Exploring a Corpus of George MacDonald's Fiction

Patrick Maiwald

"Nothing about a literature can be more essential than the language it uses."
C. S. Lewis (1964: 6)

1. The Corpus-stylistic Approach: Ways and Means

In recent years, many of George MacDonald's narrative, poetic, critical and theological works have been made available in digital formats —which has opened up new possibilities for investigating these works. The aim of the present paper is to explore some of the new possibilities by approaching George MacDonald's literary works using a quantitative method of stylistic research that has recently been termed "corpus stylistics" (Sinclair 2007; Mahlberg 2007b). This paper's main inspiration is drawn from examples such as Stubbs's (2005) and Mahlberg's (2007a) corpus-stylistic analyses of works by Joseph Conrad and Charles Dickens, respectively. Corpus stylistics itself is a relatively young discipline that is based on the analysis of literature and literary style by means of tools originally developed by linguists for empirical language research. The most important of these tools are large collections of texts or parts of texts in computer-readable form (text corpora usually millions of words in size), and computer software programs tailored to efficiently analyze such large text corpora, such as *WordSmith Tools version 5* (Smith 2008), which was used primarily for this paper. The research target of corpus linguistics itself is usually the quantitative analysis of particular languages or language varieties, but the existence of specific text corpora also allows for close investigation of the language of certain time periods, genres, or even authors.

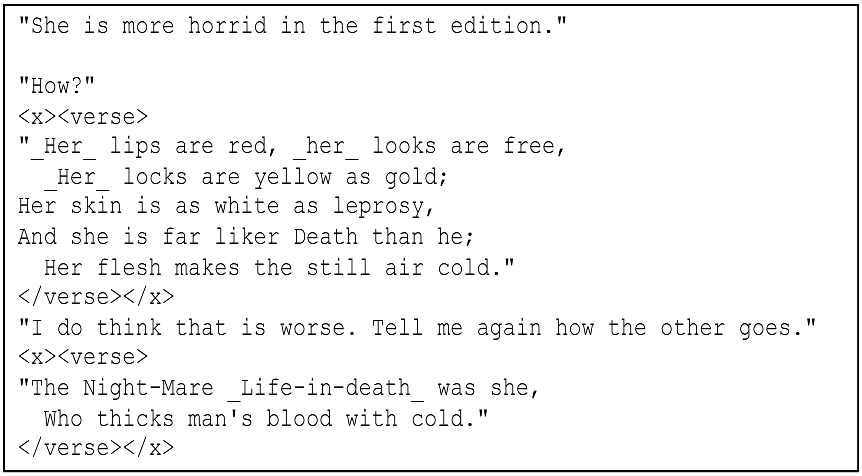
The application of corpus-linguistic tools, or of quantitative methods in general, to literary research questions might seem awkward. George MacDonald himself warns against making empirical methods absolute in his *Unspoken Sermons* (2006: 313), illustrating his point with the example of an empirical scientist who runs the risk of confusing "the facts about" a flower with what MacDonald calls "the truth of" the flower, the latter being the "idea" of a flower in the Platonic sense, compared with which the former is merely "a thing of ways and means." We might suspect a similar, if not even

North Wind 30 (2011): 50-84

a greater, danger to lie in the use of corpus-linguistic methods for the stylistic analysis of literary texts. Fortunately, however, corpus stylisticians generally seem to be aware of the limits and pitfalls of their approach, insisting on the strong need for the researcher's expertise and intuitions in corpus-based analyses of style (cf. Mahlberg 2007b: 222).

Quantitative and statistical methods in stylistic research have been around for centuries, but so have notions such as Rebecca Posner's (1963: 111-112) warning that any study of style is highly "dependent on the intuition, the sensitivity and the depth of experience of its practitioners," so that traditionally the study of style has been the focus of qualitative rather than quantitative studies. "Stylostatistics" or "stylometry," as such quantitative approaches were called around the middle of the past century (cf. Archer 2007: 245), have remained the exception. However, with the wide availability of literary texts in digital form since about 2000, corpus-stylistic studies are on the rise.

A common problem for corpus stylisticians is that it is usually hard to obtain machine-readable, i.e. digital, copies of literary texts, or the rights to use them in research. Fortunately, MacDonald's texts are all in the public domain, and they have been made available in digital forms to a large extent.[1] For the purposes of this paper, a text corpus containing nearly all the narrative fiction written and published by George MacDonald between 1858 and 1897 was compiled from forty-one text files downloaded from the *Project Gutenberg* website. Next, XML-based markup tags were used to mark both the *Project Gutenberg* "header" at the beginning of each text file and the "small print" material at the end of each as "extra-corpus material" in order to allow the software to automatically exclude these parts from the analysis. Within the texts of the novels, obvious quotations from other sources were also tagged as far as it seemed practicable.[2] In addition, tables of content, dedications, epigraphs at the beginnings of chapters, and any verse passages in the texts were tagged to allow their exclusion from the analysis. An example of some tagged text from such a corpus file is given in Fig. 1, in which two characters are discussing a passage from Coleridge in *There and Back*. Care was taken that every text be included only once.[3] The entire text of "If I Had a Father" was tagged as "drama" in order to be left out of the analysis. The resulting corpus—henceforth referred to as the George MacDonald Fiction Corpus (GMDFC)—amounts to roughly 4.5 million words from thirty-nine published works of narrative fiction (mostly novels, and a few collections of shorter tales). The exact contents of the corpus are listed in Appendix 1.



```
"She is more horrid in the first edition."

"How?"
<x><verse>
"_Her_ lips are red, _her_ looks are free,
  _Her_ locks are yellow as gold;
Her skin is as white as leprosy,
And she is far liker Death than he;
  Her flesh makes the still air cold."
</verse></x>
"I do think that is worse. Tell me again how the other goes."
<x><verse>
"The Night-Mare _Life-in-death_ was she,
  Who thicks man's blood with cold."
</verse></x>
```

Fig. 1. An example of tagged text from the GMDFC (File: There and back.txt).

2. Further In: The George MacDonald Fiction Corpus

Once a digital text corpus of MacDonald's works of fiction has been compiled and sufficiently annotated, we can begin to investigate it using corpus analysis software: It contains 4,502,892 running words (tokens) and 42,843 different word forms (types). In the following, the three basic functions of *WordSmith Tools*—namely the automatic generation of word-frequency lists, concordances and key word lists—will be briefly explained and applied to the GMDFC; a quick view at semantic tagging will round off our analysis.

Word-frequency lists, first of all, are a common starting-point in any analysis of a digital text corpus (Stubbs 2005: 11). A search for the most frequent words in the corpus naturally throws up grammatical items such as *the*, *be*, *to*, *of* and *and*, which are always the most frequent words in any English text. However, if we exclude grammatical words (i.e., pronouns, prepositions, conjunctions, determiners and primary verbs) and search only for the most frequent lexical words (defined narrowly as including only nouns, adjectives, and full verbs), the results will be more conclusive (cf. Table 1). The lists given in the tables have been lemmatized, which means that, for example, the word forms *say*, *said*, *saying* and *says* were treated as instances of the same abstract "word" (or lemma), SAY, and their frequencies were added up.[4]

Table 1
The 25 Most Frequent Lexical Items in the GMDFC

Number	Word	Frequency	Texts
1	SAY	27,105	47
2	GO	17,529	48
3	SEE	16,541	48
4	COME	15,638	48
5	KNOW	14,538	47
6	THINK	14,281	48
7	MAN	12,779	47
8	MORE	12,334	48
9	MAKE	12,199	48
10	LIKE	11,778	48
11	LOOK	10,009	48
12	TAKE	9,710	48
13	THING	9,518	47
14	LITTLE	9,500	48
15	TIME	8,594	48
16	TELL	7,941	48
17	FIND	7,915	48
18	GIVE	7,029	48
19	GET	6,991	47
20	GOD	6,852	41
21	FATHER	6,718	47
22	GOOD	6,474	48
23	LIE	6,450	43
24	LOVE	6,189	47
25	OTHER	6,066	48

As is to be seen in Table 1, the most frequent lexical items in George MacDonald's fiction are a set of full verbs that are among the most frequent ones in the English language. Not very surprisingly, the most frequent verb by far is SAY, which acts as a mediator between the frame of narration and the representation of direct speech: the most frequent combinations in which forms of the lemma SAY occur in the corpus are *he said* (3,179 hits) and *she said* (2,215 hits). It also does not take a lot of interpreting to explain the rest of the verbs in Table 1: MacDonald's characters COME and GO, they GIVE, GET and TAKE, and not only do they SAY and TELL about things, but they also SEE, THINK and KNOW.[5] Among the most frequent nouns we find MAN, THING, TIME and GOD—the latter being explicable through the

Christian subject matter of many of the dialogues, especially in the realistic novels (cf. section 4 below). Further down the list we find FATHER (no. 21), DAY (no. 27), MR (no. 30), HAND (no. 31), WAY (no. 32), EYE (no. 38), FACE (no. 40), ROOM (no. 43) and HEART (no. 44). The occurrence of the nouns MAN, FATHER and MR at such prominent places in the word frequency list suggests a prevalence of "male" nouns (and perhaps pronouns) over their "female" counterparts in general. This will briefly be investigated in the following section.

3. Gendered Nouns and Pronouns

A tentative investigation shows that the lemma MAN (12,779 hits) is about three times as frequent in the corpus as the lemma WOMAN (4,332 hits). Of course, we need to be careful since this includes cases in which MAN means "human" or "mankind." However, similar, if less extreme, relationships (in each case, the male counterpart achieves around 60% of the hits per pair) hold between FATHER and MOTHER,[6] SON and DAUGHTER, and BROTHER and SISTER (cf. Fig. 2, where all absolute and relative frequencies are given). Comparisons of MASTER against MISTRESS and UNCLE against AUNT also yield similar results. Thus, we might conclude that there is a slight dominance of male characters in MacDonald's works, at least judging on the basis of such "everyday" items. This seems to be especially true for adult characters, since the quantitative relationship between BOY and GIRL is more balanced (48% vs. 42%). Among the "everyday" items, HUSBAND and WIFE are an exception in that WIFE is more frequent than HUSBAND (60 vs. 40%).

