A Preliminary Report on the Set of Symbols Occurring in Engineering Mathematics Texts

Stephen M. Watt
Ontario Research Centre for Computer Algebra
Department of Computer Science
University of Western Ontario
London Ontario, CANADA N6A 5B7
Stephen.Watt@uwo.ca

Certain forms of mathematical expression are used more often than others in practice. We propose that a quantitative understanding of actual usage can provide information to improve the accuracy of software for the input of mathematical expressions from scanned documents or handwriting and allow more natural forms of presentation of mathematical expressions by computer algebra systems. Earlier work [1] has examined this question for the diverse set of articles from the mathematics preprint archive `arXiv.org`. That analysis showed showed the variance between mathematical areas. The present work analyzes a particular mathematical domain more deeply. We have chosen to examine second year university engineering mathematics as taught in North America as the domain. This syllabus typically includes linear algebra, complex analysis, Fourier analysis, vector calculus, and ordinary and partial differential equations. We have analyzed the set of expressions occurring in the most popular textbooks, weighted by popularity. Assuming that early training influences later usage, we take this as a model of the set of mathematical expressions used by the population of North American engineers. We present a preliminary empirical analysis of the individual symbols and of sequences of n symbols (n-grams) occurring in these expressions.

Corpus Selection The first step in our approach was to identify the most popular textbooks in the area of second year engineering mathematics. US college and university bookstore sales for spring for 2006 to fall 2006 show the most demanded texts to be Kreyszig [2] (72%), Greenberg [3] (13%), O'Neil [4] (7%), Jeffrey (5%), Harman (2%). From this we see that three titles account for more than 90% of the textbook use. We therefore built our model based on these three titles.

TEX Sources For each of the three textbooks, we obtained TEX sources for all the mathematical expressions, and then constructed MathML from the TEX. For the texts by Greenberg and O'Neil, the author and publisher (respectively) were highly cooperative and provided the TEX sources directly. The sources for the text by O'Neil corresponded to the published version in use today. The sources for the text by Greenberg had somewhat diverged from the published text but not so much as to materially affect the analysis in our opinion. For the text by Kreyszig, the publisher and author declined to provide access to the source files. To obtain the mathematical expressions of the text in electronic form, we first scanned the entire book and used the Infty system [5] to produce TEX. In most cases the TEX produced had to be edited by hand to correct errors. This was a highly labour intensive activity that spanned several months. In the end we had a TEX representation for all the mathematical expressions in all three texts.

MathML Conversion Naïve examination of TEX sources does not give the mathematical expressions of a document. This is for two reasons: The first reason is that typical TEX document markup makes use of a number of macro packages, as well as author-defined macros. These macros have to be expanded to reveal the mathematical expression. The second reason that TEX sources do not give *expressions* directly is that the TEX representation of mathematics is not grouped as required. For example, most authors would write `$a + b c$` rather than `$a + {b c}$`. We used our TEX to MathML converter [6, 7] to expand the TEX macros and properly group the expressions. We then performed our analysis on the resulting MathML. The resulting expressions treated were (for the most part) complete, well formed, and grouped appropriately. We describe the conversion process in more detail elsewhere [1, 8].

Analysis We grouped the chapters of each text into general subject categories (ODEs, PDEs, vector calculus, *etc*) and analyzed the mathematical expressions for each subject/author combination, for each author with subjects combined (weighted by author emphasis), and for each subject with authors combined (weighted by sales volume). In each case, we computed the individual symbol frequencies (normalized to total 1) and n-gram frequencies for $n = 2, 3, 4, 5$. To compute the n-grams, we converted the expressions to strings by traversing the frontier of the expression trees in writing order. The resulting strings were over the alphabet of leaf symbols extended by `<sub>`, `</sub>`, `<sup>`, `</sup>`, `<frac/>` and `<root/>`. These symbols captured transitions from the expression baseline to subscripts and superscripts as well as built up fractions and radicals. The n-grams were then tallied using sliding windows over these strings.

Results Tables 1 and 2 show extracts of the preliminary results of our analysis. Table 1 shows the frequencies of the most commonly occurring symbols in the entire corpus. These are presented with the absolute symbol count for each author and as a percentage of all symbols, weighted by author. The relative weights used were (72, 13, 7). We see that the most popular symbols were common among all the authors, although the rank of the symbols varied somewhat from author to author. The total number of mathematical symbols occurring in the texts were 368,267, 467,044 and 391,602. Table 1 also shows the most commonly occurring symbols for two representative areas. We see that the declining relative frequency is similar between the areas, with a few outlying points (such as z being very popular for complex analysis). This same pattern was observed for all subject areas. The cumulative frequency of symbols is shown in Figure 1 with one curve for each subject and one for the weighted combination. From the log plot it is possible to see that the symbol frequencies follow an approximately exponential distribution.

