A corpus-driven approach to modal auxiliaries and their didactics

Ute Römer
University of Cologne

The paper presents an example of the indirect use of corpora in language pedagogy. It centres on a comparative analysis of modal auxiliaries, their distribution, meanings, and contexts, in spoken British English corpus data and in selected texts from EFL textbooks. A focus lies on the differences observed between authentic English as used in natural communicative situations and the kind of synthetic English that pupils are often confronted with in the classroom. It is argued that, if taken seriously, corpus evidence can contribute to an improvement of teaching materials and that it is essential, especially in pedagogical contexts, to pay more attention to frequent phenomena and typical patterns of used language.

1. Introduction

Modal auxiliaries constitute one of the grammatical problem areas in teaching English as a foreign language to German learners and probably also to learners of other nationalities. This observation forms the basis of the research project reported on in this paper, which deals with the use of modals in spoken English and in English language teaching.[1]

The purpose of this paper is to summarise the results of an investigation which centred on a corpus-driven analysis of the nine central modal verbs as listed in Quirk et al. (1985: 137): *can*, *could*, *may*, *might*, *will*, *would*, *shall*, *should*, and *must*, plus the modal *ought to* in contemporary spoken English and in one of the major German textbook series used in the EFL classroom.[2] The leading questions were: “Is the English taught at German schools identical to the English which is used by native speakers?” and “How extensively does the grammar of ‘school’ English differ from authentic spoken English?”

2. Modals in spoken British English (BNC analysis)

Trying to find answers to the questions above and starting from the assumption that pupils should learn a type of English which is really used and understood by native speakers nowadays, first of all data from the 10-million-word spoken part of the British National Corpus (BNC) was collected and analysed.[3] The results of this corpus analysis were meant to show how often, in which contexts, and in which meanings the different polysemous modals are used in spoken British English. One reason for choosing exclusively spoken material was the fact that modal auxiliaries occur more frequently in spoken than in written English (cf. Quirk et al. 1985: 136). However, the main reason was the pre-eminence of spoken language in English lessons as demanded in the *Richtlinien* for teaching English as a foreign language in Germany.[4]

Using SARA's, (the BNC concordance program's) part-of-speech (POS) query option, queries on the different forms (i.e. positive, full negative, and contracted negative) of the ten modal verbs mentioned above were carried out. The query builder allows the researcher to combine different types of corpus searches, such as POS queries and SGML queries. In the present study an SGML query had to be used to make sure that exclusively spoken texts were searched (cf. Aston & Burnard 1998: 100–108). For the SGML element <catRef> (category reference) the attribute spoken_type with its values *dialogue* and *monologue* was selected. For each verb form a random set of 200 concordance lines was downloaded for further manual analysis. The "random set" box is one of the options available in SARA's "download hits" window. It is also possible to save the initial *n* solutions, all the solutions found, or only one solution per text (cf. Aston & Burnard 1998: 66–67). The main reason for choosing the random set option here was to achieve a maximum distribution of downloaded concordance lines over all the texts in the spoken part of the BNC.

2.1 Frequency analysis

With the concordancer SARA it was also possible to get frequency information about the verb forms in question. Frequencies can be very important as they show us which words or structures are central in a language. Thus they can help with decisions about what to include in teaching materials and what not. On the basis of frequency data it is possible to see which modals are the most important ones and should thus be dealt with first in EFL teaching. Figure 1 shows the frequency distribution of the central modal auxiliaries (plus *ought to*) including negative forms in the spoken part of the BNC. As can be seen in