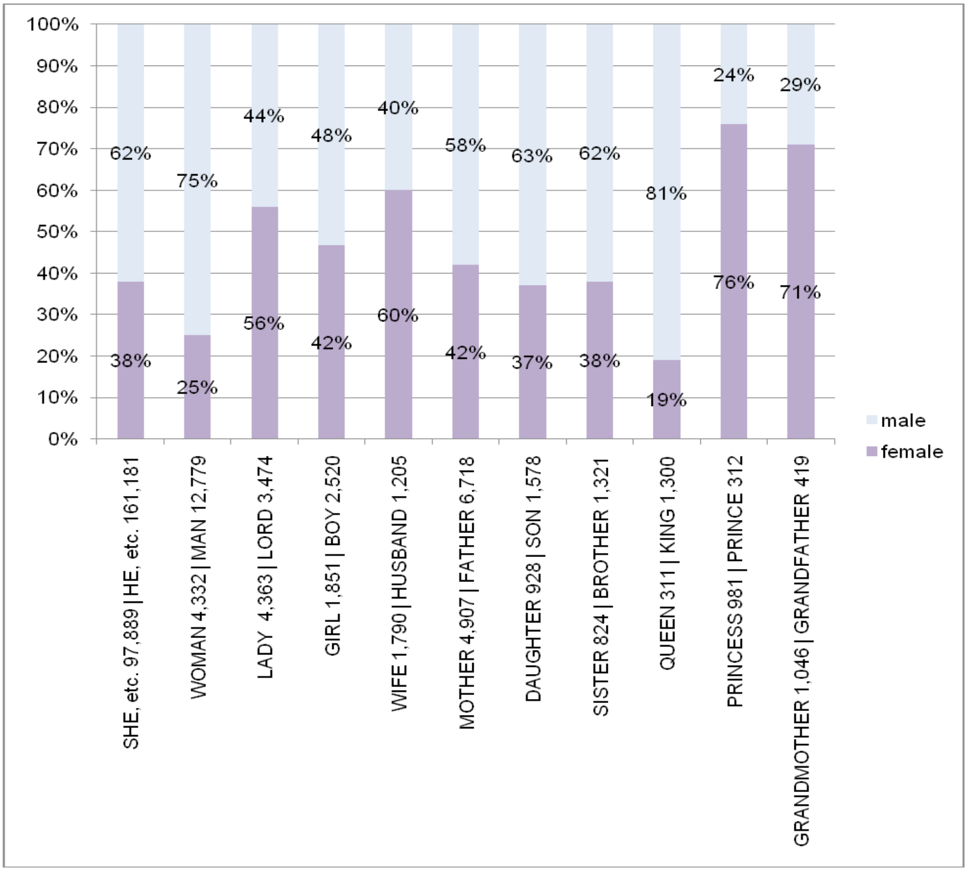


Fig. 2. The relative overall distribution of some male-female complementary nouns and pronouns in the GMDFC (in absolute numbers and percentages)

One might conclude that in MacDonald's fiction, female characters (women, sisters, daughters, etc.) are generally less present than male characters, but that when they are mentioned, there is a comparatively high chance that they will be referred to as *wives* (as opposed to men being referred to as *husbands*). This makes sense if we assume that woman characters are generally less active and thus less likely to be mentioned by name than their male counterparts, or in other words, that a sentence like (1) is more likely to occur than a sentence like (2).

(1) Joseph rushed in to his wife who had been standing at the window anxiously waiting the result of the long colloquy.

(File: At the Back of the North Wind.txt, my emphasis)

(2) Lady Margaret led her to her husband, who kissed her on the forehead, . . .

(File: St. George and St. Michael.txt, my emphasis)

With gendered pronouns, the overall percentage relation in the GMDFC (the first column in Fig. 2) is as follows: HE/HIS/HIM/HIMSELF—62% (161,181 hits) vs. SHE/HER/HERSELF—38% (97,889 hits). These figures nicely corroborate the mean percentages of all nouns mentioned so far, which are roughly 61% male—39% female.

The picture changes quite drastically, however, once we focus on words with less "everyday" connotations: LADY and PRINCESS are significantly more frequent in the GMDFC than their counterparts LORD and PRINCE (again, cf. Fig. 2). Taking into account that a number of the instances of LORD will be references to God, the relative abundance of *ladies* is all the more conspicuous.[7] PRINCESS is a frequent word simply because certain texts are about princesses: thus, roughly 42% of the instances of PRINCESS occur in the *Princess* books, 54% in other fairy tales and fantasies, and only 3% in the realistic narratives. We might thus expect PRINCESS to crop up as a statistical "key word" in certain texts (cf. section 6). Another conspicuous finding is that GRANDMOTHER (including variants such as *grandmither*, *granny*, etc.) is much more frequent than GRANDFATHER (again, including all such variants). This gives some quantitative weight to the fact that the presence and the significance of "grandmother" figures in MacDonald's works has been an object of literary study for years (e.g. Willard 1992; Hayward 1994). The word KING, on the other hand, is much more frequent than its female counterpart QUEEN—however, contrary to QUEEN, about half of the instances of KING are to be found in MacDonald's realistic novels, most notably in the "historical" novel *St. George and St. Michael*. It appears that KING is so frequent because it is a regularly occurring item in both the fantasies and the realistic novels while QUEEN occurs mainly in a few fantasies and fairy tales, especially in *The Princess and the Goblin*.

In summary of this short investigation of gendered nouns, there indeed appears to be a slight quantitative prevalence of male characters, especially with "general" and "everyday" words such as MAN, BROTHER or FATHER, notable exceptions being the items WIFE, GRANDMOTHER, LADY and PRINCESS. In other words, while male characters are mentioned more often, female characters are prone to occur in the more "specialized" forms of princesses, ladies, wives, and grandmothers in George MacDonald's works of fiction.

4. Golden Keys to MacDonald's Style

As we have seen, working with word frequency lists based on the GMDFC allows us to make statements and draw conclusions about MacDonald's fiction which, however, could just as easily be assumed to be true either for narrative fiction, or for nineteenth-century fiction, or for religious fiction, in general. Is it even sound, one might rightfully ask, to start making claims about MacDonald's use of gendered nouns without first comparing his use to some "norm" derived from similar works? In order to produce well-founded statements about MacDonald's fiction based on the quantitative analysis of the GMDFC, we need to compare it with other text corpora.

A common procedure for comparing corpora is the statistical evaluation of a text or text collection against the background of a (usually larger) "reference corpus" (cf. Mahlberg 2007b: 223). On the basis of word frequency lists obtained from both corpora, so-called "key words" can be collected, i.e. words whose frequency in the target corpus is significantly higher than would be expected on the basis of the reference corpus. Ideally, the search for such statistical key words will yield results that help us assess the stylistic "flavor" of George MacDonald's writings or a subset thereof.

Of course, the nature of the texts included in the reference corpus will influence the output of key words (cf. Archer 2007: 249; Scott 2010b: 51)—e.g. imagine comparing MacDonald's fiction with a corpus compiled from newspaper articles or from cooking recipes, as opposed to comparing it with other works of fiction: The results will probably differ to a certain extent, although experiments have shown the differences in outcome between such procedures not to be as great as we might expect.[8]

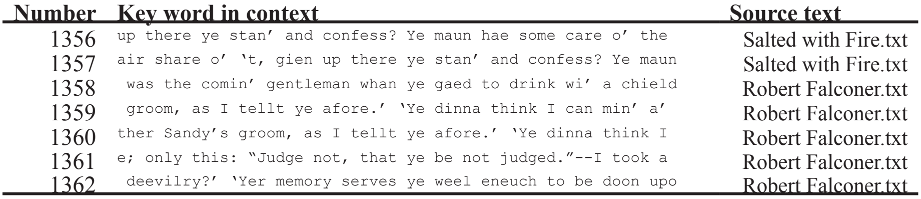
In order to obtain a sufficiently plausible reference corpus against which to compare the GMDFC, a corpus of English[9] novels from roughly the same time period (c. 1855–1900) was compiled: first, Richard D. Altick's lists of best-selling books from the Victorian period (Appendix B in Altick 1957; Altick 1969, 1986) were consulted to identify forty-two "bestsellers" published between 1855 and 1900.[10] The thirty works from this list that were available as text files at the *Project Gutenberg* website were downloaded. To these were then added thirty-three further novels published between 1855 and 1900, which were mentioned as being popular or influential in Nünning

2000, chapters 3-5. In order to maintain balance among these reference texts, a maximum of four works per author was allowed into the corpus. The files were then tagged according to the same principles as with the GMDFC (exclusion of file headers, tables of content, quotations, verse passages, etc.). The resulting corpus is roughly 9.8 million (9,818,868) words in length—about twice the size of the GMDFC—and will henceforth be referred to as the "Victorian Classics Corpus" (VCC).[11] A list of works included in this corpus is given in Appendix 2.