Table 2 shows a preliminary count of the most popular 5-grams for the three corpus authors as well as from two comparison texts. The n-grams have a qualitatively similar declining frequency pattern as the symbols, but this time in a much larger space. The total number of n-grams (for each n) was 479,388 (Kreyszig), 562,297 (Greenberg) and 477,268 (O'Neil). The total number of *different* bigrams was 5,992 (Kreyszig), 7,056 (Greenberg) and 5,442 (O'Neil). The total number of *different* 5-grams was 140,306 (Kreyszig), 146,507 (Greenberg), 126,232 (O'Neil). Figure 1 shows the cumulative frequency for all distinct n-grams occurring in the text by Kreyszig. The highest curve is for $n = 2$ and they are in order to the lowest curve for $n = 5$. We find it remarkable that even though the ranking of the particular n-grams is different for the each author, the cumulative n-gram frequency curves are almost identical from author to author.

By analyzing the population of symbols and n-grams that occur in the corpus, we are able to determine the most popular symbols and n-grams by subject. The exponential drop in number of occurrences, from the highest ranked symbols and n-grams to the lowest, means that a compact database can contain most of the frequently occurring items. Thus applications, even those for portable devices, could use these statistics to guide their recognition.

Acknowledgments We thank Michael Greenberg, Peter O'Neil, Prentice-Hall and Thomson-Nelson for the use of their materials. We also thank Robert Lopez and Maplesoft for additional materials. We thank Jeliazko Polihronov for assistance in gathering the data and Elena Smirnova for work on the n-gram analysis software. This work was supported in part by grants from the NSERC Canada, Microsoft and Maplesoft.

References

[1] C.M. So and S.M. Watt, *Determining Empirical Properties of Mathematical Expression Use*, pp. 361-375, Proc. Fourth Int'l Conf. on Mathematical Knowledge Management (MKM 2005), Springer Verlag LNCS 3863.

[2] Erwin Kreyszig, *Advanced Engineering Mathematics, 8^{th} edition*, John Wiley & Sons 1999.

[3] Michael Greenberg, *Advanced Engineering Mathematics, 2^{nd} edition*, Prentice Hall 1998.

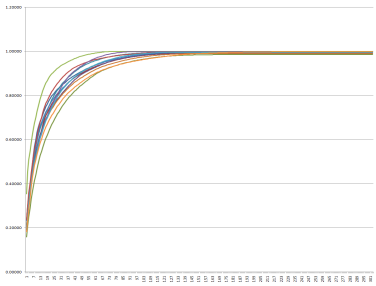
[4] Peter O'Neil, *Advanced Engineering Mathematics, 5^{th} edition*, Thomson-Nelson 2003.

[5] M.Suzuki, F.Tamari, R.Fukuda, S.Uchida, T.Kanahori, *Infty—an Integrated OCR System for Mathematical Documents*, Proceedings of ACM Symposium on Document Engineering 2003, Grenoble, 2003, pp.95-104.

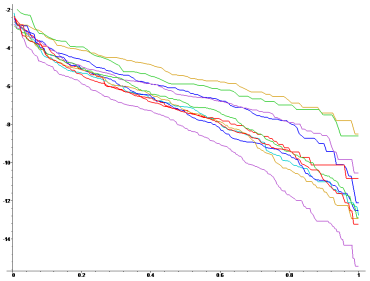
[6] ORCCA. *On-line TeX to MathML Translator.* `http://www.orcca.on.ca/MathML/texmml/textomml.html`

[7] S.M.Watt. *Implicit Mathematical Semantics in Conversion between TEX and MathML.* TUGBoat, **23**(1) 2002.

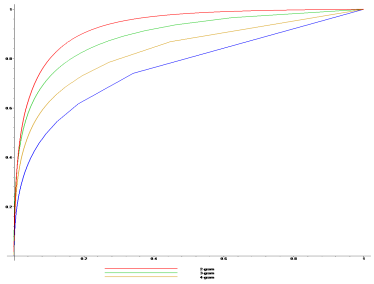
[8] E. Smirnova and S.M. Watt, *Context-Sensitive Mathematical Character Recognition.* International Conference on Frontiers in Handwriting Recognition (ICFHR 2008), (accepted).