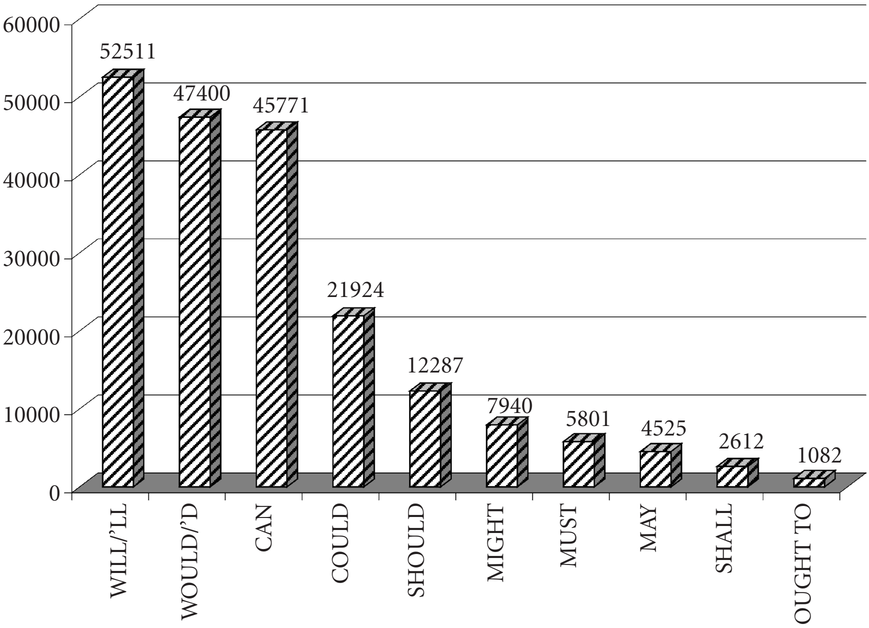

Figure 1. Frequency of modals in BNC spoken

this diagram, the three most frequent modals by far are *will/'ll*, *would/'d*, and *can* each with more than 45,000 occurrences in the spoken part of the BNC, followed by the modal *could* with 21,924 occurrences. The frequencies of the remaining modals are much lower, ranging from 12,287 occurrences (*should*) to 1,082 occurrences (*ought to*).

2.2 Different meanings analysis

Having collected the frequency data, the saved data sets (200 concordance lines for each verb form) were analysed manually with regard to different meanings and co-occurrences; i.e. the syntactic and semantic surroundings of the verbs in question were examined. The results of the different meanings approach were summarised in diagrams of the following kind (Figure 2), indicating for each modal the distribution of its different functions in spoken English.

For the modal auxiliary *can* three different meanings were found in the corpus material with the following frequency distribution: 36% ability, 31.5% possibility and 23.5% permission. The following are example sentences

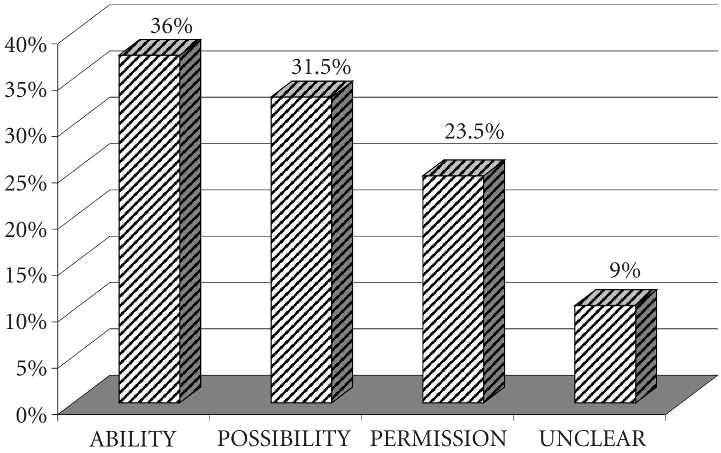

Figure 2. Different meanings of *can* (BNC spoken)

from the BNC in which *can* expresses an ability (1), a possibility (2), and a permission (3).

(1) *and when it gets to the chasing teddy bears you've got to run as fast as you* can, *so you'd better move out of the way.* (BNC, KBW, 16264)

(2) *Yeah, the whole sentence is, is a constituent itself er it and* can *be a constituent of larger sentences obviously.* (BNC, HE0, 187)

(3) Can *I have an apple please?* (BNC, FLY, 355)

In addition, there was a considerable number of indeterminate or unclear cases, where it was impossible to make a decision about the function of the modal, either because the sentence was fragmentary or because there simply was not enough context, as for example in (4) and (5).