Table 2
Top 25 Key Words in the GMDFC (Compared with the VCC), Sorted by Keyness

Number	**Key word**	**Frequency in GMDFC**	**Frequency in VCC**	**Keyness (log likelihood)**
1	YE	7,616	2,166	8,943.32
2	O	7,263	3,831	5,421.30
3	MALCOLM	2,045	8	4,640.70
4	HAE	2,370	213	4,182.18
5	DONAL	1,689	0	3,914.26
6	GOD	6,851	4,706	3,803.91
7	WAD	1,659	7	3,759.42
8	BUT	42,119	64,664	3,112.90
9	COSMO	1,188	2	2,725.06
10	LAIRD	1,117	2	2,560.76
11	DOROTHY	1,104	0	2,558.43
12	GIBBIE	1,069	0	2,477.31
13	GIEN	1,128	17	2,449.97
14	YER	1,531	212	2,417.39
15	JIST	1,046	5	2,364.37
16	WEEL	1,368	157	2,277.42
17	NOT	42,998	70,888	2,089.95
18	KEN	1,376	268	1,928.74
19	ALEC	929	34	1,884.31
20	THE	252,158	496,990	1,865.03
21	HUGH	1,074	104	1,863.86
22	CURDIE	784	0	1,816.81
23	GANG	1,063	121	1,773.40
24	ABOOT	986	81	1,772.63
25	UPO	865	30	1,764.41

An analysis of key words in the GMDFC compared with the VCC yields the results given in Table 2 (based on the lemmatized versions of word lists elicited from the two corpora). Note that these GMDFC key words are not sorted according to their absolute frequencies in the GMDFC, but according to their “keyness,” i.e. those whose frequency in the GMDFC is highest in comparison to what would be expected on the basis of the frequencies in the reference corpus (taking into account the sizes of the corpora) are at the top of the list. The figures given in the “keyness” column are the results of automatic calculations based on the log-likelihood test for statistical significance (also called G^2 test; cf. Oakes 1998: 42): The higher the log-likelihood value, the more significant is the difference between the two frequencies. Thus, values of 3.84 or higher are statistically significant at the 5% level (i.e., there is a five percent chance that the findings are due to chance), and values of 15.13 or higher are significant at the 0.01% level. The “keyness” values in Table 2 all exceed 1,000 and are thus very highly significant. Not surprisingly (cf. Scott 2010a: 166), the words that turn up with the highest “keyness” include proper nouns that are incidental to the respective narratives (e.g., MALCOLM or DONAL). More indicative of MacDonald’s style as compared to Victorian writers in general are, equally unsurprisingly, Scots dialect items such as YE, WAD and GIEN: a search for YE in the corpus shows that most instances of YE indeed occur in Scots dialog lines (cf. Fig. 3), with only a few exceptions where the archaic second-person plural pronoun is used for stylistic purposes, such as Curdie’s motivational speech in Chapter 34 of *The Princess and Curdie*. It therefore makes sense to treat YE as a Scots dialect item.



Number	Key word in context	Source text
1356	up there ye stan’ and confess? Ye maun hae some care o’ the	Salted with Fire.txt
1357	air share o’ ‘t, gien up there ye stan’ and confess? Ye maun	Salted with Fire.txt
1358	was the comin’ gentleman whan ye gaed to drink wi’ a chield	Robert Falconer.txt
1359	groom, as I tellt ye afore.’ ‘Ye dinna think I can min’ a’	Robert Falconer.txt
1360	ther Sandy’s groom, as I tellt ye afore.’ ‘Ye dinna think I	Robert Falconer.txt
1361	e; only this: “Judge not, that ye be not judged.”--I took a	Robert Falconer.txt
1362	deevilry?’ ‘Yer memory serves ye weel eneuch to be doon upo	Robert Falconer.txt

Fig. 3. Selected concordance lines for YE in the GMDFC

Among the GMDFC key words are also function words such as BUT, YET and ITS. Even though an investigation of such “key function words” might yield interesting results from a stylistic perspective (cf. Mahlberg 2007b: 223). Most immediately relevant for the sake of a quick overview will be the third type of key words, namely nouns, verbs and adjectives, which are of the sort “that human beings would recognize” as central to the texts (Scott 2010a: 166). A top-forty key word list in which proper nouns and Scots dialect words have been ruled out is given in Table 3.[12] This list can be seen as indicative of the “aboutness” of MacDonald’s fiction, and it contains a number of items that will be of particular interest to MacDonald scholars. First of all, as conjectured in section 3 above, we do find the word PRINCESS in the list (item no. 15), along with other gendered items such as FATHER (no. 10), LORD (no. 28), GRANNIE (no. 16) and GRANDMOTHER (no. 40). These findings are in line with what the bare numbers of occurrences in the GMDFC suggested to our intuition at first sight. A surprising exception is the word LADY, which has a negative keyness in MacDonald's fiction (normalized frequencies:[13] 951.35 instances pmw in GMDFC vs. 1,096.36 instances pmw in VCC; log-likelihood value: -46.65), i.e. it is significantly *less* frequent than would be expected according to the "Victorian norm." We may conclude that although LADY occurs relatively frequently in MacDonald's fiction, it is, on average, used even more frequently by other Victorian writers. The same holds true for WIFE (370.77 hits pmw vs. 466.18 hits pmw).

Table 3
Top 40 Key Words in the GMDFC, Excluding Proper Nouns and Dialect Words (Compared with the VCC), Sorted by Keyness

Number	**Key word**	**Frequency in GMDFC**	**Frequency in GMDFC**	**Keyness (log likelihood)**
1	GOD	6,851	4,706	3,803.91
2	BUT	42,119	64,664	3,112.90
3	NOT	42,998	70,888	2,089.95
4	THE	252,158	496,990	1,865.03
5	THING	9,512	11,537	1,752.71
6	WOULD	20,904	31,619	1,660.51
7	HE	73,889	134,541	1,583.44
8	YET	6,699	7,520	1,527.01
9	LENGTH	1,840	958	1,389.88
10	FATHER	6,716	7,889	1,356.07
11	SHE	44,873	78,832	1,346.46
12	ALTHOUGH	1,642	862	1,230.47
13	GROW	3,484	3,264	1,185.83
14	LIE	6,434	7,910	1,137.83
15	PRINCESS	981	304	1,096.35
16	GRANNIE	441	0	1,021.94
17	COULD	14,733	22,986	993.87
18	ITS	6,945	9,366	900.62
19	WIND	2,460	2,075	868.03

20	NOR	3,050	3,133	858.42
21	JESUS	618	154	776.63
22	SUCH	7,697	11,059	774.62
23	LOVELY	1,194	755	733.65
24	ANSWER	5,234	7,033	687.68
25	WHAT	18,653	31,993	678.57
26	FIND	7,899	11,825	658.64
27	THEREFORE	1,703	1,502	647.75
28	LORD	3,474	4,219	637.06
29	FAR	4,109	5,294	625.53
30	BELIEVE	3,749	4,719	615.80
31	HIM	32,017	58,907	607.86
32	LEAST	2,572	2,847	607.10
33	HOWEVER	2,972	3,480	605.05
34	VANISH	850	473	601.00
35	ABLE	1,762	1,649	600.79
36	IT	62,537	121,162	598.75
37	MOON	1,001	692	550.79
38	THAN	10,298	16,880	516.66
39	MOMENT	4,997	7,144	513.54
40	GRANDMOTHER	484	164	512.00

Some other nouns and verbs with a high keyness in MacDonald's works as compared to Victorian fiction in general are well worth commenting on: the items GOD (no. 1) and JESUS (no. 21) are accountable for by the Christian faith that is often the topic of dialogs in MacDonald's fiction. An analysis of the places in which GOD occurs (by means of the "dispersion plot" function of the *WordSmith Tools* concordancer) shows that about 96% of the hits are from indisputably "realistic" novels, while occurrences of the word GOD in MacDonald's fairy tales and fantasy works are either nonexistent or negligible, *Lilith* perhaps being somewhat of an exception, with 23 occurrences of GOD (0.25 hits ptw [per thousand words], which is still at the lower end of the scale). Unsurprisingly, the word GOD occurs most frequently in *The Elect Lady* (238 hits, 4.01 ptw) and the *Wingfold* trilogy—*Paul Faber, Surgeon* (559 hits, 3.35 ptw), *There and Back* (496 hits, 2.92 ptw), and *Thomas Wingfold, Curate* (432 hits, 2.66 ptw)—followed by other realistic novels that deal explicitly with issues of faith. The 618 occurrences of JESUS are completely restricted to the realistic tales, with

"The Gifts of the Child Christ" (1.2 ptw), *Thomas Wingfold, Curate* (0.7 ptw), *The Seaboard Parish* (0.55 ptw) and *The Elect Lady* (0.51 ptw) at the top of the list. At this stage one must be careful not to draw the wrong conclusions, and keep in mind that searches for key words can only highlight the "aboutness" of a text based on its formal characteristics—e.g., one could argue that *At the Back of the North Wind* is very much "about" God and "about" death, even though a search in this book yields only seven hits for GOD, and one single hit for DEATH, which occurs in the interjected story of "Little Daylight." We may point out that one of the defining characteristics of a book like *At the Back of the North Wind* is that it deals with God and with death without "formally" mentioning them.

Other key words that are connected with the issue of faith include LORD (no. 28; less that 2% of the hits are from MacDonald's fairy tales and fantasy works, which means that their dispersion is similar to that of GOD) and BELIEVE (no. 30; the dispersion of this word is slightly more balanced over the corpus). The occurrence of FATHER (no. 10) as a key word might point into a similar direction, although the word will be expected to refer to human characters within the tales in many cases. Indeed, the relative frequencies of FATHER are highest in *Ranald Bannerman's Boyhood* (4.95 hits ptw), "Port in a Storm" (3.86 ptw), *Heather and Snow* (3.2 ptw), *Salted with Fire* (3.1 ptw) and *The Vicar's Daughter* (2.9 ptw)—works in which (human) father figures play important roles. The key word WIND (no. 19) is slightly problematic because it includes instances of both the noun *wind* and (although much fewer) the verb *to wind*; however, its occurrence in this list is definitely noteworthy. *North Wind* is used as a name in *At the Back of the North Wind*, but the cluster NORTH WIND occurs only 87 times, which actually contributes only little to the total number (2,460) of occurrences of WIND. The form *wind* occurs in 46 of the 52 texts, which means that it is definitely a key word of the entire corpus. Similarly, MOON (no. 37) is a conspicuous key word. A combined search for *moon* and *moons* shows that the word crops up in almost all text files (50 of 52), and most prominently in *At the Back of the North Wind* (80 hits),[14] followed by *Lilith* (49 hits). The key word GROW (no. 13) seems to illustrate the fact that MacDonald was highly interested in the topic of spiritual growth, as several scholars have already pointed out (e.g. Gaarden 2005).

Some further items in the key word list, such as THING, LENGTH, LOVELY, FIND, FAR, LEAST and VANISH, are harder to account for and would need more thorough investigation. It would seem that these words are not indicative of "aboutness," but of MacDonald's individual writing style—e.g., on closer inspection it turns out that the overwhelming majority of instances of LENGTH (1,669 out of 1,833) occur in the phrase *at length*, and the keyness of LENGTH is thus explicable through MacDonald's comparatively frequent use of the phrase *at length* in virtually all of his texts.

