

Corpus-informed descriptions: English verbs and their collocates in science abstracts

Laura Hartwell

▶ **To cite this version:**

Laura Hartwell. Corpus-informed descriptions: English verbs and their collocates in science abstracts. Etudes en didactique des langues, 2012. hal-02955551

HAL Id: hal-02955551
https://hal.science/hal-02955551

Submitted on 2 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Corpus-informed descriptions: English verbs and their collocates in science abstracts

Laura M. HARTWELL
Maîtresse de conférences
Laboratoire LIDILEM – Université Grenoble I



In *Modes of Meaning* (1951/1957), Firth proposed an innovative approach to descriptive linguistics that embraces multiple levels of creating meaning including social context, syntax, vocabulary, phonology, and phonetics. He posited that the "collocation" of a word is part of its meaning and this within a particular literary form or genre. He made explicit the position of words that create meaning: "Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words" (1951/1957: 196). His framework contrasted with Chomsky's perspective and others' that linguists are concerned with the possible infinite generation of grammatical sentences stemming from human mental faculties. His approach suggests that language is produced in the mind and should be the center of study instead of existing texts. Performance data found in corpora are considered limited in that they fail to incorporate possible, but as yet unsaid, utterances (McEnery & Wilson, 1996; Halliday, 2004; Yallop, 2004). This paper hypothesizes that an understanding of how meaning has been created within specific corpora, and notably through collocation, is essential to developing quality teaching materials for learners of English as a foreign language as corpora can be a "supreme" tool for the observation and analysis of important quantities of natural language (Gilquin & Gries, 2009).

In the late 1950's, Randolph Quirk's *Survey of English Usage* became the first extensive language data collection project created for empirical study. In the following decades, Michael Halliday and John Sinclair propounded the importance of corpus studies. The first empirical study, the Office for Scientific and Technical Information (Osti) report included key notions such as terminology, text register, collocations and their patterns, word frequency, lexical items and statistical methods (Sinclair *et al.*, 2004). They also found that the span, that is to say, the distance from the node is an important factor of collocation. They noted that the frequency of certain common words depend on the type of text, for example, *the* was the most frequent word in both their spoken and scientific texts.

However *I* was ranked second in their corpus of general spoken texts, but ranked 241st in their corpus of written scientific texts (*ibid*.: 58). Since their study, the technical capacities to analyze collocations have dramatically progressed, while an attention to the importance of genre has also expanded (Swales, 1990; Biber *et al*, 1998; Gledhill, 2000). Today, we have not only *corpus-based* studies, which rely uniquely on empirical data drawn from corpora, but also *corpus-driven* studies which depend on corpus methodology before "intellectually processing" the data (Teubert, 2004).

This study examines verbs and their collocations in a corpus of medical and biology abstracts in English found in the on-line corpus Scientext[1]. The frequencies of both lexical and modal verbs are examined. Accepted categories of modal auxiliary verbs vary. However, semantic notions inherent to modality, often categorized as *dynamic*, *deontic*, and *epistemic*, include ability, necessity, obligation, permission, possibility, and hypotheticality (Collins, 2009; Kennedy, 2002; Nuyts, 2006). Palmer considers modality as "the grammaticalization of speakers' (subjective) attitudes and opinions" (1986: 16). This was echoed by Halliday's (1970/2005: 182) conception of modality as a form of speaker participation in the speech event stemming from the "interpersonal" function of language. Hyland (1995, 1996) highlights the role of modality in hedging, a feature that permits "precision, caution, and diplomatic deference". These are necessary ingredients to be a member of a scientific discourse community.

Scientific abstracts contain a series of rhetorical and structural aspects. Cremmins (1982) highlights purpose, scope, methods, results, or conclusions and recommendations as key components of empirical research abstracts. Furthermore, Pho (2011) suggests that abstracts of empirical studies in the fields of applied linguistics and educational technology include presenting the research, describing the methodology, summarizing the findings, and discussing the research – all of which can be identified through a cluster of linguistic features. For Gledhill (2000: 165), the salient lexical items of abstracts in pharmaceutical studies include verbs related to the data (*correlated, decreased, increased*) and reporting of past research (*studied, suggest*). Abstracts are brief but dense texts that require specific language and conceptual capacities. As Osborne (2011: 295) concludes in his study of English learners based on the PAROLE corpus, "rather than the ability to provide detail, it is often the capacity to introduce, synthesise and conclude a description" that is characteristic of fluent speakers who make more efficient syntactic and lexical choices.