(4) *I just offered to erm in Tech class* can *you?* (BNC, KPG, 5404)

(5) can *it in those rooms with the dogs* (BNC, KE6, 10457)

The percentages of the different meanings distribution for *can*, *could*, *may*, *might*, *will*, *would*, *shall*, *should*, *ought to* and *must* can be found in Table 1 below.

Table 1. Different meanings distribution of modals (BNC spoken)

	ability	possibility	permission	hypothet. meaning	prediction	volition	obligation/ advice	inference/ deduction	unclear
can	36%	31.5%	23.5%						9%
could	34%	41.5%	3.5%	14.5%					6.5%
may		83%	13%						4%
might		95%	3.5%						1.5%
will					87.5%	7.75%			4.75%
would				28.5%	50.5%	15.5%			5.5%
shall					31%	65%			4%
should				30%			62.5%		7.5%
ought to				16%			79%		5%
must							52%	39%	9%

2.3 Co-occurrence analysis

Some crucial observations could also be made in the analysis of co-occurrences of the modal verbs. Among the features examined were negations, and the occurrence of the different modals in questions, set phrases, if-clauses, and passive constructions. In the following, some of the most interesting findings are listed.

The highest percentages of negations were found with *can* (27.8%) and *could* (17.6%). Contracted forms (e.g. *can't*, 94.25%) are in all cases much more frequent than full forms (e.g. *cannot*, 5.75%). In his empirical study on modal verbs Mindt experienced a similar tendency but found much higher figures for *can* (40%) and *could* (32%) in negative contexts (Mindt 1995: 176). An explanation for these differences may lie in the different types of corpora used in the two studies. From Mindt's descriptions it does not become clear, however, what exactly his corpus consists of.

Another observation that could be made is that *shall* is used very frequently in questions (36.5% of the sentences examined), e.g. in

(6) *Well* shall *I tell you what you were going to ask?* (BNC, HMP, 112)

This finding is in accordance with the results of Mindt's analysis where *shall* tops the list of modals in interrogative contexts (1995: 177).

In 85% of the analysed concordance lines the modal *shall* is accompanied by a first person subject ("I" or "we"):

(7) *We* shall *see him says John, and we shall be like him.* (BNC, J8Y, 336)

Compared with the other modals *should* is more often (in 10% of the cases) found in passive constructions, as in the following BNC example:

(8) *But but first of all I would like to say the officers of the agencies really* should *be congratulated on*. (BNC, J9D, 145)

May is quite frequent in if-clauses (19%). In this context the occurrence of the modal in the fixed expression "if I *may*" is worth mentioning. This phrase was found in 26.3% of all if-clauses with *may,* as for example in

(9) *If I* may *come back Mr Chairman an and er express a view on behalf of Darcy*. (BNC, J42, 105)

3. Modals in EFL teaching (textbook analysis)

The second major part of this investigation is an analysis of the treatment of modal auxiliaries in *Learning English Green Line* (Vols. 1–6), a German textbook series widely used in the EFL classroom in grammar schools, and in the *Learning English Grundgrammatik*, an introductory grammar German pupils are supposed to work with and/or use as a reference grammar.[5] As an electronic version of the textbook series was not available and the analysis thus had to be carried out manually, the six volumes of *Green Line* could not be analysed completely and were not regarded as a pedagogical corpus. Instead, several texts (32 altogether) from those textbook units which mentioned one or more modal auxiliaries in their grammar sections were examined thoroughly. The 32 texts were treated as a sample of EFL textbook language – the kind of language pupils are exposed to in the EFL classroom – enabling a comparison of textbook English with authentic language material collected from a general corpus.

The major aim of this textbook analysis was to find out whether the use of modals in *Green Line* was an accurate representation of the actual language use, i.e. of the occurrence of modal verbs in the spoken part of the BNC.[6] Basically the same types of investigations were carried out as in the corpus analysis. A frequency count including an examination of the order in which the modals are introduced in *Green Line* (and which may reveal something about their prominence in grammar teaching) was followed by a different meanings and a co-occurrence analysis. In addition to several units from the textbooks (Vols. 1–6) all *Green Line* grammar sections and the *Learning English Grundgrammatik* were included in the analysis.