Further down in the "cleaned up" list of key words we find many items that seem to reflect much-studied motifs in MacDonald's works, such as ASLEEP (no. 42), LOVE (45), HEART (48), DREAM (49), MOTHER (50), STAIR (55), SUN (64), DOOR (67), CHILD (75), HORSE (77), SHADOW (78), MAMMON (87), READER (88),[15] KING (90), BABY (91), EVIL (93),[16] COTTAGE (96) and CARE (99).

Table 4
Eight to Twelve Word Clusters with a Minimum Frequency of 5 in the GMDFC, Sorted by Frequency

Number	Cluster	Frequency	Texts
1	BEEN TO THE BACK OF THE NORTH WIND	11	1
2	THE PRINCE OF THE POWER OF THE AIR	11	7
3	THERE IS NOTHING COVERED THAT SHALL NOT BE REVEALED	9	6
4	COME UNTO ME ALL YE THAT LABOUR AND ARE HEAVY LADEN	7	2
5	THAT LABOUR AND ARE HEAVY LADEN AND I WILL GIVE YOU REST	7	2
6	TOOK HER BY THE HAND AND LED HER	7	6
7	A HIDING PLACE FROM THE WIND A COVERT FROM THE TEMPEST	6	5
8	FIRST SHALL BE LAST AND THE LAST FIRST	6	4
9	I WILL ARISE AND GO TO MY FATHER	6	6
10	IN THE SECRET PLACE OF THE MOST HIGH	6	6
11	DID NOT KNOW WHAT TO MAKE OF IT	5	4
12	FOR THE FIRST TIME IN HIS LIFE HE	5	5
13	HER ARMS ROUND HIS NECK AND KISSED HIM	5	5
14	I DON'T KNOW WHAT TO MAKE OF IT	5	5
15	IN THE BODY OR OUT OF THE BODY	5	5
16	IS NOT THE GOD OF THE DEAD BUT OF THE LIVING	5	4
17	IS THERE ANYTHING I CAN DO FOR YOU	5	5
18	LIVETH AND BELIEVETH IN ME SHALL NEVER DIE	5	3
19	SHADOW OF A GREAT ROCK IN A WEARY LAND	5	5
20	THAN HE HAD EVER BEEN IN HIS LIFE	5	3
21	WHITE WITH THE WHITENESS OF WHAT IS DEAD	5	3
22	WOKE IN THE MIDDLE OF THE NIGHT AND	5	4

5. Multi-word Units

Going a step further, we can take not just single words, but multi-word units (usually referred to by corpus linguists as bundles or clusters) into account. Most of the large (eight- to twelve-word) clusters that occur at least five times in George MacDonald's fiction are phrases borrowed from the Bible; one is a line from Shelley (cf. Table 4, from which partial doublets have been removed). Others are more commonplace expressions such as FOR THE FIRST TIME IN HIS LIFE HE, which occurs five times. The results yielded by a search of the VCC, however, turn out to be fairly similar; long clusters that are particular to MacDonald are BEEN TO THE BACK OF THE NORTH WIND (11 hits—this is a reoccuring expression in *At the Back of the North Wind*, as readers of this book know.), TOOK HER BY THE HAND AND LED HER (7 hits) and IN THE BODY OR OUT OF THE BODY (5 hits), but given such low overall frequencies, it is dangerous to draw conclusions.

For the elicitation of key clusters in MacDonald's fiction, a list of all five- to twelve-word clusters occurring five times or more in the GMDFC was automatically compared with a corresponding list of clusters from the VCC. The results—the clusters that are most unexpectedly frequent in the GMDFC —are given in Table 5. The list is topped by clusters related to the phrase *the back of the north wind*, which has already been commented on. Many of the less intuitively predictable clusters in the list are parts of adverbials, e.g. IN THE MIDDLE OF THE or ON THE TOP OF THE, and are indications of expressions favored by the author. Especially conspicuous among these are such as feature combinations with particular nouns, e.g. (AT) THE TOP OF THE *STAIR* (nos. 4, 16), (ON) THE TOP OF THE *WALL* (nos. 8, 39), (IN) THE MIDDLE OF THE *NIGHT* (nos. 5, 6) and IN THE *HEART* OF (nos. 9, 22—my emphasis), some of which do not appear at all in the VCC.

Table 5
Top 40 Key Clusters in the GMDFC (Compared with the VCC), Sorted by Keyness

Number	**Cluster**	**Frequency in GMDFC**	**Frequency in VCC**	**Keyness (log likelihood)**
1	BACK OF THE NORTH WIND	48	0	111.23
2	THE BACK OF THE NORTH WIND	47	0	108.91
3	IN THE MIDDLE OF THE	221	171	103.93
4	THE TOP OF THE STAIR	42	0	97.32
5	IN THE MIDDLE OF THE NIGHT	69	16	89.73
6	THE MIDDLE OF THE NIGHT	76	23	86.11
7	IF THERE BE A GOD	37	0	85.74
8	THE TOP OF THE WALL	31	0	71.83
9	IN THE HEART OF A	30	0	69.52
10	ON THE TOP OF THE	87	43	68.97
11	OF THE SON OF MAN	28	0	64.88

12	WAS ON THE POINT OF	82	41	64.33
13	IN THE KINGDOM OF HEAVEN	27	0	62.57
14	AT THE BACK OF THE NORTH WIND	24	0	55.61
15	WHEN HE CAME TO HIMSELF	23	0	53.30
16	AT THE TOP OF THE STAIR	22	0	50.98
17	OF THE KINGDOM OF HEAVEN	22	0	50.98
18	HAD NOT YET BEGUN TO	22	0	50.98
19	I DO NOT KNOW BUT	37	8	49.65
20	HE COULD NOT HELP FEELING	21	0	48.66
21	NOT A FEW OF THE	21	0	48.66
22	IN THE HEART OF THE	45	15	48.10
23	ON THE POINT OF SAYING	20	0	46.34
24	TO THE BACK OF THE NORTH WIND	20	0	46.34
25	I DINNA KEN WHAUR I	19	0	44.03
26	I BEG YOUR PARDON MY	19	0	44.03
27	ROSE AND LEFT THE ROOM	19	0	44.03
28	HE DID NOT KNOW THAT	33	8	42.02
29	A GOOD DEAL MORE THAN	18	0	41.71
30	I DO NOT QUITE UNDERSTAND	18	0	41.71
31	TOO GOOD TO BE TRUE	18	0	41.71
32	SEEMED ON THE POINT OF	18	0	41.71
33	AND WAS ON THE POINT OF	17	0	39.39
34	NOW AND THEN HE WOULD	17	0	39.39
35	AS IF HE HAD JUST	17	0	39.39
36	HAD NOT YET LEARNED TO	17	0	39.39
37	INTO THE KINGDOM OF HEAVEN	17	0	39.39
38	AND WAS ON THE POINT	17	0	39.39
39	ON THE TOP OF THE WALL	17	0	39.39
40	HAD NOT GONE FAR BEFORE	17	0	39.39

Purely quantitative findings such as these might open the door to further qualitative research into MacDonald’s motifs, e.g., one could suppose a connection to exist between the finding of (AT) THE TOP OF THE STAIR as a key phrase in MacDonald’s narrative works and his general interest in spiritual development.[17] A glance over all instances of the phrase in MacDonald’s works shows that it crops up in various different situations, in one of which, however, metaphorical use is indeed made of the phrase:

> (3) When she went to his bedside, she found him breathing softly, and thought him still asleep. But he opened his eyes, looked at her for a moment fixedly, and then said: “Dorothy, child of my heart! things may be very different from

> what we have been taught, or what we may of ourselves desire; but every difference will be the step of an ascending stair--each nearer and nearer to the divine perfection which alone can satisfy the children of a God, alone supply the poorest of their cravings."
> She stooped and kissed his hand, then hastened to get him some food.
> When she returned, he was gone up the stair of her future, leaving behind him, like a last message that all was well, the loveliest smile frozen upon a face of peace. The past had laid hold upon his body; he was free in the Eternal. Dorothy was left standing at the top of the stair of the present.
> (File: Paul Faber, Surgeon.txt)

Although in most other instances the phrase (AT) THE TOP OF THE STAIR is used in more concrete situations, its relatively high overall frequency in MacDonald's works perhaps suggests his often subconscious use of such "developmental" imagery. In this respect, the phrase HAD NOT YET LEARNED TO is equally interesting: A search in the corpus shows that the 17 occurrences are spread out over 13 different works, and that, apart from very few unspectacular collocations including words such as *think* and *read*, the expression is usually followed by predicates associated with something good and valuable: MacDonald's characters very often "had not yet learned to" *trust God*, *care . . . about books*, *look . . . to heaven*, *obey*, *respect childhood*, *love him*, *believe*, *speak the truth.* This proves that MacDonald's characters are generally depicted as developing towards moral understanding and goodness. The twenty-three occurrences of IF THERE BE A GOD, on the other hand, are less far spread over the corpus: More than half of the hits are from the *Wingfold Trilogy*, which means that the expression is strongly associated with the recurring theme of atheism in these books.

6. The Style and Vocabulary of Fairyland

Even though corpus analysis tools are said to work best on very large amounts of text, it is also possible—and perhaps most interesting from the point of view of literary criticism—to apply them to the study of smaller amounts of text, and even to single works. In this context, it is tempting to divide up MacDonald's oeuvre into a "realistic" and a "fantastic" part, although it has been argued (e.g., Robb 1989: 111 et seq.) that this is hazardous since there are no clear-cut boundaries between fantasy and realism in MacDonald's work. However, for purposes of demonstration I have taken six texts whose essentially "fantastic" or fairy-tale-like nature is out of dispute—namely *Phantastes*, *Lilith*, the two *Princess* books, "A Double Story," "Cross Purposes," "The Shadows," "The History of Photogen and Nycteris" and "The Light Princess"—and created from them a sub-corpus which I will call "GMDFC-fant." A similar, albeit larger sub-section of the corpus (from now on referred to as "GMDFC-real") was then created out of twenty-seven realistic novels. Works of a more debatable or "mixed" nature (such as *At the Back of the North Wind* or *Adela Cathcart*) will be for now left out of the equation.

