Disentangling modal meanings with distributional semantics

Martin Hilpert and Susanne Flach

Université de Neuchâtel

Abstract

This paper investigates the collocational behavior of English modal auxiliaries such as *may* and *might* with the aim of finding corpus-based measures that distinguish between different modal expressions and that allow insights into why speakers may choose one over another in a given context. The analysis uses token-based semantic vector space modeling (Heylen et al. 2015, Hilpert and Correia Saavedra 2017) in order to determine whether different modal auxiliaries can be distinguished in terms of their collocational profiles. The analysis further examines whether different senses of the same auxiliary exhibit divergent collocational preferences. The results indicate that near-synonymous pairs of modal expressions, such as *may* and *might* or *must* and *have to*, differ in their distributional characteristics. Also different senses of the same modal expression, such as deontic and epistemic uses of *may*, can be distinguished on the basis of distributional information. We discuss these results against the background of previous empirical findings (Hilpert 2016, Flach in press) and theoretical issues such as degrees of grammaticalization (Correia Saavedra 2019) and the avoidance of synonymy (Bolinger 1968).

1 Introduction

How do speakers make a choice between modal expressions that are very similar in meaning, such as *That may be a good idea* and *That might be a good idea*? A basic assumption of cognitive-functional approaches to language is that any difference in form maps onto a difference in meaning (Bolinger 1968: 127, Goldberg 1995: 67, Croft 2001: 111). By implication, a choice such as the one between *may* and *might* should relate to differences that the two forms exhibit with regard to either propositional semantics, construal, register, discursive meaning, or social relations between speaker and hearer. Yet, when two de-contextualized utterances such as the ones above are presented side by side, it seems very hard to pinpoint what the exact nature of these differences is. Is there a contrast in the degree of certainty of the statement? Is one variant more formal than the other? Does one of the variants foreshadow a future discursive move that the other one does not? Intuitive assessments of these questions seem highly subjective and hence problematic.

In trying to address this problem, it is the aim of this paper to find corpus-based measures that distinguish between alternative modal expressions and that provide insights into why speakers may choose one expression over another in a given context. For this purpose, our analysis adopts the distributional hypothesis (Firth 1957, Turney and Pantel 2010), which holds that similarity in meaning is reflected in the way that linguistic elements co-occur with each other. Items with related meanings are thus expected to occur in similar contexts and in the vicinity of the same linguistic elements. The distributional hypothesis dovetails with work in usage-based Construction Grammar that views language as a network of symbolic units (Goldberg 1995, Diessel 2019). In that view, all linguistic forms are endowed with meaning, and linguistic knowledge is modeled exclusively in terms of form-

meaning pairings and connections between them. These connections represent different types, including categorical relations between abstract schemas and their more concrete instantiations, subpart links that represent partial overlap in meaning or form between constructions, and associative links that capture lexical collocations and interdependencies between morpho-syntactic patterns and lexical items (Hilpert 2019: 60-64). The latter kind is of particular relevance to the present study, since there is evidence to suggest that collocational preferences finely distinguish even between grammatical constructions that share some of their meaning potential (Gries and Stefanowitsch 2004, Hilpert 2008). While near-synonymous constructions may show substantial overlap in their collocational profiles, any divergences should hold important clues for the analysis of how their meanings can be differentiated.

Collocations and collocational preferences can be studied through a multitude of measures (Evert 2009). As we will explain in more detail in Section 3, our study utilizes token-based semantic vector space modeling (Schütze 1998, Heylen et al. 2015, Hilpert and Correia Saavedra 2017), which is a method that allows for the comparison of individual concordance lines on the basis of distributional information. Token-based semantic vector spaces represent a key word in context, such as the modal auxiliary *may* in *That may be a good idea*, in terms of second-order collocates, that is, the collocates of the items that are found in the immediate linguistic context of the key word. The collocational profile of *idea*, which is a word that frequently occurs with elements such as *vague*, *brilliant*, or *faintest*, thus forms part of the information that represents the specific use of *may* in the example above. Drawing on second-order collocates greatly enriches the contextual information that is present in a concordance line, and it can bring to light similarities between concordance lines that themselves do not share any lexical material (Hilpert and Correia Saavedra 2017:

10). In other work, the method has been applied to the analysis of lexical polysemy (Heylen et al. 2015). The present study will test its suitability for the analysis of grammatical elements, which have meanings that are relatively more abstract and general, and which therefore exhibit a more diffuse collocational behavior.

The rest of this paper has the following structure. Section 2 contextualizes the present study within the literature on modal meanings and their corpus-based investigation. Section 3 discusses issues of data and methodology. Our results are presented in Section 4, which is divided into two parts. In the first one, we contrast the collocational behavior of near-synonymous modal expressions. The pairs *may* and *might* and *must* and *have to* were chosen for this purpose. The second part focuses on the comparison between different meanings that one form can express. Here, we discuss epistemic and deontic uses of *may* and *must*. Section 5 concludes the paper with a number of pointers for future research.

2 Theoretical background

The English modal auxiliaries have been studied extensively from a corpus-linguistic perspective (Coates 1983, Palmer 1990, Krug 2000, Facchinetti et al. 2003, Zamorano-Mansilla et al. 2015*, inter alia*). Among the recurrent findings that this research tradition has produced, important insights include the fact that any given modal expression encodes a broad range of different meanings and that this meaning potential is modulated by contextual factors. The following uses of the modal auxiliary *would*, taken from the BNC, illustrate these points.

(1) Alright, so would you like to do this for us?

(2) Whenever I wore my brown leather shoes, people would stop me and say, with genuine amazement, 'Hey, what are those things you've got on your feet?'

(3) A: You always use the left lane.
B: Yeah sorry sorry sorry.
A: Why would you go into the wrong lane to a roundabout?