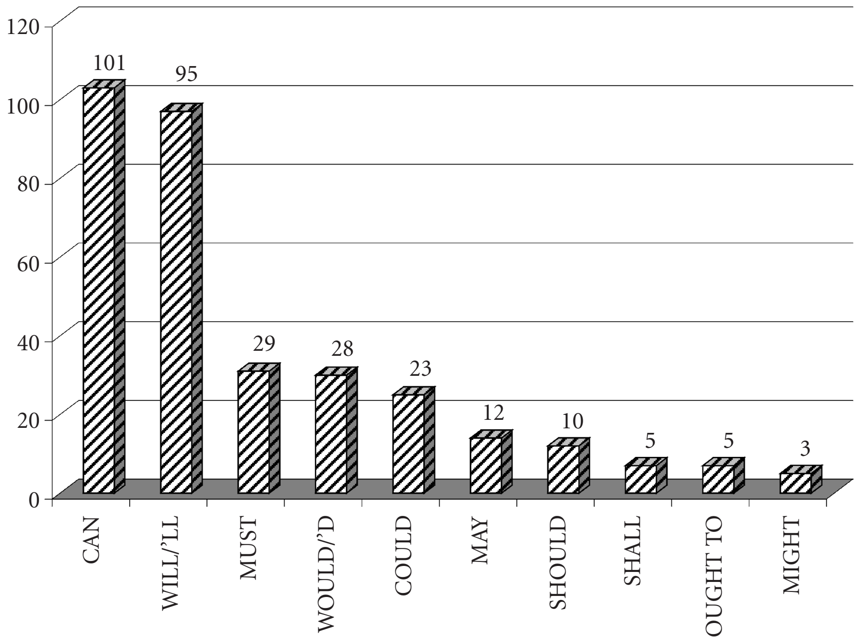

Figure 3. Frequency of modals in *Green Line*

3.1 Frequency analysis

Figure 3 shows the results of the frequency counts of the analysed texts from *Green Line* 1–6.

As we can see in this diagram, there is a huge frequency gap between *can* and *will/'ll* on the one hand and the other eight modals on the other hand. Thus in the textbook texts I found 101 occurrences of *can* and 95 occurrences of *will* and *'ll* but only between 3 and 29 instances of *could*, *would/'d, may, might, shall, should, ought to*, and *must*.

3.2 Different meanings analysis

The results of the different meanings approach were collected in diagrams comparable to those that were used in the BNC data evaluation. In the diagram for *can* shown below (Figure 4) it becomes clear that in *Green Line* the modal is most frequently used to express an ability (52.5% of the cases). The meanings "possibility" and "permission" (24.7% and 22.8%) seem to be less important.

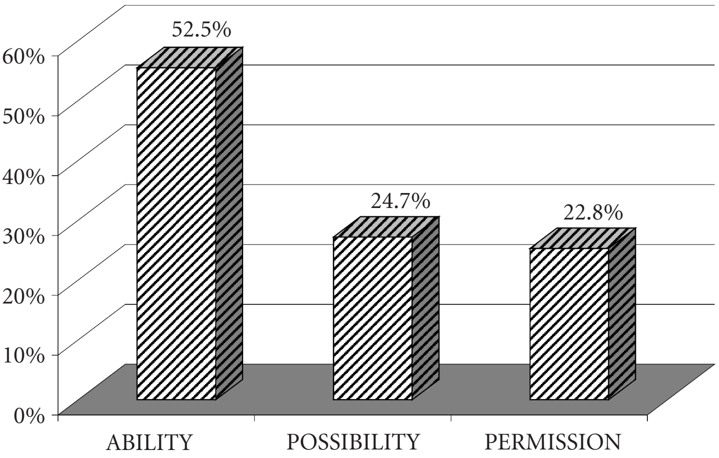

Figure 4. Different meanings of *can* (*Green Line*)

Analyses of that kind were also conducted for the other central modals and *ought to*. The percentages for all verbs under investigation have been collected in Table 2 below. Since we are dealing with invented examples here in which the sentence contexts are always constructed unambiguously and as there was always enough context available, there are no semantically indeterminate cases to be found in the textbook texts.