Table 6
Top 25 Key Words in GMDFC-fant (Compared with GMDFC-real), Sorted by Keyness

Number	**Key word**	**Frequency in GMDFC-fant**	**Frequency in GMCFC-real**	**Keyness (log liklihood)**
1	PRINCESS	896	32	4,152.01
2	CURDIE	784	0	3,871.42
3	KING	684	509	1,839.35
4	IRENE	214	0	1,056.42
5	GOBLIN	198	16	866.50
6	AND	12,193	101,214	724.57
7	LINA	138	0	681.21
8	QUEEN	205	89	667.20
9	PALACE	149	37	556.48
10	PRINCE	171	79	546.21
11	THEY	2,319	14,618	511.29
12	LOOTIE	98	0	483.75
13	MINER	102	3	476.78
14	ROSAMOND	95	0	468.94
15	SHADOW	356	806	468.21
16	RAVEN	105	12	443.05
17	GIANT	124	39	439.63
18	LONA	87	0	429.45
19	FOREST	156	119	414.91
20	WISE	228	349	412.99
21	MARA	79	0	389.96
22	NYCTERIS	79	0	389.96
23	SHE	4,657	36,590	389.24
24	LEOPARDESS	77	0	380.09
25	PHOTOGEN	69	0	340.59

Next, key word lists were created by comparing the sub-sections to each other using the methods described above. The top results of a search for statistical key words in GMD-fant are given in Table 6. Once again, proper nouns incidental to the respective stories are thrown up as key words. Table 7 shows the same results with such proper nouns (including the lemma RAVEN, all instances of which refer to the Raven/Adam character in *Lilith*) left out. At the top of this list we indeed find the item PRINCESS, whose high absolute frequency in certain texts was already noted above. In fact, much of MacDonald's "fantasy vocabulary" is very congruent with what one feels these texts to be about: fantasy and fairy tales are traditionally the domain of princesses, kings, goblins, and giants. More interesting are the function words in the list: *and*, *they*, *she*, *her* and even *the* turn up among the top key words in MacDonald's fantastic fiction. *And* could be explained as a marker of a relatively paratactic style (i.e., one in which many main clauses are linked), which in turn could be due to the fact that a number of texts in this sub-section were written especially for children (e.g., "A Double Story" or the *Princess* books) and thus prefer an easy syntax. The occurrence of the gendered pronouns *she* and *her* in this list fits in well with what has been said above, namely that female characters are featured most prominently as grandmothers, princesses, and ladies, all of which, one inclines to think, are likely to occur in fantasy and fairy tales. In fact, in the fantasy works the feminine pronouns SHE, HER, HERSELF are more frequent in total than their male counterparts (9,831 vs. 8,069 tokens or 55% vs. 45%), which verifies that female characters are featured more prominently than male characters are in MacDonald's works of fantasy.

Table 7
Top 25 Key Words in GMDFC-fant (Compared with GMDFC-real), Excluding Proper Nouns, Sorted by Keyness

Number	Key word	Frequency in GMDFC-fant	Frequency in GMCFC-real	Keyness (log likelihood)
1	PRINCESS	896	32	4,152.01
2	KING	684	509	1,839.35
3	GOBLIN	198	16	866.50
4	AND	12,193	101,214	724.57
5	QUEEN	205	89	667.20
6	PALACE	149	37	556.48
7	PRINCE	171	79	546.21
8	THEY	2,319	14,618	511.29

9	MINER	102	3	476.78
10	SHADOW	356	806	468.21
11	GIANT	124	39	439.63
12	FOREST	156	119	414.91
13	WISE	228	349	412.99
14	SHE	4,657	36,590	389.24
15	LEOPARDESS	77	0	380.09
16	FAIRY	126	93	339.83
17	THE	21,944	208,889	324.16
18	TREE	275	739	303.13
19	MOON	244	591	300.27
20	MAJESTY	115	112	272.85
21	LAMP	143	208	268.24
22	CREATURE	288	899	265.50
23	HER	4852	40,727	260.70
24	MOUNTAIN	173	337	260.34
25	RUN	398	1,589	255.88

A quick reversal of roles nicely corroborates these findings: an analysis of the key words of GMDFC-real (now using GMDFC-fant as a reference corpus) shows that, apart from the expected Scots dialect terms, items with a high keyness in the realistic novels are, on the one hand, words having to do with theology and faith (in order of their keyness: GOD, LORD, FATHER, CHURCH, JESUS, FAITH, SIN, CHRIST, SUNDAY, CURATE), and on the other, "male" nouns and pronouns (in order of their keyness: HE, HIS, MR, HIM, MAN, LORD, SIR, FATHER, LAIRD, HIMSELF, SON, UNCLE; the first "female" items in a "realistic fiction key words" list are MRS, MISTRESS and GRANNIE, ranking in the keyness vicinity of the male items LORD, SON and UNCLE, respectively). Thus, the prevalence of male characters, and of explicit references to the Christian faith, in MacDonald's realistic fiction, are facts attested to through various corpus-linguistic means.

To return to the "fantasy key words" list, the fact that the lemma THEY (no. 8 in Table 7) turns up as a key word may not have been foreseen through qualitative analysis. We must take a closer look at this finding: In the realistic novels, the word occurs between 1 and 5 times ptw (per thousand words), whereas in GMDFC-fant, it occurs more than 5 times ptw in half the texts, and up to 8.4 times ptw—its average number of occurrences being highest in both *Princess* books. The frequent occurrence of THEY in these books is thus due to the protagonists (Irene and Curdie) appearing and acting together most of the time. The appearance of the definite article *the* (no. 17) as another "fantasy key word" (on average, 60 to 75 occurrences ptw in GMDFC-fant, as opposed to only 44 to 62 occurrences ptw in GMDFC-real) is harder to account for intuitively. It is probably a stylistic feature of children's literature or of fantastic literature in general. This suspicion is corroborated by the fact that *the* does not come up as a key word if GMDFC-fant is compared with a reference corpus comprised solely of Victorian children's books: *the* is indeed a key word in (Victorian) children's literature in general.

7. A Glance at Semantic Tagging

Further corpus-linguistic tools that remain to be explored include part-of-speech tagging (also called grammatical tagging) and semantic tagging, which are possible through the use of more advanced corpus analysis tools such as *Wmatrix* (Rayson 2009; cf. McIntyre and Walker 2010). Many pages could be filled with the results of such analyses of MacDonald's fiction, but for our purposes, a short introduction to *Wmatrix* and a glance at some first results will suffice.

```
0000008 010  EX      There                    Z5
0000008 020  VBDZ    was                      A3+ Z5
0000008 030  AT1     a                        Z5
0000008 040  JJ      certain                  A4.2+ A7+
0000008 050  NN1     country                  G1.1c W3
F4/M7 K2
0000008 060  RRQ     where                    M6
0000008 070  NN2     things                   O2 X4.1 A7-
S2mf L2mf
0000008 080  VMK     used
T1.1.1[i1.2.1 A6.2+[i1.2.1
0000008 090  TO      to
T1.1.1[i1.2.2 A6.2+[i1.2.2 Z5
0000008 100  VVI     go                       M1 A2.1+
A1.1.1 A9- A1.8+ […]
0000008 110  RG      rather                   A13.5
0000008 120  RR      oddly                    A6.2-
0000008 121  .       .
```

Fig. 4. An example of part-of-speech- and semantic-tagged text from the GMDFC (created from file: A Double Story.txt).

Fig. 4 is an example of text that has been processed through the *Wmatrix* tool; the tool has separated and numbered each word and assigned different kinds of tag codes to each word; these codes are based on a tag set created for use in corpus linguistics at the UCREL research center at Lancaster University (UCREL 1993-2010). Grammatical tags are given in the third column, e.g., EX stands for "existential *there*,"VBDZ is the code for *was*, AT1 denotes a singular article, etc.; the fourth column gives the words in question; the right column contains semantic tags associated with the words —e.g., A3+ encodes the meaning "existing," A4.2+ is "detailed," A7+ means "likely," Z5 puts words into a "grammatical bin," etc. Note that more than one semantic tag can be assigned to a word, which makes the process of semantic tagging comparatively precise—the word *things* has received five different semantic tags in our example.

Thus, using *Wmatrix* means that we can now not only investigate frequencies and distributions of word forms and lemmas, but also of grammatical word classes and semantic domains. Semantic tagging is especially useful in corpus stylistics, since it can make recurrent themes appear in frequency lists even if they do not frequently "surface" on the formal level (cf. Archer 2007: 251). If, for example, we are faced with a text that any reader would feel is "about birds," but in which the actual word *bird* is avoided while words like *wing*, *feather*, *talon* and *beak* abound, "birds" will probably still crop up as a "key semantic domain" even if the word *bird* will not be among the "regular" key words. We should therefore expect the analysis of semantic tags to throw up "semantic domains" which are not necessarily congruent with the top "key words."

Due to technical limitations, *Wmatrix* at present can only handle text amounts below a million words, which is why in the following we will not analyze the entire GMDFC, but only sub-sections. Table 8 shows a list of "key semantic domains" elicited from the "fantastic" section (GMDFC-fant), compared against a subset of the VCC (namely about 30,000 words randomly selected from 30 VCC files). In this table we see at once the advantages and some disadvantages of using automatic semantic tagging.