In example (1), *would* is used in its function of a tentative form (Palmer 1990: 58). It appears in the syntactic frame of a polar interrogative that expresses a polite request. Example (2) shows *would* in a different syntactic context with a different meaning. It appears as the predicate of a main clause that is preceded by a temporal subordinate clause headed by the element *whenever*. The interpretation that results is that of habitual meaning (Palmer 1990: 155). Yet another meaning of *would* is shown in example (3), where it is embedded in a question headed by *why*. Here, its epistemic quality is no longer tentative but factual, and the auxiliary expresses subjectified meanings of disapproval and irritation (Furmaniak and Larreya 2015: 116). In each of the three examples, formal cues in the environment of the modal auxiliary allow the hearer to rely on more or less conventionalized form-meaning mappings in order to figure out what the speaker means. These cues involve syntactic patterns, collocations with lexical or grammatical elements, or a combination of the two. Above and beyond that, there are of course contextual cues that are not part of the written representation of language that is provided by the corpus. Prosody, gesture, and other multimodal aspects of language use allow the hearer to identify the intended modal

meaning with even greater precision and reliability. Our point here is that even when our data is restricted to lexical and morphosyntactic information, it is possible to disentangle modal meanings with some measure of success.

An important question to ask in this context is how contextual elements can be made useful for the distinction of different modal expressions. The differentiating potential of lexical elements has been investigated for example by Hilpert (2013), who focuses on the lexical verbs in the infinitive that are projected by the modal auxiliaries in a study that compares the nine core English modals *can*, *could*, *may*, *might*, *must*, *shall*, *should*, *will* and *would*. Using data from the COHA (Davies 2010), Hilpert (2013: 75) finds that collocational profiles reflect degrees of similarity in the meaning potentials of the auxiliaries, so that for example *must* and *should*, which both encode obligation, show substantial overlap in their verbal collocates. The same holds for *might* and *could*, which encode possibility. The auxiliary *would* exhibits the collocational profile that overlaps the least with those of the other auxiliaries, which suggests that its meaning potential cannot easily be subsumed under any major category of modal meaning, such as deontic, epistemic, or dynamic modality (cf. Palmer 2003: 7). Diachronic shifts in the collocational preferences of the auxiliary *may* further point to a shift in its meaning potential towards epistemic modality. Specifically, the verbal collocates that are most strongly attracted to *may* in the 19th century lend themselves to the expression of permissive meaning, whereas the top collocates in more recent data combine with *may* to express possibilities. This finding is corroborated by a follow-up study (Hilpert 2016) that uses a distributional approach to monitor how the meaning potential of *may* has changed. The study represents the meaning potential of *may* as a semantic vector space of its verbal collocates. Over time, several areas of the semantic vector space change with regard to the density of types and their co-occurrence frequency

with *may*. In line with earlier findings, semantic areas that contain verbs with abstract and epistemic meanings are diachronically on the rise (Hilpert 2016: 81). These findings support the conclusion that collocating verbs differentiate between modal expressions and modal meanings.

Besides lexical verbs, adverbs are another highly informative class of contextual elements. In a study on the collocational links between modal auxiliaries and adverbs, Flach (in press) contrasts combinations such as *would rather* or *may as well*, which exhibit strong mutual attraction, with pairings such as *would well* or *will rather*, which are grammatically possible but underrepresented in authentic language use. Flach argues that these combinations can be arranged along a cline of idiomaticity. At one end of this cline, collocations such as *may as well* encode non-compositional meanings and thus exhibit a high degree of idiomaticity. In pairings such as *could possibly*, auxiliary and adverb have meanings that harmonically converge on a single idea, which results in their frequent co-occurrence and strong mutual attraction. Further along the continuum are chance collocations such as *could later* or *must also*. These are commonly used, but the meanings of these collocations are clearly transparent and compositional. Using collostructional analysis (Stefanowitsch and Gries 2003), Flach further identifies combinations that are significantly underrepresented in data from the COCA (Davies 2008), including *can never*, *would well*, or *will rather*. Hence, when speakers plan an utterance, the choice of a modal auxiliary can be influenced by an adverb that the speaker wishes to include in that utterance and the associative ties that this adverb has to different auxiliaries.

Co-dependencies between modal expressions and contextual elements have further been observed at the level of syntactic structure. Hohaus (2019) presents the results of a study in which corpus examples of English modal auxiliaries are annotated for the syntactic

context in which they occur. Controlling for variables such as the animacy of the subject, grammatical aspect, and the meaning of the lexical verb, Hohaus (2019) shows, among other things, that the auxiliary *could* is overrepresented in adverbial clauses while *must* is underrepresented in relative clauses. A multifactorial analysis reveals that the English modal auxiliaries are conventionally associated with specific configurations of syntactic and semantic features. The choice between alternative modal expressions is thus governed by associations that target linguistic units at different levels of abstraction, including lexical elements, semantic classes of elements, and syntactic patterns.

A combination of these cues is at work in conventionalized modal expressions that convey specific semantic and pragmatic information. Cappelle and Depraetere (2016) study patterns such as *You can say what you want about X, but Y*. This semi-fixed arrangement of syntactic structures and lexical elements globally encodes a concessive relation: Despite any number of things one could possibly say, something else is the case. This interpretation is related to the modal meaning of *can*, but it is non-compositional and associated with the entire lexico-grammatical pattern. Cappelle and Depraetere show that more examples of this kind are easily found. Expressions such as *I can't complain*, *Not if I can help it*, or *If I could just say a few words* all involve conventionalized pairings of form and meaning.

In summary, contextual elements are a key factor that determines how speakers choose between alternative modal expressions and how hearers distinguish different meanings of the same modal expression. The available evidence shows that relevant information is extracted from different structural levels of linguistic organization, so that both concrete linguistic units and more abstract patterns are taken into account. Crucially, all available studies rely on first-order collocates to make this point. The analyses that will be presented in the following sections will test whether similar results can be obtained by a

corpus-linguistic method that is sensitive to second-order collocates of a key word in context.