Descriptive grammar analyses are essential to language teaching (Kennedy, 2002; Oakey, 2002) and especially within contexts of language learning for specific purposes (Gledhill, 2011). McEnery and Wilson (1996) refer to the studies of Holmes (1988), Kennedy (1987 a & b), Ljung (1990), and Mindt (1992) who have compared the vocabularies or grammatical structures of non-

[1] <http://scientext.msh-alpes.fr/scientext-site-en/spip.php?article9>.

empirically-based textbooks to data derived from corpora analysis. They highlighted the substantial differences between language use as empirically revealed through corpora study and the descriptions found in textbooks.

> Some textbooks have been found to gloss over important aspects of usage or variations in usage, and sometimes textbooks may even foreground less frequent stylistic choices at the expense of more common ones. The more general conclusion which scholars such as Mindt and Kennedy have drawn from these exercises is that non-empirically based teaching materials can be positively misleading and that 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, 1996: 104).

Hartwell (2011) also noted a lack of attention in two textbooks designed for students in the sciences and technologies to the modal verb *may*, common to hedging, while the less common *must* is emphasized. Furthermore, Henderson & Barr (2010) found the comparison of a corpus of student writing in psychology to a corpus of published research articles useful for supporting teaching and learning.

Methodology

The on-line corpus Scientext includes published and unpublished works in both French and English (Tutin *et al*, 2009; Falaise *et al*, 2011). The data of the study discussed here were gathered from the 787,276 words from the abstracts of 3,381 research articles in English. The peer-reviewed articles, collected by the LiCorn team at the Université de Bretagne-Sud, were originally published on-line by the independent editor BioMed Central comprising sixty-two subthemes from the fields of biology and medicine, such as medical genomics, genomics, bioinformatics, genetics and women's health. Scientext has three integrated search modes: semantic search (semantic grammars), assisted search (parts of speech and syntactic relations) and advanced search (queries with grammars).

The first step of this study was to conduct an assisted search for the modal verbs under the categories "form" and "lemma". However, as the parsing software Syntex identifies modal verbs only as part of a unit with a lexical verb, this search provided limited results. Second, "tensed verbs" were searched among the "verbs" using the "category" option of the assisted search mode. After manually removing unwanted nouns, a total of 23,970 entries of 542 different lexical verbs were found. Among these 542 verbs, the 50 most frequently occurring verbs of these 542 verbs constituted approximately ninety percent of all the verbs. Third, these 50 most frequent verbs were then searched using the "lemma" option of the assisted search mode. This third step revealed a total of 35,704 tokens of these most common 50 verbs, including past participles used to modify a noun, as the word *reduced* in the following quote.

> We found a **reduced** birth weight for the offspring of mothers who had a PCB concentration ≥ 25 microg / L (adjusted birth weight = 2,958 g, p = 0.022).

This search method also revealed an additional 287 modal verbs for a total of 1942 modal verb tokens (cf. Appendix 1). The difference of results between the search methods may be explained by the different objective of each search. For example, the second search included only tensed verbs. It also contained a large quantity of nouns that were manually eliminated. The third search included only the top 50 verbs. However, it included a wider range of verb forms of each lemma. The highest figures for each modal verb were included in the final results.

Then, the lexical collocations of the two verbs *provide* and *play* were examined in detail. The frequency of modal verbs, nouns, adjectives, and adverbs were studied within a five-word span to the right and the left of the nodes *provide* and *play*. Several complementary tests of independence were conducted for certain collocations of the verb *provide*, including a Pointwise Mutual Information test (Biber *et al*, 1998), a t-test (Hunston & Francis, 1999: 231), a log-likelihood (Ellis & Simpson-Vlach, 2009; Sinclair *et al.*, 2004), and a Mutual Information test. The Chi-squared test was not conducted as it is considered unreliable for small frequencies (*ibid.*, 2004). The scores of several statistical tests are included as they display slight differences. They offer the reader the opportunity to compare the scores of each test and also to compare them with other corpora studies that rely upon only one of these tests. For example, Hunston & Francis employ the t-score software available with the corpus Bank of English (1999: 231). In contrast, Biber *et al* consider t-score software, such as that in concordancing packages found in Corpus Bench, inappropriate for identifying a single word's most important collocates (1998: 268).