Table 2. Different meanings distribution of modals (*Green Line*)

	ability	possibility	permission	hypothet. meaning	prediction	volition	obligation/ advice	inference/ deduction	unclear
can	52.5%	24.7%	22.8%						
could	78.3%	13%		8.7%					
may		58.3%	41.7%						
might		100%							
will					82.1%	17.9%			
would				39.3%	28.6%	32.1%			
shall						100%			
should				20%			80%		
ought to				20%			80%		
must							93.1%	6.9%	

3.3 Co-occurrence analysis

Interesting findings from the co-occurrence examination of *can*, *could*, *may*, *might*, *will*, *would*, *shall*, *should*, *must*, and *ought to* in *Green Line* texts are the following ones:

Very high incidence of negation is found with *can* (36.7%), *may* (33.3%), *could* (21.7%), and *must* (20.7%). On the other hand, there are no negative forms of *might*, *shall*, and *ought to*. The modal *shall* is found exclusively in questions (100%). *Could* (30.4%) and *may* (25%) also show rather high percentages of questions. *May* does not occur in if-clauses; e.g. there is not a single instance of the set phrase "if I *may*" in any of the textbook texts. Very high percentages of if-clauses are found with *would* (35.7%) and *might* (33.3%), and the modal *shall* is always used with a first person singular subject (100%).

4. Comparison: The use of modals in "real" English and in "school" English

The third part of this investigation consists of a comparison of the results of corpus analysis (BNC spoken) and textbook analysis (*Green Line*). Differences of the findings were pointed out again with regard to frequencies, different meanings and co-occurrences. This comparison made it clear that there are huge discrepancies between the use of modal auxiliaries in authentic English and in the English taught in German schools.

The frequency distribution of the modals in *Green Line*, for instance, differs quite a lot from the one found in the spoken part of the BNC. As we can see in Figure 5, the modals *will/'ll*, *can,* and *must* are overused in *Green Line* while there is an underuse of *would/'d*, *could*, *should,* and *might*. This underuse is especially significant in the case of *would/'d*. In the BNC the modal (including its contracted form *'d*) is the second most frequent one with 23.48%, whereas in the textbook series it only comes in fifth place (relative frequency: 9%).

More differences can be found if we compare the diagrams showing the different meanings distribution of each modal verb for the spoken part of the BNC and for the textbooks. For *can* and *could* expressing an ability for instance the percentages in *Green Line* (52.5% and 78.3%) are much higher than in the BNC (36% and 34%). In the sentences from the BNC *could* more frequently expresses a possibility (in 41.5% of the cases) than an ability. Concerning *may* we get a much higher share of the permission meaning in *Green Line* (41.7%) than in spoken English (13%), although the modal is mainly used to convey