3 Data and methodology

The methodological basis for the present analysis is semantic vector space modeling (Turney and Pantel 2010, Kiela and Clark 2014, Lebani and Lenci 2016), which operationalizes meaning in terms of frequency vectors that capture how often a given linguistic unit co-occurs with other units in a given corpus. The rationale for such an approach is that semantically related units will appear in similar contexts. For example, the words *shocked* and *astonished* have similar meanings and may both be preceded by *He was* and followed by *to hear about the incident*. Semantic representations of linguistic units are usually constructed in such a way that concordance lines are obtained from a corpus, a context window of a certain size around the key expression is chosen, and the frequencies of all elements within that context window are determined. This procedure will aggregate the co-occurrence frequencies of context elements that appear in the neighborhood of a given key expression, which will for example reveal that the word types *shocked* and *astonished* have very similar collocational profiles in a corpus such as the British National Corpus (Leech 1992). A description of these working steps is offered by Levshina (2015: 323).

The present analysis will use an extension of this method that focuses not on the comparison of word types, but rather on concrete word tokens in their specific contexts. It is our aim to investigate how multiple usage events of the same linguistic form differ from one another. For this purpose, we utilize token-based semantic vector space modeling, which

draws on the same general logic, but which constructs semantic representations of linguistic units on the basis of second-order collocates. For a given use of a key expression in context, the method takes the elements that are found within a specified context window, creates collocate vectors for all of these elements, and merges those vectors into a single one. The following paragraphs provide a basic description of the application of token-based semantic vector space modeling that is implemented in the present study. Hilpert and Correia Saavedra (2017: 4-9) provide a detailed explanation of the method.

Our study uses data from the British National Corpus (Leech 1992). The BNC is a balanced tagged corpus that comprises 100 million words of different text types. About 10% of the corpus represents spoken English. In the following paragraphs, we describe the construction of a type-based semantic vector space that forms the basis for our analyses. That vector space has been constructed with co-occurrence frequencies that have been extracted from the BNC.

For the purposes of our study, we aimed to construct context vectors for a set of about twenty thousand vocabulary items. A lexical set of this size is not only large enough to contain the most commonly used words of a language, but also many rarer items. Our starting point was a list of the most frequent 20,200 elements in the BNC. From this list, we removed a set of stop words that included grammatical items, punctuation, single-letter lexemes, numbers, forms of the grammatical verbs *be*, *have*, and *do*, and elements that were tagged as unclear. Tags were preserved, so that for example the infinitive and the basic form of the verb *love* are represented as two different types, namely *<w vvi>love* and *<w vvb>love* respectively.

After removing all unwanted items, our set of vocabulary items had been reduced to 19,429 types. For these types, it was determined how often each element of the list

occurred with each other element within a span of two words to the left and two words to the right. For example, given a concordance of the word *digestive*, what elements appear in a context window to its immediate left and right in the corpus data, and how frequently do these elements appear? The raw co-occurrence frequencies that we obtained from our concordances were transformed by means of a collocation measure, for which we chose Pointwise Mutual Information (PMI). The formula we used is shown in (4).

(4)

$$PMI = \ln \frac{p(x,y)}{p(x) * p(y)}$$

The formula yields a value of mutual association strength for bigrams such as *digestive biscuits*. The joint probability of *digestive* and *biscuits* (p(x,y)) is divided by the product of the individual probabilities of *digestive* (p(x)) and *biscuits* (p(y)). These probabilities are derived from the observed frequencies of *digestive*, *biscuits*, and their co-occurrence. Table 1 shows the relevant frequencies for this example.

Table 1: Joint and individual frequencies and probabilities of *digestive* and *biscuits*

	observed frequencies				probabilities		
	digestive	¬ digestive	Total		digestive	¬ digestive	Total
biscuits	12	3,998	4,010	biscuits	0.00000008	0.00002519	0.00002527
¬ biscuits	1,124	158,658,614	158,659,738	¬ biscuits	0.00000708	0.99996765	0.99997473
Total	1,136	158,662,612	158,663,748	Total	0.00000716	0.99999284	1.00000000

Our database contains 12 concordance lines in which the words *digestive* and *biscuits* occur in close proximity, that is, within the specified context window of two words to the left and right of the key word. While *digestive* occurs in 1,136 combinations in total, *biscuits* occurs in 4,010 combinations. The right panel of Table 1 shows probabilities that obtain when one of the 158 million word combinations of our database is drawn at random. As is shown in (5), these probabilities yield a PMI value of 6.09 for the combination of *digestive* and *biscuits*, which reflects the fact that the two indeed form a conventionalized collocation.

(5)

$$PMI = \ln \frac{0.00000008}{0.00000716 * 0.00002527} = 6.09$$

The size of the co-occurrence matrix was reduced by deleting all columns and rows that fell short of reaching at least one highly informative PMI value. For this, a cut-off point of PMI=5.5 was selected. In order to illustrate the cut-off point, Table 2 shows the context items with the highest PMI values for the element *biscuits*.

Table 2: Context items with high PMI values for the element *biscuits*

context item	*PMI*	*context item*	*PMI*	*context item*	*PMI*
<w nn2>biscuits	7.08	<w nn1>chocolate	5.87	<w nn1>tin	5.38
<w nn2>cakes	6.30	<w aj0-nn1>ginger	5.87	<w nn1>cheese	5.25
<w aj0>digestive	6.04	<w nn2>sweets	5.83	<w nn1-vvb>jam	5.21
<w nn2>crisps	5.93	<w nn1>packet	5.73	<w nn2>chocolates	5.16
<w nn2>packets	5.89	<w nn1>biscuit	5.60	<w nn2>tins	5.12

After excluding the relevant columns and rows, we are left with a type-based semantic vector space that consists of 12,621 columns and 12,619 rows. This vector space is the basis for the analytical steps that are described below.

As was pointed out above, token-based semantic vector spaces compare context vectors of individual usage events. The usage events that are analyzed in the present study are concordance lines of the modal auxiliaries *may*, *might*, *must*, and *have to*, which were extracted from the BNC. For each auxiliary, 10,000 usage events were randomly selected from the corpus. Negated forms such as *mightn't* and *mustn't* were not included. Since *have* inflects for person and tense, its variant forms *has*, *had*, and *having* were included. Each concordance line contained the modal auxiliary with ten words to the left and to the right. The concordance lines were stripped of any elements that were not represented in our type-based semantic vector space, which led to the exclusion of punctuation signs, most grammatical words, and low-frequency lexical items. Concordance lines with fewer than four remaining context items were discarded, which greatly reduced the number of concordance lines that were kept in the database.