Table 8
Top 10 Key Semantic Domains in GMDFC-fant (Compared with 30,000 Words from VCC), Sorted by Keyness

Number	**Key semantic domain**	**Frequency in GMDFC-fant**	**Frequency in VCC sample**	**Keyness (log likelihood)**
1	living creatures: animals, birds, etc.	3,265	88	75.24
2	moving, coming and going	11,195	486	60.38
3	geographical terms	2,314	60	57.13
4	in power	7,745	131	47.02
5	plants	1,783	46	44.40
6	the universe	1,071	25	31.28
7	pronouns	75,541	4,230	31.03
8	size: big	1,141	28	30.86
9	sensible	450	4	30.38
10	quantities: many/much	1,746	57	26.53

"Living creatures, animals, birds" tops the list because it includes the lemma RAVEN, which is used as the name of a character in *Lilith*. Apart from that, the statistical "key semantic domains" in MacDonald's fantasy tales are very much what a qualitative analysis might also elicit: the characters are constantly "moving" through "geographical" realms (the latter semantic domain is assigned to words like RIVER, FOREST, MOUNTAIN, STREAM and HILL). The semantic domain labeled "in power" indicates the prevalence of KINGs, QUEENs and PRINCESSes in these works. "Size: big" includes the lemma GROW, "quantities: many/much" includes the phrase AT LENGTH (both discussed above), and even MacDonald's famous "wise women" make their appearance in the semantic tag "sensible." The keyness of the domains "living creatures," "geographical terms," "plants" and "the universe" (the latter represented mostly by the word MOON) is due to the fact that MacDonald's fantasy deals very much with "nature" compared to the Victorian norm.

The use of different reference corpora demonstrates that the choice of the reference corpus is also important for the elicitation of key semantic domains: if, for example, GMDFC-fant is compared to the most plausible of the default options in *Wmatrix*, namely the "imaginative" subset taken from the British National Corpus (BNC) Sampler, which consists of roughly 223,000 words from works of imaginative fiction published between 1960 and 1974, the results differ considerably: the top semantic key domains thrown up in a comparison with 20th-century fiction are "light" and "darkness." Apparently MacDonald used words like LIGHT, SHINE, RAY, GLEAM, GLIMMER, MOONLIGHT and SUNLIGHT, as well as DARK and DARKNESS considerably more often than 20th-century writers did. However, these key domains disappear completely from the list when MacDonald is compared to his contemporaries. A direct comparison of the VCC sample with the sample of 20th-century fiction again yields the key domains "light" and "darkness" among the top six positions,[18] thus proving that this preoccupation with light and "visuality" is not particular to MacDonald, but to Victorian writers in general, and that this is the reason why "light" and "darkness" do not come up as key semantic domains when MacDonald is compared to the more plausible VCC sample.[19]

Table 9

Top 12 Key Semantic Domains in Six Selected Works (Compared with 30,000 Words from VCC), Sorted by Keyness

Phantastes	*Lilith*	*At the Back of the North Wind*
plants	living creatures: animals, birds, etc.	substances and materials: solid
geographical terms	pronouns	weather
colour and colour patterns	moving, coming and going	(discourse bin)
music and related activities	anatomy and physiology	negative
light	geographical terms	pronouns
quantities: many, much	the universe	existing
moving, coming and going	plants	degree: boosters
the universe	size: big	exclusivizers/particularizers
seem	negative	likely
location and direction	quantities: many/much	the universe
entire; maximum	colour and colour patterns	plants
substances and materials: solid	fear/shock	quantities: many/much
The Princess and the Goblin	*The Princess and Curdie*	*Sir Gibbie*
in power	in power	(unmatched)
(unmatched)	living creatures: animals, birds, etc.	quantities: many/much
industry	parts of buildings	geographical terms
degree: boosters	objects generally	objects generally
quantities: many/much	industry	the media: books
geographical terms	moving, coming and going	drinks and alcohol
objects generally	no power	food
negative	food	weather
sensory: sound	quantities: many/much	lack of food

degree: approximators	geographical terms	degree: approximators
strong obligation or necessity	drinks and alcohol	getting and possession
fear/shock	(unmatched)	linguistic actions, states and processes; communication

Table 9 summarizes the top-twelve “key semantic domains” in some selected individual works by George MacDonald; the table nicely displays some thematic similarities and differences between these works. Thus, the adult fantasy romances *Phantastes* and *Lilith* share “plants,” “moving, coming and going,” “colour and colour patterns,” “the universe” and “geographical terms” at prominent positions; the latter is also shared with the two *Princess* books, which in turn mutually share “in power” (through words like *princess*) and “industry” (through words like *miner*). *Sir Gibbie*, being a “realistic” work, shows a completely different character than the others in its top key domains. In *At the Back of the North Wind*, the top two key semantic domains (“substances and materials” and “weather”) mainly have to do with the main characters’ names, Diamond and North Wind—which highlights a minor weakness of the automatic allocation of semantic tags. In the “discourse bin” we find many elements of the realistic dialogues, and the keyness of the “universe” domain is mainly due to the word *moon* (cf. section 4 above). Note that the key semantic domains of *At the Back of the North Wind*, still do not contain the items “God/divinity” (although North Wind can be said to represent the divine) or even “death” (although many would claim that *At the Back of the North Wind* is a book “about death”), which were conjectured about in section 4. This is due to the fact that the semantic key domains are elicited based solely on the semantic domains of the individual words used in the novel, and it demonstrates that digital text analysis tools can go very far, but still they cannot take over the (human reader’s) task of interpreting or “reading between the lines” of a text. Table 10 illustrates this inability of corpus-stylistic tools to represent the whole capacity of human interpretation: while corpus-based studies succeed in describing level I (the linguistic “surface level”) and tools like *Wmatrix* even reach level II (the “below-surface level”) through the assignment of semantic tags, which can already be seen as a kind of automatic text interpretation, there are no digital tools that can reach the deepest level of interpretation—the themes, topics and motifs often written “between the lines,” that human beings are so successful in finding (in the case of *At the Back of the North Wind*, the fact that the book deals with death would be on level III). In spite of this, and although the *Wmatrix*-based results presented here are very selective, I hope it has become clear that working with semantic tags is a useful tool for the "characterization" of MacDonald's works in regard to their content, even if not on the deepest level of interpretation.

Table 10
Three Levels of Textual Analysis of Literary Works, and Corpus-stylistic Tools Helpful on These Levels

Levels of textual analysis of literary works	**Corpus-stylistic tools for analysis**
I. Linguistic 'surface' level	Untagged, tagged or parsed text corpora; searches for high-frequency words, key words, clusters, key clusters, parts of speech, grammatical structures, etc.
II. Linguistic-interpretative 'below-surface' level	E.g. semantically tagged corpora (Wmatrix); searches for semantic domains, key semantic domains, etc.
III. Interpretative level	???

9. Conclusion

This paper has made the first attempt to assess George MacDonald's works of fiction from a corpus-stylistic perspective, using the most basic functions of linguistic and stylistic corpus-analysis software. In conclusion, it is safe to say that although much of what crops up among the results of such a computer-assisted empirical research is either irrelevant, redundant, or all too obvious, there are also fascinating findings. Any scholar versed in MacDonald's works of literature will be able to expand upon what is to be found in the word frequency lists and key word lists presented above. Especially the search for "key clusters" in the George MacDonald Fiction Corpus against the backdrop of the Victorian Classics Corpus has elicited interesting and sometimes surprising expressions that could indeed be taken to represent the author's "stylistic fingerprint." Of course, as already suggested above, one has to be aware that in and of themselves, such lists do not mean a lot. The use of corpus-stylistic tools can never replace, but only complement the thorough qualitative analysis and interpretation of any text (cf. the three-level model in Table 10). "In order to fully understand the lists produced by a computer tool," stylistics experts McIntyre and Walker (2010: 522) write, "we must return to the text. Quantitative analysis guides qualitative analysis, which might guide further quantitative analysis." George MacDonald's aforementioned warning that the mere gathering and scrutinization of hard data does not necessarily lead the researcher closer towards the "truth" of a matter can be read in a similar vein. In this respect, corpus stylistics as a sub-discipline needs to remain modest in its aims. In the words of Gerbig and Müller-Wood (2006: 87), applying corpus-linguistic tools to literature will not "lead to ultimate truths," but what it can hopefully do is "bring precision to otherwise often impressionistic treatments of texts." In general, it seems that the aims of corpus-based studies must indeed lie in the realms of "precision" and corroboration of old knowledge rather than in the search for fundamentally 'new' insights.[20] We have almost exclusively focused on the GMDFC as a whole as well as on "fantastic" and "realistic" sub-corpora, but corpus-stylistic tools are equally helpful and enlightening in the analysis of individual texts. The elicitation of "key semantic domains" through semantic tagging seems especially promising in this respect, provided that the target and reference corpora are carefully chosen (cf. the semantic domains "light" and "darkness," which are not peculiar to MacDonald, but rather Victorian key concepts). Other areas that remain to be explored in future papers include, for example, the grammatical analysis through part-of-speech tagging, which we have only mentioned briefly.

To end on a somewhat lighter note—a non-academic, yet interesting web-based application of statistical text analysis software is *Wordle* (Feinberg 2009), a website which can be used to generate visually appealing "word clouds" based on the respective frequencies of word forms in a text. The automatically generated images are comprised of the source text's most frequent word forms, which are given different sizes to reflect their relative frequencies. Fig. 5 is a "word cloud" created from the text of *The Princess and the Goblin*, with "common English words" (Feinberg 2009 is not too explicit about what exactly this means) having been automatically removed. Among other things, the image demonstrates visually the fact that, in terms of "aboutness," *The Princess and Curdie* would actually have been an apt title for this book!

