.0000000082
demand	856	157.54	91	89.16	31.1	0.0000000218
wished	438	80.61	35	34.29	30.3	0.0000000343
advice	839	154.41	91	89.16	28.7	0.0000000818
surely	692	127.36	74	72.50	24.7	0.0000006691
obligation	252	46.38	16	15.68	24.6	0.0000007169
deemed	221	40.67	13	12.74	23.6	0.0000011687
intend	245	45.09	16	15.68	23.0	0.0000015961
doubtful	192	35.34	12	11.98	19.1	0.0000124849
allow	866	159.38	107	104.84	18.6	0.0000158652
unable	540	99.38	59	57.81	18.1	0.0000214326
entitlement	71	13.07	1	0.98	17.6	0.0000273597
purpose	695	127.91	83	81.32	17.1	0.0000350081
advise	215	39.57	16	15.68	16.7	0.0000429299
certain	1,648	303.31	237	232.21	15.8	0.0000704088
chances	318	58.53	30	29.39	15.7	0.0000726175
permitted	223	41.04	18	17.64	15.1	0.0000993929
unlikely	592	108.95	70	68.59	15.1	0.0000996674
needed	1,377	253.43	196	192.04	14.1	0.0001688611
allowed	1,017	187.17	140	137.17	12.9	0.0003363111
suggest	588	108.22	74	72.50	11.7	0.0006279354
probabilities	33	6.07	0	0	11.4	0.0007499971
advisable	95	17.48	5	4.90	11.4	0.0007164479
promise	349	64.23	39	38.21	10.9	0.0009798151
desirable	217	39.94	21	20.58	10.1	0.0014723970
obligations	155	21.77	13	12.74	9.8	0.0017328333
entails	45	8.28	1	0.98	9.5	0.0020043203
probable	115	21.17	9	8.80	8.2	0.0040896023
intention	386	71.04	48	47.03	8.1	0.0044707572
unsure	97	17.85	7	6.86	7.9	0.0048939781
preferable	87	16.01	6	5.88	7.6	0.0058721923

Keyword	*If*-BNCw		FLOB		LL	*p*≤
	Freq.	/mil	Freq.	/mil.		
impossible	575	105.83	79	77.40	7.4	0.0067015006
doubted	123	22.64	11	10.78	6.9	0.0088623865
advised	223	41.04	25	24.49	6.9	0.0087655494
possibility	533	98.10	73	71.52	6.9	0.0084355818