Table 3 presents a summary of the concordance lines that were obtained for each auxiliary and the number of context items present in those concordance lines.

Table 3: Concordance lines for *may*, *might*, *must*, and *have to*

modal auxiliary	n concordance lines	n context items					
		4	5	6	7	8	9+
may	1914	1114	491	225	50	28	4
might	1268	783	331	113	29	12	0
must	1537	887	445	137	51	12	5
have to	1133	718	274	110	23	8	0

Each concordance line in the resulting database was used to construct a semantic representation in the shape of a vector of PMI values. The following example serves as an illustration. The concordance line with *may* in (6) contains five context elements that are represented in our type-based semantic vector space. These elements are shown in (7).

(6) emotions. The desire to love and be physically close may be repressed if it is seen to pose a threat

(7) <w nn2>emotions, <w vvi>love, <w av0>physically, <w vvi>pose, <w nn1>threat

Each of the five context items was looked up in our type-based semantic vector space, and the five corresponding vectors of 12,619 PMI values were copied into a table. In order to merge them into a single vector, the PMI values were averaged as shown in Table 4. The

elements in the first column of Table 4 represent the vocabulary items in our type-based semantic vector space in alphabetical order.

Table 4: A concordance line represented as a vector of averaged PMI values

	<w nn2>emotions	<w vvi>love	<w aj0-av0>physically	<w vvi>pose	<w nn1>threat	ø
<w aj0-av0>above	0	0	0	0	1.09	0.218
<w aj0-av0>alike	0	0	0	0	0	0
<w aj0-av0>away	0	2.38	0	0	1.91	0.858
<w aj0-av0>brightly	0	0	0	0	0	0
<w aj0-av0>cheap	0	1.19	0	0	0	0.238
<w aj0-av0>cold	0	2.35	0	0	0	0.47
…	…	…	…	…	…	…

The rightmost column in Table 4 constitutes a vector that semantically represents the concordance line shown in (6). When the elements in that vector are ordered in terms of their PMI values, it becomes apparent that the vector actually captures a semantic image of the concordance line. This is shown in Table 5, which lists the second-order collocates of concordance line (6) with the highest averaged PMI values. The items in Table 5 represent a semantic cross-section of associations with the words *emotions*, *love*, *physically*, *pose*, and *threat*, and they capture that the concordance line relates to the mind, the body, desire, and problems associated with these concepts.

Table 5: Second-order collocates of *may* in example (6)

collocate	PMI	collocate	PMI	collocate	PMI
<w av0>psychologically	2.470	<w aj0-vvd>handicapped	1.912	<w vvg>loving	1.698
<w aj0>knowledgeable	2.354	<w aj0>fake	1.840	<w aj0>stubborn	1.690
<w nn1-vvb>fear	2.078	<w av0>emotionally	1.826	<w aj0>frail	1.674
<w vvi>pose	2.034	<w aj0-vvn>released	1.822	<w nn2>adventurers	1.668
<w nn2-vvz>fears	1.970	<w np0>kylie	1.802	<w vvi>love	1.666
<w nn1>threat	1.942	<w nn1-vvb>desire	1.738	<w aj0>appalling	1.648

An advantage that token-based semantic vector spaces have over their type-based alternatives is that they can capture the similarity between concordance lines that are semantically related but do not share any lexical material (Hilpert and Correia Saavedra 2017: 10). Table 6 illustrates this. The table shows two concordance lines that relate to the same semantic domain of property lease. The context items that are extracted from the respective concordance lines do not overlap. By contrast, three of the eight most strongly attracted second-order collocates, shown in boldface, capture the semantic relation between the two concordance lines.

Table 6: Different context items, overlap in second-order collocates

concordance lines	is someone's home. To a landlord a dwelling may be a unit of accommodation or an item of investment	occupier of the leased premises. A right of re-entry may therefore be exercised by L or L2 against S. Enforceability
context items	<w nn1>landlord, <w nn1>dwelling, <w nn1>accommodation, <w nn1>investment	<w nn1>occupier, <w nn2>premises, <w vvn>exercised, <w np0>s.
second-order collocates	<w aj0>rented, <w nn1>tenancy, **<w nn1> occupier**, <w nn1 dwelling, <w nn1-vvb>rent, **<w nn2>premises**, **<w nn1>tenant**, <w nn1>exemption	**<w nn1>occupier**, <w nn0>s., **<w nn1>tenant**, **<w nn2>premises**, <w nn1>inference, <w vvn>exercised, <w aj0>liable, <w aj0>unfit

Token-based semantic vector spaces are extremely sensitive to semantic similarity due to shared linguistic material, but they are nonetheless able to differentiate between concordance lines with limited overlap. Table 7 shows two concordance lines that both include the polysemous lexical item *wave*. In the first concordance line, *wave* is used metaphorically. In the second one, *wave* refers to the movement of water. While the two concordance lines share *wave* as a context item, the semantic convergence within each respective set of context items results in sets of second-order collocates that do not show any overlap at all. The first is clearly crime-related while the second one relates to natural phenomena.

Table 7: Overlapping context item, different second-order collocates

concordance lines	ALDOUS/Wiltshire Police Voice over Government optimism that the crime **wave** may at last be under control will be little consolation to Franc	responsible for damaging marine structures. A pocket of air may be trapped between the breaking **wave** and the cliff,
context items	<w nn2>police, <w nn1>voice, <w avp-prp>over, <w nn1>crime, **<w nn1>wave**, <w nn1>franc	<w aj0-vvg>damaging, <w aj0>marine, <w nn1>pocket, <w nn1>air, **<w nn1>wave**, <w nn1>cliff
second-order collocates	<w aj0-nn1>video-tape, <w np0>det, <w nn1>crackdown, <w np0>insp, <w nn2>burglars, <w nn1-np0>crime, <w np0>gloucestershire, <w np0>graeme, <w nn2>detectives, <w nn2>arrests	<w aj0>coastal, <w nn1>cliff, <w nn1>erosion, <w nn2>pollutants, <w vvd>slapped, <w nn1>pollution, <w nn2>waves, <w nn1>ozone, <w aj0>marine, <w nn1>wave,

These observations motivate our hypothesis that similarities and differences in meaning between different modal auxiliaries will be reflected in collocational patterns that a token-based semantic vector space can capture.