Finally, the lexico-grammatical patterns of the verb *play* are examined. Hunston & Francis define a word's pattern as "all of the words and structures which are regularly associated with the word and which contribute to its meaning" (2000: 37). Taking this approach, lexis and grammar are not treated as separate categories. Lexical patterns, woven into grammatical structures, are essential to understanding a language, as words are "primed" for use by fluent speakers (Hoey, 2005). For example, Ellis & Simpson-Vlach (2009) found that native speakers are "tuned" to the regularities of formulaic expressions as these speakers predict the endings of phrases with higher Mutual Information scores.

Results

By far, the most common verb was the lemma "be", with 9,984 tokens of different forms found. Almost one-third of these took the form "is" at 3,346 tokens. This figure does not include lemmas of "be" found in verbal constructions of other lexical verbs, as in the passive voice found in the quote.

> Testosterone and estrogen **are** no longer **considered** male only and female only hormones.

Although second in frequency, *use* was far behind with only 3,263 tokens, followed by *have* with 1,654 tokens. Several of these 50 most frequent verbs were

related to scientific research, including *show, compare, suggest, report, determine, examine, describe, investigate, indicate, demonstrate, reveal, confirm, support, contribute, measure,* and *discuss.* A second category of verbs is related to the cause and effect results, including *increase, reduce, decrease, affect, lead, improve,* and *remain.* Finally, other verbs are directly related to the fields of medicine and biology, including *induce, regulate,* or *inhibit* (Appendix 3).

Modal verbs

A search on Scientext also facilitated an analysis of the 1,943 modal verbs found within the abstracts. It identifies modal verbs that are part of a passive voice construction as in the following quote on radiation exposure.

> Radiation exposure **may be associated** with risks to physician, patient and personnel. While there have been many studies evaluating the risk of radiation exposure and techniques to reduce this risk in the upper part of the body, the literature is scant in evaluating the risk of radiation exposure in the lower part of the body.

The parsing system identifies modal verbs even when they are separated from the lexical verb as in this quote on gender awareness.

> Physicians' degree of gender awareness **may**, as one of many factors, **affect** working climate and the distribution of women and men in different specialties. Therefore, to improve working climate and reduce segregation we suggest efforts to increase gender awareness among physicians, for example educational programs where continuous reflections about gender attitudes are encouraged.

These two examples also draw attention to the notion of hedging (Hyland, 1995 & 1996) in which researchers position themselves within a discourse community by the acknowledgment of opposing claims. Precision and caution are also rhetorical elements in the previous quote about radiation exposure. The authors create their research niche by noting the abundant attention paid to upper body studies of radiation exposure, while highlighting their consideration to lower body exposure. The risk of upper body exposure is acknowledged, but the previous lack of attention to lower body exposure is put forward. In an example on gender awareness, the *may affect* diplomatically introduces the notion of gender awareness. In the following sentence, the authors reaffirm the validity of gender awareness by suggesting appropriate educational programs.

In the current study, *can* was the most frequent modal verb and constituted more than one-third (37.5%) of the modal verbs, its most common pattern being *can be used*, as in the quote on HIV detection, followed by *can be*, as in the following quote on self-hypnosis. It is noteworthy that although the lemma *be* (9,984 tokens) was over three times more frequent than the lemma *use* (3,263 tokens), there were nearly twice as many tokens of *can be used* (121 tokens) than *can be* (61 tokens).

> They further **can be used** for improvement of oligo-probe based HIV detection techniques.

Self-hypnosis **can be** a useful skill in the treatment of a patient with anxiety and asthma.

The results showed a variation between the frequency of modal verbs found in Collin's Corpus of general oral and written texts in English (Collins, 2009; Aijmer & Simon-Vandernbergen, 2008) and the specific sections of texts from Natural and Pure Sciences and from Applied Sciences from the British National Corpus (BNC) (Kennedy, 2002). For example, the most common modal verb within the Collin's Corpus and the BNC's Applied Sciences section was *will* (24% and 27.5% respectively). However, *will* accounts for only 11.7% of the modal verbs in the Scientext corpus, which is closer to the 17.6% found in the Natural and Pure Sciences corpus of the BNC. In contrast, within both the Scientext corpus and the Natural and Pure Sciences texts of the BNC, *can* was the most frequent modal verb (37.5% and 27.3% respectively). The second most frequent modal verb found in this study was *may* (17.3%), which was almost three times more present than in the Collin's Corpus, but similar to that of the Natural and Pure Sciences section of the BNC (17.4%) (cf. Table 1).