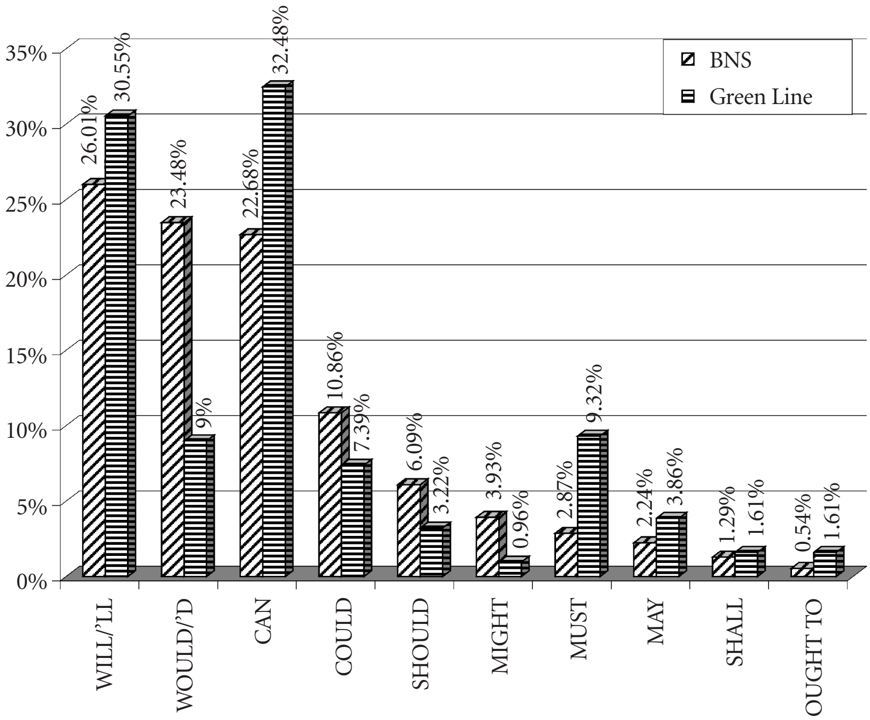

Figure 5. Relative frequencies of modals in BNC and *Green Line*

the meaning "possibility" in actual language use (83%). In the textbooks *might* is exclusively used to express a possibility. Even if this is also the most frequent meaning in the BNC data, there are some instances of the modal where it is used to ask for permission (3.5%). *Will* and *would* are less frequently used in *Green Line* in their prediction meaning than they are in "real" English. Another striking phenomenon is that *shall* is never used to make predictions in the textbooks and school grammars although this is an important meaning of the modal in the BNC (with 31%). There is another mismatch between "school" English and "real" English concerning *must*. Although the modal expresses an inference/deduction in 39% of the BNC concordance lines analysed, this meaning is only expressed in very few sentences in *Green Line (6.9%)*.

Beside the differences related to the polysemy of the modal verbs, a couple of interesting observations could be made concerning the modals' syntactic surroundings in corpus and textbook data. On the whole, we can say that the percentages of modal negation are much higher in *Green Line* than in spoken English. It is striking, however, that some modals (*might*, *shall*, *ought to*) are not negated at all in the textbook texts analysed. There are also some differences with respect to the percentages of questions. In *Green Line might*, *must*, and *ought to* never occur in questions, whereas *could*, *may*, *will*, *shall,* and *should* are much more frequently used in questions in the textbooks than in the spoken part of the BNC. In the case of *shall* the textbooks even imply that this modal is exclusively used in questions despite the fact that 63.5% of the BNC concordance lines are statements. The shares of if-clauses in *Green Line* as compared to the BNC data are much too high concerning *might* and *would*, but the other modals are too rarely, or even not at all, used in if-clauses. Another mismatch between corpus and textbooks is the non-occurrence of set phrases and word clusters like "if I *may*", "*must* admit", or "*might* as well" in the latter.

5. Suggestions for the improvement of teaching materials

From these and other discrepancies between corpus and textbook data some consequences can be drawn and suggestions for an improvement of teaching materials on the basis of the findings from this corpus-driven approach can be made. The central questions are: "How can we come closer to achieving a high degree of authenticity in English language teaching as called for in the *Richtlinien*?" and "How can we reach the aim of teaching pupils an English which is comparable to native speaker English?".

Assuming that the collection of EFL textbook texts used in the present study indicates the kind of English prioritised in English language teaching in German schools, a couple of changes concerning the use of modal verbs might be helpful to make the English that is taught more natural and native-like. First of all, I would suggest changing the order in which the modals are introduced from

> *can → must → may → could → would/'d → should → will/'ll → shall → ought to → might*

to

> *will/'ll → would/'d → can → could → should → might → must → may → shall → ought to,*

an order which is based on corpus findings. In my opinion, other things being equal the more frequent verbs (i.e. the more important verbs, at least from a communicative point of view) should be introduced at an earlier stage in the learning process than the less frequent ones. Secondly, I consider it important, if we want to enable pupils to communicate successfully, not to leave out some of the different meanings modal auxiliaries can have, e.g. the permission meaning of *might* and *could*, and to stress the inference/deduction meaning of *must*. To achieve a higher degree of authenticity, we might want to use similar proportions of the different senses of a polysemous verb in the English used in schools as found in the English used in real-life situations. Hence it would be a move in the right direction to present *can*, *could*, and *may* more frequently in contexts where they express possibilities.