To allow pairwise comparisons of the modal auxiliaries in our database, we combined the vectors of all concordance lines with *may* and *might* into one matrix and the vectors of all concordance lines with *must* and *have to* into another. We downsampled the data for the more frequent modals, so that *may* and *might* are each represented by n=1268 concordance lines and *must* and *have to* by n=1133 concordance lines. Cosine distances were obtained for all mutual pairings of concordance lines within those two datasets, yielding two distance matrices. The resulting distance matrices allow us to test the hypothesis that individual usage events of *may* and *might* and *must* and *have to* can be distinguished on the basis of second-order collocates. Our findings are discussed in Section 4.1.

Our second main research question is whether different meanings of the same modal expression are reflected in diverging collocational profiles. We address this question with the same data that were discussed above. Concordance lines that were obtained for *may* (n = 1914) and for *must* (n = 1537) that contained at least four usable context items (cf. Table 3) underwent manual annotation for a distinction between deontic and epistemic meaning. Deontic uses of *may* and *must* express permission and obligation respectively, as illustrated in examples (8) and (9), epistemic uses express possibility and certainty, as shown in (10) and (11).

(8) grey eyes lit with amusement. "You may say it, my lord. I'm shock-proof

(9) or by pressing the help key. All errors must be resolved before the acceptance

(10) bacteria have already been isolated, and may potentially contain useful new

(11) key to who meets AC Milan - and surely it must be the Italians - in the final

The semantic annotation was performed by both authors. Subsets of the data (10%) were doubly annotated to allow tests of interrater reliability. The tests showed a Cohen's kappa value of 0.52 (83.2% agreement) for deontic and epistemic meanings of *may* and a value of 0.84 (93.5% agreement) for deontic and epistemic meanings of *must*. Cases of disagreement in the annotation were discussed and resolved. Table 8 shows that *may* and *must* show inverse tendencies with regard to the distribution of deontic and epistemic meaning. While

may occurs more frequently with epistemic meaning in our dataset, *must* is used relatively more often with deontic meaning.

Table 8: Examples of *may* and *must* with deontic and epistemic meaning

	deontic	epistemic
may	428	1486
must	1233	304

We downsampled the examples with the more frequent meaning, so that deontic and epistemic *may* are each represented by n=428 concordance lines and deontic and epistemic *must* by n=304 concordance lines. For both annotated sets of concordance lines, the analytic steps that were outlined above provide token-based context vectors that allow us to test the hypothesis that different meanings of the same modal auxiliary are reflected in different second-order collocates. In Section 4.2, we discuss each of the two auxiliaries individually and explore how their deontic and epistemic meanings differ in terms of second-order collocates.

4 Results and discussion

4.1 Pairwise comparisons of modal auxiliaries

In order to test how accurately second-order collocates differentiate between uses of the modal auxiliaries *may* and *might* in the BNC, we used binary logistic regression (Levshina

2015: 253). The dependent variable of the analysis represents the choice between *may* and *might*. The analysis uses the collocational data described in section 3 and assigns each concordance line to one of the two modal auxiliaries. The predictor variables of the regression analysis represent the collocational information that is contained in the token-based semantic vector space we described in the previous section. We used metric multidimensional scaling (Wheeler 2005) to transform the high-dimensional collocational data into a lower-dimensional space. With this method, we reduced the token-based semantic vector space to 20 dimensions that we subsequently used as predictor variables for the logistic regression. Table 9 shows that for *may* and *might*, the analysis yields a classification accuracy of 64.4%, up from a chance level of 50%. Between the two auxiliaries, classifications for *may* are relatively more reliable. While a majority of concordance lines with *may* and *might* are classified correctly, this result indicates that the two auxiliaries commonly occur across similar lexical contexts. To contextualize this finding, the classification accuracy of our analysis can be compared against results obtained by Hilpert and Correia Saavedra (2017: 15-16), who report classification accuracies of 89.8% for two semantically unrelated lexical items and 64.2% for the near-synonymous adjectives *happy* and *glad*. Given the broad meaning potentials of *may* and *might*, it is to be expected that second-order collocates distinguish between the two only to a moderate degree.

Table 9: Classification of *may* and *might* on the basis of second-order collocates

	classified as *may*	classified as *might*

may	859	409
might	493	775

As was discussed in the introduction, *may* and *might* can occur in identical lexical contexts. Isolated constructed examples do of course not tell us whether the two modal auxiliaries have any preferences for different contexts, and how strong these preferences might be. Table 9 shows that lexical context influences the choice between *may* and *might* probabilistically, but that it does not determine it. This can be illustrated with the examples in (12).

(12) a. particularly rich in haemopoietic stem cells. Thus placental transfusion may be important in constitution of the preterm infant's bone

b. resulted in smooth muscle contraction. Receptors for 5HT 2 may alter pyloric and cecal function.

c. proportion of active and inactive tissue between patients (such as might result if one patient's mucosa was atrophic)

All three examples come from medical texts and include specialist vocabulary. Examples (12a) and (12b) represent typical uses of *may* in medical texts. Example (12c) includes *might* but is misclassified as *may* by the regression because of its overlap in second-order collocates with examples (12a) and (12b).

Misclassification also occurred in examples such as (13). The concordance line contains the set phrase *If I may*, which is an idiomatic expression with a distinct meaning.

(13) my, of my ri-- remarks Mr Mayor if I may, to the liberal democrats er-- party

The context items that are extracted from the concordance line (*remarks*, *Mayor*, *liberal*, and *democrats*) are associated with uses of *might* and thus lead to the misclassification. Due to the exclusion of punctuation and grammatical elements, the analysis is blind to highly conventionalized patterns such as parenthetical *if I may*, which is of course a drawback of this approach. Merely retaining the stop words in the analysis would not be a solution, since it is the specific order of the grammatical elements that is meaningful in this case. To solve the problem, the approach would have to go beyond a bag-of-words model and include sequential information. Since the aim of this paper is to work out whether and how token-based semantic vector spaces without sequential information can be used to distinguish between modals and modal meanings, that route is not pursued here. It is, however, an important avenue for future research.