Symbols
X: Symbol rank number
Y: Cumulative frequency



Symbols
X: Symbol rank number
Y: Log frequency



2-, 3-, 4- and 5-grams
X: $n-$gram rank number
Y: Cumulative frquencey

Figure 1: **Symbol and n-gram frequencies**

All areas combined				
Symbol	Weighted Freq. (%)	Symbol Counts Kreyszig	Greenberg	O'Neil
1	6.16415	24519	23209	20345
2	6.15918	24436	22613	21886
$=$	5.89883	22906	26202	19275
0	5.13055	20436	19623	16164
(	5.08432	18162	26262	27777
)	5.08387	18158	26257	27804
x	4.97402	18271	28243	17918
$-$	3.82436	14609	15625	17152
$+$	3.12976	11906	14648	11711
y	2.94812	11400	13191	9996
,	2.53506	9796	12571	6784
n	2.11526	8016	9681	8577
z	1.88590	7447	7238	6593
3	1.87252	7225	7603	7706
$\frac{\square}{\square}$	1.73059	6386	7715	9163
t	1.71003	5771	9800	11446
.	1.62134	6234	4510	10083
4	1.42027	5694	4119	6097
f	1.30925	4926	6522	4874

Complex Analysis	
Symbol	Weighted Freq. (%)
z	11.28007
$=$	6.19577
)	5.76133
(	5.75744
1	5.59297
2	5.21226
$-$	4.02399
0	3.88584
$+$	3.71409
i	2.95919
n	2.94910
|	2.78406
x	2.45995
f	1.98821
,	1.69837
y	1.60176
π	1.30730
C	1.18192
3	1.13346

PDEs	
Symbol	Weighted Freq. (%)
$=$	7.22187
x	7.04832
(	6.44756
)	6.43967
2	5.54914
0	4.82981
u	4.28608
1	3.35806
n	3.33931
t	3.10607
y	2.63211
$-$	2.38819
$+$	2.02753
,	2.00038
c	1.68841
r	1.67920
π	1.66333
f	1.38602
∂	1.32067

Table 1: **Most popular symbols, by weighted frequency, for entire corpus and two sample areas**