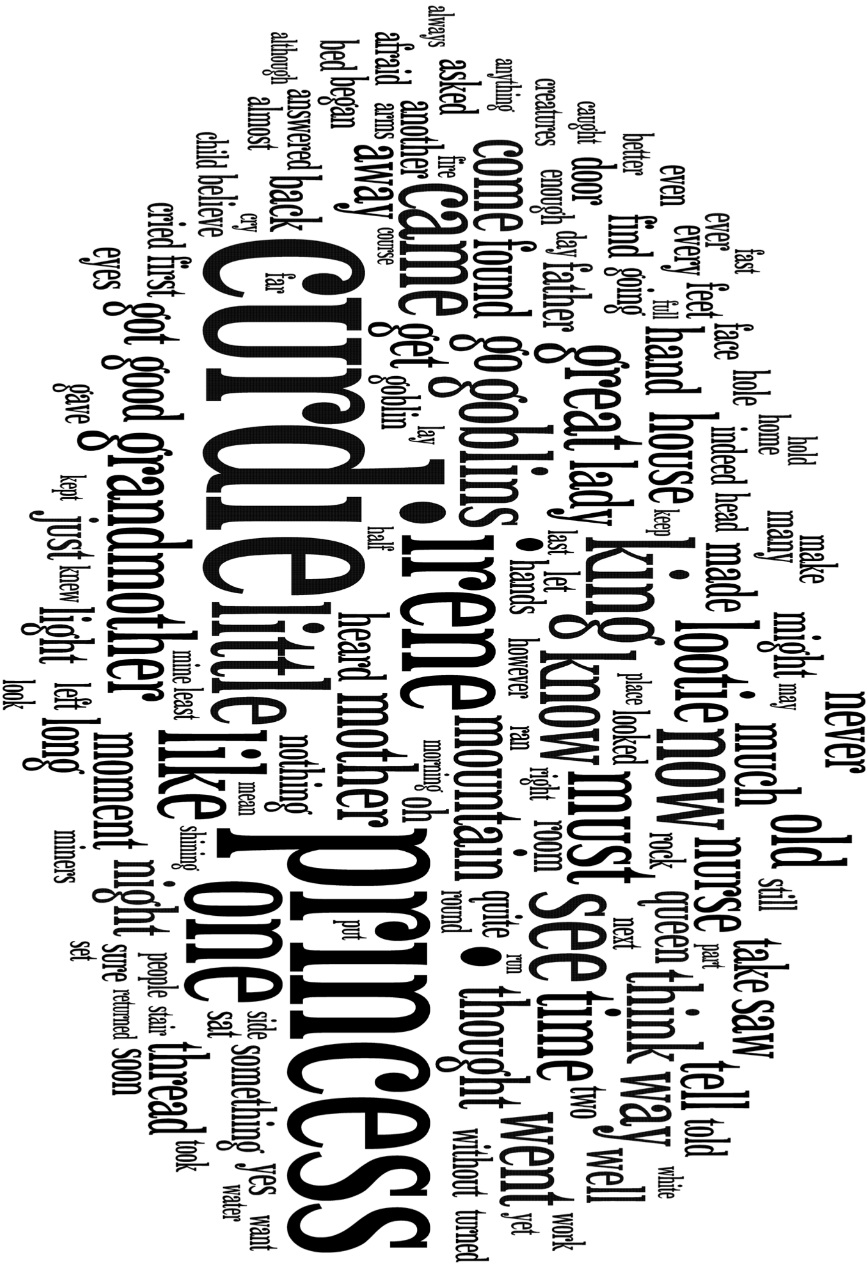


Fig. 5. A *Wordle* word cloud created from the 150 most frequent lexical items in *The Princess and the Goblin,* in roughly alphabetical order

Appendix 1. Files in the George MacDonald Fiction Corpus

A Double Story.txt; A Rough Shaking.txt; Adela Cathcart 1.txt; Adela Cathcart 2.txt; Adela Cathcart 3.txt; Alec Forbes of Howglen.txt; Annals of a Quiet Neighborhood.txt; At the Back of the North Wind.txt; Cross Purposes and The Shadows.txt; David Elginbrod.txt; Donal Grant.txt; Far above Rubies.txt; Gutta-Percha Willie.txt; Heather and Snow.txt; Home Again, a Tale.txt; Lilith.txt; Malcolm.txt; Mary Marston.txt; Paul Faber, Surgeon.txt; Phantastes.txt; Ranald Bannerman's Boyhood.txt; Robert Falconer.txt; Salted With Fire.txt; Sir Gibbie.txt; St. George and St. Michael.txt; Stephen Archer 1 Stephen Archer.txt; Stephen Archer 2 The Gifts of the Child Christ.txt; Stephen Archer 3 Photogen.txt; Stephen Archer 4 The Butcher's Bills.txt; Stephen Archer 5 Port in a Storm.txt; Stephen Archer 6 If I Had a Father.txt; The Elect Lady.txt; The Flight of the Shadow.txt; The Light Princess.txt; The Marquis of Lossie.txt; The Portent and other stories 1 The Portent.txt; The Portent and other stories 2 The Cruel Painter.txt; The Portent and other stories 3 The Castle.txt; The Portent and other stories 4 The Wow o'Rivven.txt; The Portent and other stories 5 The Broken Swords.txt; The Portent and other stories 6 The Gray Wolf.txt; The Portent and other stories 7 Uncle Cornelius His Story.txt; The Princess and Curdie.txt; The Princess and the Goblin.txt; The Seaboard Parish.txt; The Vicar's Daughter.txt; There and Back.txt; Thomas Wingfold, Curate.txt; Warlock O'Glenwarlock.txt; Weighed and Wanting.txt; What's Mine's Mine.txt; Wilfrid Cumbermede.txt

Appendix 2.
Works included in the Victorian Classics Corpus, sorted chronologically

Year of publication:	**Author:**	**Title:**
1855	Charles Kingsley	*Westward Ho!*
1855	Charles Dickens	*Little Dorrit*
1855	Anthony Trollope	*The Warden*
1856	Charles Reade	*It Is Never Too Late to Mend*
1856	John Henry Newman	*Callista*
1857	Thomas Hughes	*Tom Brown's School Days*
1857	Charles Kingsley	*Two Years Ago*
1857-59	William Makepiece Thackeray	*The Virginians*
1859	Frederic William Farrar	*Julian Home*
1859	George Eliot	*Adam Bede*
1859	George Meredith	*The Ordeal of Richard Feverel*
1859-60	Wilkie Collins	*The Woman in White*

1860	George Eliot	*The Mill on the Floss*
1860-1	Charles Dickens	*Great Expectations*
1861	Mrs. Henry Wood	*East Lynne*
1861	Charles Reade	*The Cloister and the Hearth*
1862	Mrs. Henry Wood	*The Channings*
1862	Mrs. Henry Wood	*Mrs. Halliburton's Troubles*
1862	Mary Elizabeth Braddon	*Lady Audley's Seret*
1862-63	Wilkie Collins	*No Name*
1863	Charles Kingsley	*The Water-Babies*
1863	Margaret Oliphant	*The Rector*
1863	Margaret Oliphant	*The Doctor's Family*
1864	Charles Dickens	*Our Mutual Friend*
1865	Lewis Carroll	*Alice's Adventures in Wonderland*
1868	Wilkie Collins	*The Moonstone*
1869	R.D. Blackmore	*Lorna Doone*
1870	Charles Dickens	*The Mystery of Edwin Drood*
1870	Benjamin Disraeli	*Lothair*
1871	Edward Bulwer Lytton	*The Coming Race*
1871-72	George Eliot	*Middlemarch*
1872	Samuel Butler	*Erewhon*
1874	Thomas Hardy	*Far from the Madding Crowd*
1875	Anthony Trollope	*The Way We Live Now*
1876	George Eliot	*Daniel Deronda*
1879	George Meredith	*The Egoist*
1881-83	Robert Louis Stevenson	*Treasure Island*
1882	Margaret Oliphant	*A Little Pilgrim*
1884	George Meredith	*Diana of the Crossways*
1885	H. Rider Haggard	*King Solomon's Mines*
1886	Robert Louis Stevenson	*Dr. Jeckyll and Mr. Hyde*
1886	Marie Corelli	*A Romance of Two Worlds*
1887	H. Rider Haggard	*She*
1887	H. Rider Haggard	*Allan Quartermain*
1887	Fergus Hume	*The Mystery of a Hansom Cab*
1888	Mrs. Humphrey Ward	*Robert Elsmere*
1888	H. Rider Haggard	*Maiwa's Revenge*
1890-91	William Morris	*News from Nowhere*
1891	James M. Barrie	*The Little Minister*
1891	George du Maurier	*Peter Ibbetson*
1891	Thomas Hardy	*Tess of the d'Urbervilles*
1892	Mrs. Humphrey Ward	*The History of David Grieve*

1894	Ian Maclaren	*Beside the Bonnie Brier Bush*
1895	Thomas Hardy	*Jude the Obscure*
1895	H.G. Wells	*The Time Machine*
1896	James M. Barrie	*Sentimental Tommie*
1896	James M. Barrie	*Margaret Ogilvy*
1897	Hall Caine	*The Christian*
1897	Bram Stoker	*Dracula*
1897-98	H.G. Wells	*The War of the Worlds*
1898	Mrs. Humphrey Ward	*Helbeck of Bannisdale*
1899	Joseph Conrad	*Heart of Darkness*
1900	Marie Corelli	*The Master Christian*

Endnotes

[1] "First generation" public-domain e-books were often made from scans that had been poorly formatted and proofread. While I was writing this article, author and book restoration specialist Charles Seper told me that he was in the process of creating thoroughly proofread e-book versions of *Phantastes*, *Lilith* and some other works, which will be made available for use on Kindle, Nook, etc. in the near future.

[2] Of course, it would be a nearly impossible enterprise to try to exclude (marked or unmarked) Bible quotations from a corpus of George MacDonald's works. One could indeed argue that it is an important part of the style of the novels that their characters and narrators often speak "in Biblical terms."

[3] E.g., the tale "The Shadows" is both featured in *Adela Cathcart* and in a fairy tale collection—exclusion tags were added to make this text appear only once in the analysis.

[4] Lemmatization is also known as "stemming" (cf. Feinberg 2009). In this case it was conducted automatically with the help of Someya's (1998) lemma list. The GMDFC contains 32,605 lemmas as opposed to 42,845 word forms. Of course, one needs to be aware that the process of lemmatization, if done automatically, cannot be expected to be completely accurate. Ideally, e.g., one would have to look at all 5,847 instances of the word form *thought* in the corpus (which are here automatically counted as being part of the lemma THINK) and in each case decide whether it really is a form of THINK or rather an instance of the noun THOUGHT.

[5] Cf. Stubbs 2005: 11 for more about SAY and "mental verbs" being very frequent in fictional texts; LIKE is a more problematic finding, since it is probable that a large proportion of the hits are instances of the adverb, not the verb. In order to keep the two words apart, it would be necessary to "part-of-speech-tag" the corpus.

[6] The lemmas FATHER and MOTHER here include nicknames (e.g. *daddy*) and dialectal variants (e.g. *mither*).

[7] However, if we include LAIRD (1,117 hits) in the count, the balance tips towards the male side again.