Another result that the analysis brings to light is that conversational examples are correctly classified as *might* with high confidence. While *might* is frequently found in casual conversation, *may* shows the opposite tendency. The examples in (14) show three relevant concordance lines. High-ranking second-order collocates that are shared by the examples and that influence their classification include the elements *mummy*, *alright*, *fucking*, *telly*, as well as *gonna* and *wanna*.

(14) a. excuse by saying you wanna -- co-- converse. She might be still asleep though.

b. of I'm gonna be a pop star might be a bit much for some. But listen up you

c. No, I'm gonna go upstairs in a minute, I might go into Rupert's room

What this shows is that the analysis detects genre effects through the collocational preferences of those genres. In the case of medical texts, relevant second-order collocates constitute a set of highly specific elements. The examples in (14) show that also elements with broader meanings can be informative, so long as they are strongly indicative of a given genre.

For the analysis of *must* and *have to*, we followed the same analytical steps that were carried out in the comparison of *may* and *might*. As Table 10 shows, a logistic regression applied to our data distinguishes between *must* and *have to* with 62.4% classification accuracy, which means that their respective collocational profiles exert an influence that is measurable although somewhat limited. The classification is relatively more accurate for *must*.

Table 10: Classification of *must* and *have to* on the basis of second-order collocates

	classified as *must*	classified as *have to*
must	742	391
have to	462	671

Concordance lines that are correctly classified as *must* and that are highly similar include the examples in (15), which are from a computing manual, and which therefore share highly specific second-order collocates such as *lifespan*, *offline*, and *update*. The use of *must* in these contexts relates to direct and explicit instructions to the computer user. If the steps outlined by the computing manual are not followed, the user will not obtain the desired result.

(15) a. been closed via another SSR. In addition, the SSR must refer either to an SPR or to a valid module

b. privileges are required to use this option, but you must have been requested to endorse the SSR for specific package(s)

c. including spaces. Any future reference to this Client must be made by either the existing Client identifier

Similarly, concordance lines with legal terminology are associated with *must*. Accordingly, examples with *have to* that contain elements with related second-order collocates are misclassified. This includes examples such as (16c).

(16) a. orders for costs made below will stand and the prosecutor must pay the defendant's costs in this House.

b. of their duty, it is clear that the prosecutor must show that the defendant was aware that the person

c. is for the judiciary. It is true that somebody has to appoint the Chief Inspector as the Lord Chancellor appoints

We observed above that *may* and *might* are distributed unevenly across formal and informal texts, with *might* occurring more frequently in face-to-face conversation. A similar finding pertains to *must* and *have to*, as the latter is found in conversational examples that are correctly classified by the logistic regression. The examples in (17) illustrate this.

(17) a. not really gonna stand up with the men -- they have to become, referees, they have to become coaches.

b. A: oh I'll wash

B: (yawn) (pause) Sorry?

A: Alright? That's the way it has to be. Gonna look

B: Shut up!

A: a little bit

The examples do not share a common semantic domain, as (17a) is about football and (17b) is taken from a casual conversation about clothing. What unites them is the informal tone, which is captured by shared second-order collocates such as *gonna* and *wanna*.

4.2 Pairwise comparisons of modal meanings

We followed the same analytic procedure that was described above for the distinction and pairwise comparison of modal meanings. The only difference between the analyses described in the previous section and the ones that will be presented in this section is the dependent variable of the logistic regression. In the contrasts between *may* and *might* and *must* and *have to*, that variable concerned a difference in linguistic form. Here, the dependent variable is a semantic distinction that we hand-coded into the datasets. A logistic regression analysis based on the collocational profiles of deontic and epistemic *may* classifies the 856 concordance lines in our database with 65.9% classification accuracy, up from a chance level of 50%. Table 11 summarizes the results.

Table 11: Classification of deontic and epistemic *may* on the basis of second-order collocates

	classified as deontic	classified as epistemic
deontic may	272	156
epistemic may	136	292

Concordance lines that contain medical vocabulary are correctly and reliably classified as epistemic, which indicates that medical texts typically encode possibilities, rather than permissions. Two representative examples are given in (18).

(18) a. Crohn's disease in an indolent chronic inflammatory disorder that may affect the entire alimentary tract, with inflammation

b. particularly in sufferers from alcoholism (in whom these features may be indications of their primary disease of alcoholism)

Conversely, the logistic regression identifies concordance lines with computer-related lexis as having deontic meaning. Example (19b) refers to user privileges, example (19b) describes the functionality of a software.

(19) a. option 4.5.1, Relate SSR to Modules. Option 4.5.1 may only be used by the submitter of the SSR.

b. packages within a named package. Or modules may be read individually and automatically updated to a new version.

Deontic meaning is further assigned to concordance lines that deal with contracts. Example (20a) describes a legal privilege of landlords with its first instance of the auxiliary (*may from time to time*). The regression therefore misclassifies concordance line (20b), which shares second-order collocates such as *covenants*, *premises*, and *clause* with (20a), but which nonetheless expresses possibility and hence epistemic meaning.

(20) a. account in the United Kingdom that the Landlord may from time to time nominate. There may be occasions where the tenant

b. original tenant is forced to join a guarantor, it may be advisable for the guarantor's covenant to be amended

Moving on to a contrast of deontic and epistemic *must*, a logistic regression analysis distinguishes the two meanings with 74% classification accuracy, which suggests that the distinction between these two meanings is tied relatively strongly to collocational profiles.

Table 12: Classification of deontic and epistemic *must* on the basis of second-order collocates

	classified as deontic	classified as epistemic
deontic must	218	86
epistemic must	72	232

Sets of collocates that are highly indicative of deontic meaning have been discussed above in relation to the examples in (15), which represented handling instructions from a computer manual, and (16), which contained legal language. In both of these domains, agents are

required to carry out certain actions in order to obtain a desired result, so that *must* is used with the meaning of obligation.