As most of the modals were found to be used too frequently in negative contexts in *Green Line* texts, this overuse ought to be avoided if possible. This is not supposed to mean that some of the modals' negative forms are to be excluded from the teaching materials. It is, however, important to mention how often each verb occurs in affirmative and negative contexts.

Other changes that might lead to a higher degree of authenticity and thus to an improvement of textbooks like *Green Line* are to use *might* and *would* less frequently but the other modals more often in if-clauses, to mention the fact that *shall* does not only occur with a first person singular subject, and to avoid using *shall* exclusively in questions as it also occurs in statements. It may also be worth mentioning that some of the modals could be presented more frequently in tag-questions and in set expressions like "if I *may*" or "*might* as well" and that *should* and *must* could be used more often in passive constructions in the textbooks.

As many of the examples taken from *Green Line* sound rather unnatural, I would like to stress the importance of banishing invented sentences from textbooks and suggest to use preferably authentic material from a corpus instead. This approach to base English language teaching on real examples taken from corpora and to expose pupils to natural language was already formulated by Dave Willis in *The Lexical Syllabus*.[7] Willis gives a description of the *Collins COBUILD English Course*, a lexically-based course making extensive use of spontaneously produced examples and stresses the importance of an "exposure to authentic language materials" (Willis 1990: 46). Another supporter of the authenticity principle is John Sinclair who advises language teachers to "[p]resent real examples only" and considers it "most unwise to offer examples which are unattested, or to make major changes to actual instances" if example sentences are supposed to serve as models of English language usage (Sinclair 1997: 31).

Moreover it could be considered an improvement to present the modals as a group rather than treating them separately. In this context the differences to full verbs such as the so-called "NICE properties" ought to be stressed to give pupils a clearer picture of how modal verbs are used differently from other verbs.[8] Finally I would like to suggest to focus more on the connection between past-tense-modals and politeness, an important concept which is still very much neglected in the EFL classroom.

6. Conclusion

As we have been able to observe, the results of the analysis make it clear that corpus-driven approaches to language learning and teaching can be very helpful for teachers and schoolbook publishers and that, to cite Dieter Mindt, "corpus-based studies of grammar (...) can do much to bring the teaching of English into accordance with actual language use" (Mindt 1997: 50). The way the topic of "modal auxiliaries" is treated in English lessons in German grammar schools and the way *can*, *could*, *may*, *might*, *will*, *would*, *shall*, *should*, *ought to*, and *must* are presented in teaching materials differ considerably from the use of those verbs in contemporary spoken British English. At the expense of quite frequent and important aspects (e.g. certain modal meanings) which are underrepresented or sometimes even left out completely some minor and less important features of usage are over-emphasised in the textbooks. I fully agree with McEnery and Wilson, who say "... non-empirically based teaching materials can be positively misleading and [....] corpus studies should be used to inform the production of materials, so that the more common choices of usage are given more attention than those which are less common." (McEnery & Wilson 2001: 120) This postulate suggests that a lot of corpus-driven work still has to be done to reach the aim of enabling both pupils and teachers to learn and teach an English which is more authentic and closer to that of native speakers.

Notes

[1] A more detailed account of the investigations reported on in the present paper can be found in Römer 1999.

[2] For the convenience of the reader the modals under investigation (i.e. *can*, *could*, *may*, *might*, *will*, *would*, *shall*, *should*, *ought to*, and *must*) will always be italicised.