	can	***may***	***could***	***will***	***should***	***might***	***would***	***must***	***shall***
Scientext	**729 37.5%**	**336 17.3%**	242 12.5%	227 11.7%	146 7.5%	113 5.8%	76 3.9%	74 3.8%	1 0.05%
Collin's Corpus	7,663 21.6%	2,261 6.4%	3,557 10%	**8,505 24%**	2,432 6.9%	1,499 4.2%	**7,775 22%**	1,367 3.9%	343 1%
Natural and Pure Sciences BNC	27.3%	**17.4%**	7.5%	17.6%	7.3%	4%	11.8%	5.4%	1.2%
Applied Sciences BNC	22.6%	12.2%	8%	**27.5%**	8.3%	3.2%	12.2%	5%	0.4%

Table 1 – **Frequency of modals in Scientext and Collin's Corpus and sections of the British National Corpus (BNC)**

For the other modal verbs, *could, should, might, must* and *shall,* there were similar rates of frequency across the different corpora. Hence, there are variations in the use of certain modal verbs in scientific abstracts as compared to scientific texts and especially as compared to general English texts.

Tense and modal verbs with **provide** *and* **play**

The general frequencies displayed in Table 1 do not imply that individual verbs are employed with the same frequency even within science abstracts. Some verbs offer little variation, but the differences for some verbs, including tense, use with modal verb, and collocation are important, such as with the verbs *provide*

and *play*. A closer look at these two verbs suggests that the frequencies of tense and modal verbs vary according to the verb. As noted in Table 2, the vast majority of the tokens of these two verbs were in the present tense (70% and 81.7% respectively). *Play* (5.2%) occurred in the past and present perfect tense, but *provide* did not. In contrast, *provide* (3.7%) was conjugated with *will*, but *play* was not. The modal verb *may* occurred three times more often with the verb *play* (18.9%) than it did with *provide* (6.1%).

	present	(*has/have*) *-ed*	*can*	*could*	*may* (*have -ed*)	*will*	*would*
provide 458	374 81.7%	0	18 3.9%	9 2%	28 6.1%	17 3.7%	5 1.1%
play 233	163 70%	12 5.2%	3 1.3%	4 1.7%	44 18.9%	0	1 0.4%

Provide, less than 1%: should (1), must (2), might (3), did (1).
Play, less than 1%: -ing (5), might (2).

Table 2 – **Frequency of tense, aspect and modal verbs**

The verbs *provide* and *play* also displayed contrasting collocational patterns. While *provide* was linked to a wide range of nouns, *play* was significantly collocated with the noun *role*, which in turn was collocated with a specific range of adjectives.

The nouns collocated with *provide* (cf. Appendix 4) were mainly associated with three categories of meaning: the first related to data (*evidence* 48, *information* 33), the second related to method or means (*tool* 22, *means* 12, *method* 18), and the third related to understanding (*insight* 31, *explanation* 7). These nouns were collocated to a range of adjectives, a common collocation being *useful information* as in the following quote on patient beliefs.

> Most patients believe the test **will provide useful information** in making treatment decisions, despite probable lack of insurance coverage, and appear willing to experience some discomfort for the overall gain of the results obtained from undergoing the session.

This example also highlights the use of the modal verb *will* that was relatively frequent with this verb. In comparison, the following quote on colon cancer shows the node *provide* with both the modal verb *may* and the compound *exposure estimates*. Though compounds are frequent in scientific discourse, they are beyond the scope of this study.

> Use of colon cancer controls **may provide valid exposure estimates** in studies of many occupational risk factors for cancer, but not for studies on exposure related to farming.

Frequency of collocation with **provide**

Frequency of collocation can be evaluated through a range of statistical tests. A high frequency of occurrence with a given node does not always indicate a high level of collocation. For example, *method-s* was found 18 times in collocation with the node *provide,* but was present 973 times in the corpus. In contrast, *means* occurred 12 times with *provide*, but was present only 50 times in the corpus. Four statistical tests suggest that the word *means* occurs with the node *provide* with greater frequency than *method-s* occurs with the same node (Table 3). It should be noted that the Pointwise Mutual Information test and the t-test (Hunston & Francis, 1996) give higher totals for words of low frequency. For this reason, these tests place *insight-s* as having the strongest co-occurrence, however the log-likelihood (Ellis & Simpson-Vlach, 2009) and Mutual Information tests place *evidence* as having the greatest frequency of collocation with *provide.*