Among the concordance lines that are classified as epistemic with high probability, a group of examples represent texts that deal with historical facts and events. Examples (21a) and (21b) both include the context item *century*, and they share second-order collocates such as *dating*, *ninth*, and *thirteenth*.

(21) a. here during the 17th century. Many more single coins must have been found and have not been recorded.

b. in the east since the seventeenth century. The changes must have been as great as those in western Europe

Other epistemic uses of *must* that are reliably classified as such include conversational examples such as the ones shown in (22). While conversational data shows a preference for *have to* over *must*, when *must* does occur, it tends to encode epistemic meaning rather than deontic meaning. Consequently, the regression misclassifies example (22c), which represents fictional dialogue but expresses obligation.

(22) a. Even if we're not gonna succeed -- it must have stuck in their throat that there was Jesus lying asleep.

b. A: Well that's (unclear).

B: Aye. Mhm.

A: That must have gone on a lot (unclear)

B: Oh aye. Aye and then he'd leave.

c. this unexpected confession. 'Now, listen. You must never think of such a wicked thing again. Promise

Summing up this section, the deontic and epistemic meanings of *may* and of *must* are to some extent contingent on the lexical material that appears in their linguistic context. The analyses reveal both specific contextual cues, such as the mutual attraction between computer-related lexis and deontic meanings of *must*, and more general tendencies, including the result that informal and conversational data will be associated with epistemic uses of *must*.

5 Conclusions

This paper started out with the question of how speakers decide between different modal expressions when several near-synonymous choices are available. Based on corpus-based findings from the literature (Facchinetti et al. 2003, Zamorano-Mansilla et al. 2015, Cappelle and Depraetere 2016), it is clear that speakers' choices are influenced by the linguistic context, which includes both lexical collocates (Flach to appear) and syntactic structures (Hohaus 2019). Also, different meanings of the same modal expression are reflected by elements that appear in the linguistic context (Hilpert 2016). Following up on these observations, the present paper tested the hypothesis that token-based semantic vector spaces (Heylen et al. 2015, Hilpert and Correia Saavedra 2017) can be used as a method for the corpus-based distinction between alternative modal expressions and modal meanings.

The analyses that have been carried out compare concordance lines in terms of second-order collocates, which are hypothesized to capture differences in meaning.

The results indicate that second-order collocates allow us to better understand the choices that speakers make. In our data, the choice between modal expressions is influenced by contextual elements that reflect specific semantic domains such as computing, medicine, or legal privileges and obligations. We have also observed broader genre effects, such as a preference for *might* over *may* in face-to-face conversation. In both cases, relevant concordance lines converge on similar sets of second-order collocates. We have further shown how a strong association between specific second-order collocates and a given modal expression can occasionally lead the analysis astray and result in misclassifications when a modal expression is used in an unusual semantic context. With regard to the distinction of different modal meanings our analysis yields similar results. Second-order collocates provide a statistical signal that facilitates the discrimination of deontic and epistemic modal meaning. Again, both specific lexical items and more general characteristics of different text types can be shown to play a role. These findings align with theoretical claims in usage-based construction grammar that portray linguistic knowledge as a network of symbolic units that are mutually interconnected at different levels of schematicity (Diessel 2019). Overall, our study thus offers another illustration of how token-based semantic vector spaces can be used to address questions that pertain to linguistic theory.

An important caveat that we need to discuss concerns the ways in which aggregate data from linguistic corpora can be used to make inferences about the cognition of individual speakers. Blumenthal-Dramé (2012: 28-29) cautions against modelling cognitive phenomena on the basis of corpus-based collocation measures. Given that the analyses presented in this paper were based on corpus data but purported to investigate how

speakers choose between modal expressions, a few clarifying comments are in order. First and foremost, we do not assume that our results reflect knowledge that all speakers of English share. Familiarity with a text type such as legal language is specialist knowledge that is available only to speakers who have had sufficient exposure to this genre. On a more positive note, the proposal that speakers are sensitive to contextual cues in the form of multi-word sequences actually receives empirical support from studies that triangulate corpus-based work with psycho-linguistic experiments. For example, Arnon and Snider (2010: 76) asked participants to judge the grammaticality of four-word sequences such as *I don't know why* (grammatical) or *I saw man the* (ungrammatical). The results indicate a frequency effect. Grammatical sequences that occur more often in corpus data yield shorter reaction times. These findings indicate that aggregate frequencies from corpora can be meaningfully related to cognitive phenomena.

Another note of caution concerns the fact that the effect of second-order collocates is probabilistic and exerts an influence that can be overridden. Across many contexts, our analysis estimates that alternative modal expressions such as *may* and *might* are about equally appropriate. This is illustrated by the concordance line shown in (23), which includes *may* and for which the logistic regression analysis returns a probability estimate that indicates a relative absence of informative second-order collocates.

(23) Intensive and uncontrolled fishing in Russia's Far Eastern seas may soon lead to the total exhaustion of fish stocks

When *may* is replaced with *might* in this example, the resulting sentence is perfectly acceptable, even though *might* and *soon* actually statistically repel each other (Flach to

appear). A high degree of potential interchangeability is in fact to be fully expected with forms such as modal auxiliaries. In other domains of English grammar, alternating forms such as the ditransitive construction and the prepositional dative construction (Gries and Stefanowitsch 2004) or future constructions with *will* and *be going to* (Hilpert 2008) exhibit similar profiles of partial collocational overlap and construction-specific collocational preferences. The relatively modest predictive power of second-order collocates is further in line with the finding that diversity in collocational profiles correlates positively with degrees of grammaticalization (Correia Saavedra 2019: 98): As linguistic forms grammaticalize, their meanings become broader and more abstract, so that they can occur across a wider set of contexts. As highly grammaticalized elements, the English modal auxiliaries show a fair amount of collocational variability. What analyses such as the ones presented in this paper can offer is the insight that below the surface of this variability, there is a multitude of associations that conventionally tie a given modal expression to a particular set of lexical elements, or a specific modal meaning to a particular linguistic context.