Kreyszig		**Greenberg**		**O'Neil**		**Lopez**		**MSKit**	
Freq%	**Sequence**	**Freq%**	**Sequence**	**Freq%**	**Sequence**	**Freq%**	**Sequence**	**Freq%**	**Sequence**
0.00104	(x, y)	0.00142	$\int^{\langle sub\rangle}0^{\langle /sub\rangle\langle sup\rangle}$	0.00152	(x, y)	0.00284	00000	0.00442	$lim^{\langle sub\rangle}x$
0.00095	$y^{\langle sup\rangle\prime\prime\langle /sup\rangle}$	0.00130	(x, y)	0.00149	$\int^{\langle sub\rangle}0^{\langle /sub\rangle\langle sup\rangle}$	0.00136	(x, y)	0.00406	$im^{\langle sub\rangle}x\rightarrow$
0.00082	$x^{\langle sub\rangle}1^{\langle /sub\rangle}+$	0.00077	$0^{\langle /sub\rangle\langle sup\rangle}\infty^{\langle /sup\rangle}$	0.00106	$y^{\langle sup\rangle\prime\prime\langle /sup\rangle}$	0.00120	$x^{\langle sup\rangle}2^{\langle /sup\rangle}+$	0.00320	$x^{\langle sup\rangle}2^{\langle /sup\rangle}+$
0.00081	$f(x) =$	0.00077	$x^{\langle sup\rangle}2^{\langle /sup\rangle}+$	0.00102	$0^{\langle /sub\rangle\langle sup\rangle}\infty^{\langle /sup\rangle}$	0.00104	$\int^{\langle sub\rangle}0^{\langle /sub\rangle\langle sup\rangle}$	0.00285	$dy^{\langle frac/\rangle}dx$
0.00080	$\int^{\langle sub\rangle}0^{\langle /sub\rangle\langle sup\rangle}$	0.00071	(x, t)	0.00100	$\sum^{\langle sub\rangle}n = 1$	0.00096	$f(x) =$	0.00200	$f(x) =$
0.00073	$0^{\langle /sub\rangle\langle sup\rangle}\infty^{\langle /sup\rangle}$	0.00070	$1^{\langle /sub\rangle},\ldots,$	0.00100	$^{\langle sub\rangle}n = 1^{\langle /sub\rangle}$	0.00074	$x^{\langle sup\rangle}2^{\langle /sup\rangle}-$	0.00200	$sin(x$
0.00072	$^{\langle /sub\rangle\langle sup\rangle}/^{\langle /sup\rangle} =$	0.00067	$^{\langle sub\rangle}1^{\langle /sub\rangle},\ldots$	0.00100	$n = 1^{\langle /sub\rangle\langle sup\rangle}$	0.00070	$(x, y,$	0.00190	$x^{\langle sup\rangle}2^{\langle /sup\rangle}-$
0.00072	$x^{\langle sup\rangle}2^{\langle /sup\rangle}+$	0.00062	$y(x) =$	0.00094	$1^{\langle /sub\rangle\langle sup\rangle}\infty^{\langle /sup\rangle}$	0.00066	$^{\langle sup\rangle}2^{\langle /sup\rangle}+1$	0.00188	$in(x)$
0.00071	$^{\langle sup\rangle}/^{\langle /sup\rangle}+$	0.00060	$^{\langle sub\rangle}0^{\langle /sub\rangle\langle sup\rangle}\infty$	0.00093	$= 1^{\langle /sub\rangle\langle sup\rangle}\infty$	0.00065	$,\ldots,$	0.00154	$2x^{\langle sup\rangle}2^{\langle /sup\rangle}$
0.00064	$-z^{\langle sub\rangle}0^{\langle /sub\rangle}$	0.00057	$^{\langle /sub\rangle}(x) =$	0.00090	$)sin($	0.00064	$f(x, y$	0.00141	$cos(x$
0.00060	.....	0.00056	$(0) = 0$	0.00086	$x^{\langle sup\rangle}2^{\langle /sup\rangle}+$	0.00064	$+y^{\langle sup\rangle}2^{\langle /sup\rangle}$	0.00133	$os(x)$
0.00057	$^{\langle sub\rangle}1^{\langle /sub\rangle\langle sup\rangle}/$	0.00056	$f(x, y$	0.00084	$^{\langle sup\rangle}/^{\langle /sup\rangle}+$	0.00061	x, y, z	0.00131	$x^{\langle sup\rangle}3^{\langle /sup\rangle}+$
0.00055	$z^{\langle sub\rangle}0^{\langle /sub\rangle})$	0.00052	$^{\langle sub\rangle}1^{\langle /sub\rangle}(x$	0.00084	$(-1)^{\langle sup\rangle}$	0.00060	$, y, z)$	0.00124	$^{\langle sup\rangle}2^{\langle /sup\rangle}+1$
0.00055	$y^{\langle sub\rangle}1^{\langle /sub\rangle\langle sup\rangle}$	0.00052	$1^{\langle /sub\rangle}(x)$	0.00082	$sin(n$	0.00059	$2x^{\langle sup\rangle}2^{\langle /sup\rangle}$	0.00122	$log^{\langle sub\rangle}a$
0.00054	$y(0) =$	0.00052	$1^{\langle /sub\rangle\langle sup\rangle}\infty^{\langle /sup\rangle}$	0.00076	$y^{\langle sup\rangle}/^{\langle /sup\rangle} =$	0.00058	$^{\langle sup\rangle}2^{\langle /sup\rangle}+y$	0.00122	$og^{\langle sub\rangle}a^{\langle /sub\rangle}$
0.00054	$1^{\langle /sub\rangle\langle sup\rangle}/^{\langle /sup\rangle}$	0.00051	x, y, z	0.00073	$y^{\langle sup\rangle}/^{\langle /sup\rangle}+$	0.00058	$)sin($	0.00117	$y^{\langle frac/\rangle}dx =$
0.00051	$^{\langle sub\rangle}2^{\langle /sub\rangle\langle sup\rangle}/$	0.00050	$f(x) =$	0.00072	$,\ldots,$	0.00058	$^{\langle /sup\rangle}cos($	0.00116	$^{\langle root/\rangle}2x^{\langle sup\rangle}2$
0.00051	$z - z^{\langle sub\rangle}0$	0.00049	$(x, y,$	0.00070	(x, t)	0.00057	$^{\langle /sup\rangle}sin($	0.00116	$(x^{\langle sup\rangle}2^{\langle /sup\rangle}$
0.00050	$, y^{\langle sub\rangle}2^{\langle /sub\rangle}$	0.00044	$^{\langle sub\rangle}n^{\langle /sub\rangle}(x$	0.00068	$-1)^{\langle sup\rangle}n$	0.00055	$0^{\langle /sub\rangle\langle sup\rangle}\infty^{\langle /sup\rangle}$	0.00116	$du^{\langle frac/\rangle}dx$

Table 2: **Most popular 5-grams**