Collocate	*Tokens with node*	*Tokens*	*Pointwise MI*	*t-test*	*Loglike*	*Mutual Information*
insight-s (right only)	31	70	9.882	170.86	550.1	0.00028
evidence	48	453	7.819	103.65	**627.8**	**0.00032**
means (way) (right only)	12	50	**8.998**	78.19	186.4	0.00010
information (right only)	33	500	7.135	67.64	**383.1**	**0.00020**
may (left only)	27	336	**7.420**	67.58	**328.7**	**0.00017**
tool-s	22	307	**7.255**	57.58	**260.0**	**0.00013**
result-s	20	1016	5.391	28.28	161.1	0.00008
method-s	18	973	5.301	26.96	141.7	0.00007
model-s	11	862	4.765	16.66	74.8	0.00004
analysis	10	1509	3.820	11.04	49.9	0.00003

Table 3 – **Collocates of the word *provide* (454)**

Some words collocate only to the left or only to the right of the node. In Table 3, we can see that the collocates *insight-s, means,* and *information* are only found to the right of the node *provide.* In contrast *may* is only found to the left of the node. The other collocates can be found both to the left and to the right of the node.

Patterns with **play**

The patterns in conjunction with *play* contrast to those with *provide. Play* was present 233 times in the corpus, all but ten of these tokens were collocated with *role.* The most frequent adjective was *important* (64 tokens) as in the following quote on breastfeeding.

> Breastfeeding **plays a very important role** in protecting infants from intestinal infections.

This collocation fell mainly within the basic pattern *PLAY a/n […] [adj] role* (212 tokens). The brackets and ellipse […] indicate that there may be a word or several words at the given position within the phrase. The adjectives found within this pattern fall into two main categories: (1) related to *level* or *quantity*, such as *important* or (2) *critical* or adjectives having a *qualitative* function, as in *physiological* or *biological* (Appendix 4). Although the lemma *role* (637 tokens) occurs within the corpus without relation to the verb *play*, the two words remain significantly collocated (Pointwise M1 10.499, t-test 566.41, Loglike 4692.1, MI 0.00242).

This basic lexicogrammatical pattern encompassed a series of parallel patterns. These patterns comprise the collocation of *play* and *role* in five specific sequences, as seen in Patterns 1-5.

1. […] play (lemma) a/n […] role-s (183 tokens, 130 with adjectives, 45 with modal verbs);
2. may play a/n […] role (37 tokens, 18 with adjectives);
3. […] play-s a/n […] role-s in the (67 tokens);
4. […] play-s a […] role-s in –ing (32 tokens, 22 with adjectives, 6 modal verbs);
5. role played by (5 tokens).

In Pattern 1, the lemma *play* is preceded in 45 occurrences by a modal verb. As can be seen in Pattern 2, 37 of these modal verbs are *may,* as in this quote on oxidative stress.

> Oxidative stress **may play a critical role in the** vascular disease of end stage renal failure and hemodialysis patients. Studies, analyzing either discrete analytes and antioxidant substances, or the integrated total antioxidant activity of human plasma during hemodialysis, give contradictory results.

As discussed (*infra*), *may* evokes a notion of possibility, but this sequence, like others found within this corpus, contained an adjective with a strong connotation. Other adjectives were *key, crucial, critical, pivotal,* and *important.* The preposition *in* was often followed either by *the* or a gerund in an *–ing* form. *In* and *the* are common grammatical words. Gledhill (2000) found that *the* was the most common word in his study of pharmaceutical research articles and *in* was ranked fourth, after *of* and *and.*

Frequent grammatical words present a specific challenge to language learners, including those with a relatively good command of the language. More than half of these patterns end with "in". This suggests that expressions such as

"play a role" should be given to English learners in the more complete pattern "play a role in", so that those difficult, often untranslatable prepositions be learned in context. Learners would benefit from learning these grammatical words in relation with frequent lexical verbs. For example, Appendix 5 lists the 50 most frequent verbs in their most frequent tense and some recurrent patterns. We find *compared with, induced by, involved in,* and *contributes to*. Learning materials based on patterns, supported by data from genre-specific corpora may lead to greater fluency.

Conclusion

These results reveal interesting notions about English found in abstracts of research articles in the fields of medicine and biology. These notions should be taken into account within contexts of teaching and learning to members of this discourse community. It has been seen that the 50 most frequent verbs found within the corpus accounted for approximately 90% of the verbs. Their acquisition is essential for learners to obtain a minimum level of comprehension.