This last point brings us back to the idea that any linguistic form should map onto a different meaning (Bolinger 1968: 127). Does the meaning of example (23) change when *may* is replaced with *might*? Our analysis suggests that this might be the wrong question to ask. The collocational evidence supports previous research that has argued that *may* and *might* differ in their respective meaning potentials (Coates 1983, Palmer 1990). In the sense that each auxiliary appears in contexts in which the other one is not appropriate, different forms clearly map onto different meanings. At the same time, the high degree of schematicity that both *may* and *might* have acquired through grammaticalization means that both forms can be used in contexts to which neither of the two is particularly attracted. In this view, replacing *may* with *might* would indeed bring about a change, but only insofar

as a set of weak links between *may* and its lexical context items is replaced with a set of equally weak links between these items and *might*.

To conclude, associations between linguistic units can go a long way not only towards explaining why speakers choose the variants that they do, but also towards identifying contexts in which a choice might be relatively more open. The approach taken in this paper thus opens up a new perspective for studies of modality and other grammatical domains in which alternations between near-synonymous forms can be observed. Rather than focusing exclusively on the differences between two forms, it would be a fruitful avenue of research to consider in more detail when and how linguistic units share an ecological niche. As we hope to have shown, token-based semantic vector spaces provide useful tools for the further exploration of this issue.

References

Arnon, I. and Snider, N. (2010). More than words: Frequency effects for multi-word phrases, *Journal of Memory and Language* 62/1: 67–82.

Blumenthal-Dramé, A. (2012). *Entrenchment in Usage-Based Theories. What Corpus Data Do and Do Not Reveal About the Mind*. Berlin: de Gruyter.

Bolinger, D. (1968). Entailment and the meaning of structures. Glossa, 2(2): 119-127.

Cappelle, B. and Depraetere, I. (2016). Short-circuited interpretations of modal verb constructions. Some evidence from The Simpsons. *Constructions and Frames*, 8(1): 7–39.

Coates, J. (1983). *The semantics of the modal auxiliaries*. London: Croom Helm.

Correia Saavedra, D. (2019). Measurements of Grammaticalization: Developing a quantitative index for the study of grammatical change. PhD Dissertation, Université de Neuchâtel.

Croft, W. (2001). *Radical Construction Grammar. Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

Davies, M. (2008). The Corpus of Contemporary American English (COCA): 400+ million words. Available online at http://corpus.byu.edu/coca.

Davies, M. (2010). The Corpus of Historical American English (COHA): 400+ million words, 1810– 2009. Available online at http://corpus.byu.edu/coha.

Diessel, H. (2019). *The Grammar Network. How Linguistic Structure is Shaped by Language Use*. Cambridge: Cambridge University Press.

Evert, S. (2009). Corpora and collocations. In Lüdeling, A. and Kytö, M. (eds.), *Corpus Linguistics: An International Handbook*, Vol. 2. Berlin/New York: Mouton de Gruyter, pp. 1212–1248.

Facchinetti, R., Krug, M., and Palmer, F.R. (eds.) (2003). *Modality in Contemporary English*. Berlin: de Gruyter.

Firth, J.R. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*. Blackwell, Oxford, pp. 1–32.

Flach, S. (in press). Beyond modal idioms and modal harmony: A corpus-based analysis of gradient idiomaticity in modal-adverb collocations. *English Language and Linguistics*.

Furmaniak, G. and Larreya, P. (2015). On the uses of *would* in epistemic contexts. In Zamorano-Mansilla, J.R., Maiz, C., Dominguez, E., and Martin De la Rosa, V. (eds.), *Thinking Modally: English and Contrastive Studies on Modality*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 105-124.

Goldberg, A.E. (1995). *Constructions. A Construction Grammar Approach to Argument Structure*. Chicago: Chicago University Press.

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

Heylen, K., Wielfaert, T., Speelman, D. and Geeraerts, D. (2015). Monitoring Polysemy. Word Space Models as a Tool for Large-Scale Lexical Semantic Analysis. *Lingua*, 157: 153-172.

Hilpert, M. (2008). *Germanic future constructions. A usage-based approach to language change*. Amsterdam: John Benjamins.

Hilpert, M. (2013). Die englischen Modalverben im Daumenkino: Zur dynamischen Visualisierung von Phänomenen des Sprachwandels. *Zeitschrift für Literaturwissenschaft und Linguistik*, 42: 67-82.

Hilpert, M. (2016). Change in modal meanings: Another look at the shifting collocates of *may*. *Constructions and Frames*, 8(1): 66-85.

Hilpert, M. (2019). *Construction Grammar and its Application to English*. 2nd edition. Edinburgh: Edinburgh University Press.

Hilpert, M. and Correia Saavedra, D. (2017). Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. *Corpus Linguistics and Linguistic Theory.*

Hohaus, P. (2019). Subordinating Modalities – A Quantitative Analysis of Syntactically Dependent Modal Verb Constructions. PhD Dissertation. University of Hannover.

Kiela, D. and Clark, S. (2014). A Systematic Study of Semantic Vector Space Model Parameters. *Proceedings of EACL 2014, Second Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, Gothenburg, Sweden, pp. 21-30.

Krug, M. (2000). *Emerging English modals: A corpus-based study of grammaticalization*. Berlin: de Gruyter.

Lebani, G. and Lenci, A. (2016). "Beware the Jabberwock, dear reader!" Testing the distributional reality of construction semantics. *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex-V)*, pp. 8-18.

Leech, G. (1992). 100 million words of English: the British National Corpus. *Language Research*, 28(1): 1–13.

Levshina, N. (2015). *How to do Linguistics with R: Data exploration and statistical analysis*. Amsterdam: John Benjamins.

Palmer, F.R. (1990). *Modality and the English Modals*. 2nd edition. London: Longman.

Palmer, F.R. (2003). Modality in English: Theoretical, descriptive and typological issues. In Facchinetti, R., Krug, M., and Palmer, F.R. (eds.) 2003. *Modality in Contemporary English*. Berlin: de Gruyter, pp. 1-20.

Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1): 97-124.

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

Turney, P.D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics, *Journal of Artificial Intelligence Research*, 37: 141-188.

Wheeler, E.S. (2005). Multidimensional scaling for linguistics. In Koehler, R., Altmann, G., and Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook*. Berlin: De Gruyter, pp. 548–553.