[3] The BNC is a fully part-of-speech-tagged corpus of over 100,000,000 words of both written and spoken British English. The spoken subcorpus makes up 10% of the whole corpus

and contains e.g. interviews, lectures, radio programmes, and everyday conversations. (cf. Aston & Burnard 1998: 31–36)

4. The *Richtlinien* serve as guide-lines which tell teachers what they are supposed to teach and how they are supposed to teach it. They are similar to the National Curriculum in Great Britain.

5. At the time they use the six volumes of *Green Line*, German pupils are between 10 and 16 years old. For most of the pupils English is the first foreign language.

6. According to one of the editors, *Green Line* is committed to a close representation of today's English (cf. Tegethoff 1984: L5).

7. The "lexical syllabus" was first described by A. Renouf and J. Sinclair in their 1988 article "A lexical syllabus for language learning" (in R. Carter & M. McCarthy (Eds.) *Vocabulary and Language Teaching*. London: Longman) and then further explained in Willis' book.

8. NICE is an acronym which stands for **n**egation, **i**nversion, **c**ode, **e**mphasis (cf. Coates 1983: 4). Full verbs and modal verbs differ considerably with respect to the NICE properties.

References

Aston, G. & L. Burnard (1998). *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Bald, W.-D. (1991). Modal auxiliaries: Form and function in texts. In C. Uhlig & R. Zimmermann (Eds.), *Anglistentag 1990. Proceedings* (pp. 348–361). Tübingen: Niemeyer.

Beile, W., A. Beile-Bowes, R. Hellyer-Jones, & P. Lampater (Eds.). (1984 [1989]). *Learning English. Green Line 1 [-6]. Unterrichtswerk für Gymnasien*. Stuttgart: Klett.

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

Hermerén, L. (1978). *On Modality in English. A Study of the Semantics of the Modals*. Lund: CWK Gleerup.

Leech, G. & J. Coates (1980). Semantic indeterminacy and the modals. In S. Greenbaum, G. Leech, & J. Svartvik (Eds.), *Studies in English Linguistics: For Randolph Quirk* (pp. 79–90). London: Longman.

McEnery, A. M. & A. Wilson (2001). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Mindt, D. (1987). *Sprache – Grammatik – Unterrichtsgrammatik*. Frankfurt: Diesterweg.

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

Mindt, D. (1996). A corpus-based empirical grammar of English modal verbs. In C. E. Percy, C. F. Meyer, & I. Lancashire (Eds.), *Synchronic Corpus Linguistics. Papers from the Sixteenth International Conference on English Language Research on Computerized Corpora (ICAME 16)* (pp. 133–141). Amsterdam: Rodopi.

Mindt, D. (1997). Corpora and the teaching of English in Germany. In A. Wichmann, S. Fligelstone, A. M. McEnery, & G. Knowles (Eds.), *Teaching and Language Corpora* (pp. 40–50). London: Longman.

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen (Ed.). (1993). *Richtlinien und Lehrpläne für das Gymnasium – Sekundarstufe I – in Nordrhein-Westfalen*. Frechen: Ritterbach.

Mitchell, K. W. (1988). Modals. In W.-D. Bald (Ed.), *Kernprobleme der englischen Grammatik – Sprachliche Fakten und ihre Vermittlung* (pp. 173–192). Berlin: Langenscheidt-Longman.

Palmer, F. R. (1990). *Modality and the English Modals*. London: Longman.

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

Römer, U. (1999). Das System englischer Modalverben und seine Stellung im Unterricht. Unpublished MA thesis, English Department, University of Cologne, Germany.

Sinclair, J. (1997). Corpus evidence in language description. In A. Wichmann, S. Fligelstone, A. M. McEnery, & G. Knowles (Eds.), *Teaching and Language Corpora* (pp. 27–39). London: Longman.

Tegethoff, E. (Ed.). (1984). *Learning English. Green Line 1. Lehrerbuch*. Stuttgart: Klett.

Ungerer, F., P. Pasch, P. Lampater, & R. Hellyer-Jones (1989). *Learning English Grundgrammatik. Ausgabe für Gymnasien*. Stuttgart: Klett.

Willis, D. (1990). *The Lexical Syllabus. A New Approach to Language Teaching*. London: Harper Collins.