Second, modal verbs within this corpus do not follow the frequencies of general English. The modal verbs *can* and *may* have higher frequency in this context of academic research, and should be given specific attention, including their use in the passive voice. This study confirms the use of *may* as a means of hedging when presenting results. The rhetorical nuances of *may* and other modal verbs offer a challenging, but essential task for both teachers and learners, who seek to become articulate members of a scientific discourse community. Furthermore, the frequency of tense and the collocations are not uniform across all verbs. Hence, learning materials would better mirror English for medicine or biology if these forms and collocations were taken into account.

Finally, this study identifies the grammatical patterns that may be useful for improving fluency, because the mastery of these patterns will help learners replicate English within highly competitive disciplines such as biology or medicine. Instead of centering learning on isolated vocabulary or general rules that aid in the analysis of an utterance, identifying patterns can help learners to draw links between lexis and grammar. Instead of memorizing, for example, the 50 most frequent verbs, these verbs should be studied in connection with their collocations.

Appendixes

Appendix 1: **Frequency of modal verbs found using different Scientext search methods**

	will	*would*	*can*	*could*	*may*	*should*	*must*	*might*	*shall*

Assisted search of all tensed verbs and by modal verb	164	40	729	242	209	146	74	51	0
Assisted individual search of the 50 most common verbs	227	76	363	119	336	81	26	113	1

Appendix 2: **Fifty most frequent verbs**

1	*be*	9,984		26	*decrease*	399
2	*use*	3,263		27	*regulate*	378
3	*have*	1,654		28	*make*	351
4	*show*	1,424		29	*predict*	349
5	*compare*	1,063		30	*affect*	337
6	*identify*	1,026		31	*occur*	331
7	*increase*	843		32	*allow*	310
8	*suggest*	813		33	*lead*	310
9	*report*	735		34	*improve*	309
10	*determine*	729		35	*give*	305
11	*induce*	672		36	*cause*	302
12	*express*	655		37	*encode*	275
13	*involve*	559		38	*appear*	270
14	*examine*	538		39	*represent*	259
15	*include*	503		40	*remain*	247
16	*contain*	493		41	*inhibit*	245
17	*describe*	483		42	*activate*	243
18	*investigate*	461		43	*play*	232
19	*reduce*	459		44	*confirm*	218
20	*provide*	456		45	*support*	206
21	*require*	445		46	*contribute*	204
22	*indicate*	439		47	*measure*	186
23	*present*	432		48	*become*	181
24	*demonstrate*	430		49	*discuss*	141
25	*reveal*	418		50	*consist*	139
						35,704

Appendix 3: **Collocates and words occasionally found with the verb *provide***

Right context collocates

Evidence (45), information (33), insight (31), tool (16), means (12) (*does not include the meaning of "average"*), method (10), model (7), resource (7), basis (7), system (9), estimate (8), alternative (7), opportunity (8), explanation (7),

overview (7), clue (5), analysis (4), assessment (3), graphical representation (3), guide (3), answer (2), image (2), interface (2), map (1), description (4), picture (2), result (5), background (2), foundation (1), reflection (1).

Left context collocates
Result (15), method (8), tool (6), analysis (6), model (4), evidence (3), hypothesis (2), estimate (2), map (2), alternative (1), interface (1).

Appendix 4: **Collocates and occasional modifiers of the pair *play* and *role***

Adjectives related to ***level*** or ***quantity***: important (64), critical (15), key (11), significant (10), crucial (8), pivotal (7), central (5), fundamental (2), prominent (2), vital (2), essential (3), no primary (1), cardinal (1), major (5), likely (1), different (2), at most a subtle (1), only a minor (1), diverse (1), more than one (1), multiple (1).
The adverbs modifying ***important*** include increasingly (1), more (2), very (2), such an (1), most (1).

Adjectives having a ***qualitative*** function: physiological (2), biological (1), causal (1), direct (1), active (1), antagonistic (1), as yet an recognized (1), an immune and inflammatory (1), an immune modulatory (1), incompletely understood (1), more specialized (1), no catalytic (1), an evolutionarily conserved and critical (1), the same role (1).

Patterns containing both ***important*** and a qualitative adjective, (important [adj] role): regulatory (2), functional (1), pathogenic (1), physiological (1), pathological (1).