8. Cf. Scott 2010b: 51: "[Key words] indentified even by an obviously absurd [reference corpus] can be plausible indicators of aboutness, which reinforces the conclusion that keyword [sic] analysis is robust. That is to say, there is a set of common [key words] identified both by a plausible and by an implausible [reference corpus]; the implausible one will also throw up some additional (and probably implausible) [key words]." Also cf. Archer 2007: 249-250.

9. Of course, it would have been possible to consider the fact that many characters in MacDonald's realistic novels speak in a Scots dialect and thus create a reference corpus of novels by novelists who grew up in Scotland, or who also included Scots dialogues in their works, etc. The choice of texts to be included in a reference corpus depends on which features of the target corpus one wishes to focus on in the analysis.

10. Bestsellers were chosen as a starting-point for the compilation of the reference corpus because popular books are relatively likely to be found in electronic formats. Altick's lists are very useful since they contain "[b]oth varieties of best-sellers, those which enjoyed immense short-term sales and those which sold steadily over a long span of time" (Altick 1957: 381).

11. The GMDFC and the VCC were not designed as comparable or "parallel" corpora; they are actually quite dissimilar in that the VCC is about twice as long as the GMDFC (reference corpora used to elicit key words are generally larger than the target corpus), and the VCC is more varied (e.g., it will have a richer vocabulary) than the GMDFC because it contains texts from thirty-four different authors.

12. Such manually-edited tables are never made without compromises; e.g., the item TURKEY was ruled out because it is used mostly as a proper name, which meant that the (relatively few) references to actual turkeys had to be neglected; GOD and JESUS were left in the list, since these are not names of characters, but mostly subjects of discussion and were therefore felt to contribute to the "aboutness" of the texts —MARY, on the other hand, was excluded because most *Marys* in the corpus refer to fictional characters and not to biblical figures.

13. When comparing corpora of different sizes, it makes sense to normalize the frequencies of findings. In the following, frequencies of items will be given "per million words" (pmw) when compared across corpora.

14. This can be taken as numerical evidence of what Catherine Persyn (2003: 78) identifies as North Wind's hidden "lunar identity"—and it goes to show that corpus-stylistic analyses have the power to uncover hidden structures in fiction.

15. The most frequent collocation involving READER is MY+READER (239 instances), indicating the habit of MacDonald's narrators to apostrophize the reader (cf. Mahlberg 2007b: 222 for more about collocations).

16. The fact that EVIL turns up as a key word and GOOD does not might come as a surprise to some, since it has long been a commonplace that MacDonald was more successful in his depiction of good characters than of bad characters (cf. Lewis 2001: xxxiii). However, one could argue that this is not surprising, since we should expect MacDonald's "good" characters or actions to be described with a greater variety of

epithets (LOVING, GENTLE, CARING, etc.) while the "bad" characters or actions are more often simply EVIL.

17. Many thanks to Magnus Huber for reminding me to look at both variants, *stair* and *stairs*. A search for *top of the stair** (with * as a wildcard) yields the following results: 57 tokens (12.66 pmw) in GMDFC vs. 30 tokens (3.06 pmw) in VCC. The phrase is thus about four times as frequent in MacDonald's works as it is in Victorian fiction in general; its keyness in a "lemmatized" key cluster list—if such a list were easily possible to create—would still rank relatively high.

18. Other "Victorian" key semantic domains (compared with 20th-century fiction) include: "Degree", "time: general," "religion and the supernatural," "alive," "failure," "unexpected," "expected," "unethical," "psychological actions, states and processes," "money: affluence" and "work and employment: professionalism."

19. The Victorians' "fascination with visuality" (Griem 2008: 245) has been an active area of study in recent years; the Victorian preoccupation with light and vision seems to have been instigated, among other things, by a scientific "shift in the study of optics" that happened in the early nineteenth century (Garrison 2008: 199); also cf. Spear 2002: 191 et seq.

20. Cf. Stubbs 2005: 6: "[E]ven if quantification only confirms what we already know, this is no bad thing. Indeed, in developing a new method, it is perhaps better not to find anything too new, but to confirm findings from many years of traditional study, since this gives confidence that the method can be relied on."

Works Cited

Altick, Richard D. *The English Common Reader: A Social History of the Mass Reading Public, 1800-1900.* Chicago: U of Chicago P, 1957. Print.

—. "Nineteenth-century English Best-sellers: A Further List." *Studies in Bibliography: Papers of the Bibliographical Society of the University of Virginia* 22 (1969): 197-206. Print.

—. "Nineteenth-century English Best-sellers: A Third List." *Studies in Bibliography: Papers of the Bibliographical Society of the University of Virginia* 39 (1986): 235-41. Print.

Archer, Dawn. "Computer-assisted Literary Stylistics: The State of the Field." Ed. Marian Lambrou and Peter Stockwell. *Contemporary Stylistics.* London: Continuum, 2007. 244-256. Print.

Feinberg, Jonathan. *Wordle.* 2009. Web. 12 July 2011. <http://www.wordle.net>.

Gaarden, Bonnie. "*Die Aufhebung* in George MacDonald." *North Wind: Journal of George MacDonald Studies* 24 (2005): 41-50. Web.

Garrison, Laurie. "Interdisciplinarity in Victorian Studies." *The Victorian Literature Handbook.* Ed. Alexandra Warwick and Martin Willis. London: Continuum, 2008. 190-203. Print.

Gerbig, Andrea and Anja Müller-Wood. "Introduction: Conjoining Linguistics and Literature." *College Literature* 33.2 (2006): 85-90. Print.

Griem, Julika. "Visuality and Its Discontents: On Some Uses of Invisibility in Edgar Allan Poe, George Eliot, and Henry James." In: Brosch, Renate (ed.). *Victorian Visual Culture.* Ed. Renate Brosch. Heidelberg: Winter, 2008. 245-265. Print.

Hayward, Deirdre. "The Mystical Sophia: More on the Great Grandmother in the Princess Books." *North Wind: Journal of the George MacDonald Society* 13 (1994): 29-33. Print.

Lewis, C. S. *The Discarded Image: An Introduction to Medieval and Renaissance Literature.* Cambridge: Cambridge UP, 1964. Print.

—. *George MacDonald: An Anthology.* San Francisco: HarperCollins, 2001 (1946). Print.

McIntyre, Dan, and Brian Walker. "How can corpora be used to explore the language of poetry and drama?" *The Routledge Handbook of Corpus Linguistics.* In: O'Keefe, Anne and Michael McCarthy (eds.). London/New York: Routledge, 2010. 516-530. Print.

MacDonald, George. *Unspoken Sermons: Series I, II, III in One Volume.* Charleston: BilioBazaar, 2006 (1867-91). Print.

Mahlberg, Michaela. "A Corpus Stylistic Perspective on Dickens' *Great Expectations*." *Corpus Stylistics.* In: Lambrou, Martiba and Peter Stockwell (eds.). London: Continuum, 2007a. 19-31. Print.

—. "Corpus Stylistics: Bridging the Gap between Linguistic and Literary Studies." In: Hoey, Michael, Michaela Mahlberg, Michael Stubbs and Wolfgang Teubert (eds.). *Text, Discourse and Corpora: Theory and Analysis.* London: Continuum, 2007b. 219-246. Print.

Nünning, Vera. *Der englische Roman des 19. Jahrhunderts.* Stuttgart: Klett, 2000. Print.

Oakes, Michael P. *Statistics for Corpus Linguistics.* Edinburgh: Edinburgh UP, 1998. Print.

Persyn, Catherine. "A Person's Name and a Person's Self; or, Just *Who* Is North Wind." *North Wind: A Journal of George MacDonald Studies* 22 (2003): 60-83. Print.

Posner, Rebecca. "The Use and Abuse of Stylistic Statistics." *Archivum Linguisticum* 15 (1963): 111-139. Print.

Project Gutenberg. N.d. Web. 29 April 2011. <http://www.gutenberg.org>.

Rayson, Paul. *Wmatrix: A Web-based Corpus Processing Environment*, Computing Department, Lancaster University, 2009. Web. 12 July 2011. <http://ucrel.lancs.ac.uk/wmatrix>. Web.

Robb, David S. *God's Fiction: Symbolism and Allegory in the Works of George MacDonald.* Masterline Series 4. Eureka: Sunrise, 1989. Print.

Scott, Mike. *WordSmith Tools Version 5*. Liverpool: Lexical Analysis Software, 2008. Print.

—. *WordSmith Tools Help*. Liverpool: Lexical Analysis Software, 2010a. Print.

—. "Problems in Investiganting Keyness, or Clearing the Undergrowth and Marking Out Trails…" *Keyness in Texts.* In: Bondi, Marina and Mike Scott (eds.): Amsterdam/Philadelphia: John Benjamins, 2010b. 43-57. Print.

Sinclair, John. "The Search for Units of Meaning." *Corpus Linguistics: Critical Concepts in Linguistics.* In: Teubert, Wolfgang and Ramesh Krishnamurthy (eds.). Vol. 3. London: Routledge, 2007 (1996). 3-29. Print.

Someya, Yasumasa. *e_lemma (Ver.2 für WordSmith 4).* 1998. Web. 29 April 2011. <http://www.lexically.net/downloads/BNC_wordlists/e_lemma.txt>. Web.

Spear, Jeffrey. "The Other Arts: Victorian Visual Culture." *A Companion to the Victorian Novel.* In: Brantlinger, Patrick and William B. Thesing (eds.). Oxford: Blackwell, 2002. 189-206. Print.

Stubbs, Michael. "Conrad in the computer: Examples of Quantitative Stylistics methods." *Language and Literature* 14, 1 (2005): 5-24. Print.

UCREL Home Page, Lancaster, UK. 1993-2010. Web. 23 August 2011. <http://ucrel.lancs.ac.uk>. Web.

Willard, Nancy. "The Goddess in the Belfry: Grandmothers and Wise Women in George MacDonald's Books for Children." In: McGillis, Roderick (ed.): *For the Childlike: George MacDonald's Fantasies for Children.* Ed. Roderick McGillis. Metuchen: Scarecrow, 1992. 67-74. Print.