Appendix 5: **Frequent verbs in biology and medical abstracts**

Top 50 most frequent verbs presented in their most frequent tense, frequent collocation with a modal verb (if applicable) and some frequent patterns

is/are/may be, had, using/can be used to, showed that/a-n/no, compared to/with, identify, increased, suggest-s that, determine-s, reported, induced by, expressed, involved in, we (also) examined, included, containing, we describe, investigate-s whether, reduced, provide-s/may provide, required for, indicate-s that, we/this study present-s/patients presenting, we/this study demonstrated, revealed that, decreased, gene/up/down regulated, make-s, predicted, affect-s/may affect, occurred/can occur, leading to, allow-s, improve-s/may improve, a given, caused by, genes encoding, appear-s to, represent-s a, remain-s + adjective (unclear), inhibited, activated, play-s/may play a role in, confirmed, support-s, contribute-s to the/may contribute, to measure, become-s/has become, discuss-es the, consist-s of.

Other frequent verbs
review, offer, depend, exist, aim, predict, seem, consist, review, offer, introduce, depend, enable, bind, utilize, reflect, interact, vary, focus, means, continue, facilitate, promote, differ, highlight, summarize, exhibit, generate, prevent, stimulate, comprise, take, alter, constitute, mediate, modulate, assess, rely, confer, evaluate, permit, produce, suppress, carry, help to maintain, illustrate, resemble, yield, correspond, localize, serve, act, develop, explore, hold, incorporate, catalyze, combine, correlate, cover, exert, extend, fail, imply

Corpus

Scientext On-line Corpus. Consulted from May to December 2011: <http://scientext.msh-alpes.fr/scientext-site-en/spip.php?article9>.

Bibliographical references

AIJMER, KARIN & ANNE-MARIE SIMON-VANDERNBERGEN. 2008. *Topics in English Linguistics: The Semantic Field of Modal Certainty: A Corpus-Based Study of English Adverbs.* Berlin: Mouton de Gruyter.

BIBER, DOUGLAS, SUSAN CONRAD & RANDI REPPEN. 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

COLLINS, PETER. 2009. *Modals and Quasi-modals in English.* Amsterdam: Rodopi B. V. Editions

CREMMINS, EDWARD T. 1982. *The Art of Abstracting.* Philadelphia: ISI Press.

ELLIS, NICK C. & RITA SIMPSON-VLACH. 2009. Formulaic language in native speakers: triangulating psycholinguistics, corpus linguistics, and education. GRIES, STEPHAN TH. & ANATOL STEFANOWITSCH (eds.). *Corpus Linguistics and Linguistic Theory* 5 : 1, 61-78.

FALAISE, ACHILLE, AGNES TUTIN & OLIVIER KRAIF. 2011. Exploitation d'un corpus arboré pour non spécialistes par des requêtes guidées et des requêtes sémantiques. *Proceedings of TALN, Montpellier 2011. Language and Linguistics* 14 : 2, 187-215.

FIRTH, J.R. 1957. Modes of Meaning. *Papers in Linguistics 1934-51.* Oxford: Oxford University Press, 190-215.

GILQUIN, GAETANELLE & STEPHAN T. GRIES. 2009. Corpora and Experimental Methods: A State-of-the-Art Review. GRIES, STEPHAN T. & ANATOL STEFANOWITSCH (eds.) *Corpus Linguistics and Linguistic Theory* 5 : 1, 1-26.

GLEDHILL, CHRISTOPHER J. 2000. *Collocations in Science Writing.* Tübingen: Gunter Narr Verlag.

GLEDHILL, CHRISTOPHER J. 2011. The "lexicogrammar" approach to analysing phraseology and collocation in ESP texts. *ASp* 59, 5-23.

HALLIDAY, M. A. K. 1970/2005. Functional Diversity in Language, as Seen from a Consideration of Modality and Mood in English. *Studies in English Language.* London: Continuum International Publishing, 164-204.

HALLIDAY, M. A. K. 2004. Lexicology. HALLIDAY, M. A. K. WOLFGANG TEUBERT, COLIN YALLOP & ANNA ČERMÁKOVÁ (eds.) *Lexicology and Corpus Linguistics.* London: Continuum, 1-22.

HARTWELL, LAURA M. 2011. Learning on-line about modality in written and oral English for science and technology. *Proceedings of ICT for Language Learning, 4th edition*. <http://www.pixel-online.net/ICT4LL2011/common/download/Paper_pdf/SLA21-121-FP-Hartwell-ICT4LL2011.pdf> (consulted November 201).

HENDERSON, ALICE & ROBERT BARR. 2010. Comparing indicators of authorial stance in psychology students' writing and published research articles. *Journal of Writing Research* 2 : 2, 245-265.

HOEY, MICHAEL. 2005. *Lexical Priming: A New Theory of Words and Language.* London and New York: Routledge.

HOLMES, JANET. 1988. Doubt and certainty in ESL textbooks. *Applied Linguistics* 9 : 1, 21-44.

HUNSTON, SUSAN & GILL FRANCIS. 1999/2000. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English.* Amsterdam: John Benjamins Publishing Company.

HYLAND, KEN. 1995. The author in the text: hedging in scientific writing. *Hong Kong Papers in Linguistics and Language Teaching* 18, 33-42.

HYLAND, KEN. 1996. Writing without conviction? Hedging in science research articles. *Applied Linguistics* 17 : 4, 433-454.

KENNEDY, GRAEME. 1987a. Expressing temporal frequency in academic English. *TESOL Quarterly* 21 : 1, 69-86.

KENNEDY, GRAEME. 1987b. Quantification and the use of English: a case study of one aspect of the learner's task. *Applied Linguistics* 8 : 3, 264-286.

KENNEDY, GRAEME. 2002. Variation in the distribution of modal verbs in the British national corpus. REPPEN, RANDI, SUSAN M. FITZMAURICE & DOUGLAS BIBER (eds.). *Using Corpora to Explore Linguistic Variation.* Amsterdam: John Benjamins Publishing Company, 73-90.

LJUNG, M. 1990. *A study of TEFL vocabulary*. Stockholm: Almqvist and Wiksell International.

MCENERY, TONY & ANDREW WILSON. 1996. *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

MINDT, DIETER. 1992. *Zeitbezug im Englischen.* Tübingen: Gunter Narr Verlag.

NUYTS, JAN. 2006. Modality: overview and linguistic issues. FRALEY, WILLIAM (ed.). *The Expression of Modality*. Berlin: Walter de Gruyter, 1-25.

OAKEY, DAVID. 2002. Formulaic language in English academic writing: a corpus-based study of the formal and functional variation of a lexical phrase in different academic disciplines. REPPEN, RANDI, SUSAN M. FITZMAURICE &

DOUGLAS BIBER (eds.) *Using Corpora to Explore Linguistic Variation.* Amsterdam: John Benjamins Publishing Company, 111-129.

OSBORNE, JOHN. 2011. Fluency, complexity and informativeness in native and non-native speech. *International Journal of Corpus Linguistics* 16 : 2, 276-298.

PALMER, FRANK ROBERT. 1986/2001. *Mood and Modality.* Cambridge: Cambridge University Press.

PHO, PHUONG DZUNG. 2008. Research article abstracts in applied linguistics and educational technology. *Discourse Studies* 10 : 2, 231-250.

SINCLAIR, JOHN, SUSAN JONES & ROBERT DALEY. 2004. English collocation studies: the OSTI report. KRISHNAMURTHY, RAMECH (ed.) *Studies in Corpus and Discourse*. London: Continuum.

SWALES, JOHN M. 1990/2004. *Genre Analysis: English in Academic and Research Settings.* Cambridge: Cambridge University Press.

TEUBERT, WOLFGANG. 2004. Language and corpus linguistics. HALLIDAY, M. A. K. WOLFGANG TEUBERT, COLIN YALLOP & ANNA ČERMÁKOVÁ (eds.) *Lexicology and Corpus Linguistics.* London: Continuum, 1-22.

TUTIN, AGNÈS, FRANCIS GROSSMANN, ACHILLE FALAISE & OLIVIER KRAIF. 2009. Autour du projet Scientext: étude des marques linguistiques du positionnement de l'auteur dans les écrits scientifiques. Unpublished document: <http://w3.u-grenoble3.fr/lidilem/labo/file/Lorient_vfinale.pdf>.

YALLOP, COLIN. 2004. Words and meaning. HALLIDAY, M. A. K., WOLFGANG TEUBERT, COLIN YALLOP & ANNA ČERMÁKOVÁ (eds.). *Lexicology and Corpus Linguistics.* London: Continuum, 23-72.

Acknowledgements

The author would like to thank Nolwena Monnier for organizing the Journée d'études LAIRDIL of December 2011. She also sincerely thanks Oliver Karif for the statistical help provided as well as Achille Falaise and Agnès Tutin for the roles they have played in creating Scientext.