An Empirical Measure 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

E-mail: watt@orcca.on.ca

Abstract

Certain forms of mathematical expression are used more often than others in practice. A quantitative understanding of actual usage can provide additional information to improve the accuracy of software for the input of mathematical expressions from scanned documents or handwriting and more natural forms of presentation of mathematical expressions by computer algebra systems. Earlier work 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. We have analyzed the set of expressions occurring in the most popular textbooks, weighted by popularity. Assuming that early training influences later mathematical usage, we take this as a model of the set of mathematical expressions used by the population of North American engineers. We present an empirical analysis of the symbols and n-grams occurring in these expressions.*

1 Introduction

This paper concerns computer analysis of mathematical documents. Unlike natural language text, dictionary-based techniques cannot be used easily to guide recognition—there is no fixed vocabulary of mathematical "words" that may appear in expressions. There is, however, a well-established tradition of common mathematical usage: some sub-expressions occur in practice more than others. We wish to use this information to guide the recognition of mathematical text.

In earlier work [1] we have reported on the analysis of some 20,000 articles from the mathematics pre-print server arXiv.org [2]. These articles were classified by area using the MSC subject classification, and we were able to observe that mathematical usage varied considerably by area. We then were able to use the information gathered to construct n-grams to improve the recognition accuracy of a pen-based mathematics interface [3, 4]. The construction of n-grams from tree-structured data used a linearization technique to traverse the tree frontier and insert sufficient geometric symbols to keep track of the expression baseline.

We wish to explore further the general approach of using statistical analysis of mathematical corpora to improve the recognition of mathematical expressions by software systems. While our interest is primarily in the area of mathematical handwriting recognition, the same models should be useful in improving the analysis of mathematics in scanned documents by systems such as Infty [5].

One of the difficulties in using the arXiv server is that certain specialized mathematical areas receive an unrepresentative number of articles by particular authors and their idiosyncrasies skew the analysis. For the current work, we have therefore taken an area of general interest that nevertheless exhibits a wide diversity of mathematical notation. We have selected the domain of engineering mathematics as taught in the second year of North American university programs as the scope for the current study.

Second year engineering mathematics is taught as a collection of applied mathematical subjects including such topics as elementary complex analysis and vector calculus. The population that uses these techniques measures in the millions of individuals, so any progress in handling documents with this mathematical content would be useful.

In addition to the use in document analysis and recognition, a statistical study of mathematical expression usage can be of interest in other areas. In particular, we would suggest that computer algebra systems could make use of information about what are the preferred forms in practice in order to present their output in the most desirable way.

Earlier work on optical character recognition for typeset mathematical documents touches upon aspects of the current paper. One study [6] considered a collection of 30 English works on pure mathematics and analyzed the scanned images for visual properties of the mathematical characters, such as whether they were touching or abnormal in shape. Another study [7] analyzed a database of 400 document images and noted that expression symbols differed from normal text, that a set of 12 two-dimensional layout structures were used and that the top 150 n-grams or so were highly representative of the subject categories. The present article does not consider at all the printed appearance of the mathematical text. Instead we take the document source text (ground truth) as given and analyze the symbol and n-gram frequences that occur.

The paper is organized as follows: Section 2 presents the problem we study in more detail. Section 3 outlines the methodology we have used to collect and analyze our data. Section 4 presents our first results. Section 5 discusses future work and presents our conclusions.

2 The Problem

We are interested in analyzing documents that use engineering mathematics as presented in the second year of the North American university education. Such documents would include engineering documents in professional practice, mathematical textbooks, student assignments and hand-written mathematics by both students and practicing professionals.

While elementary engineering mathematics includes a broad range of activity, the range of mathematical notation used is limited, at least when compared to range of notations for mathematics as a whole. We make the assumption that the notations used in practice will follow to a large those that the practitioners learned while students.

Under this assumption, we are ultimately studying the set of expressions occurring in the collection of textbooks used to teach second year engineering students. We may model the population of expressions used in practice by analyzing the individual textbooks and weighting them by their popularity.

The problem we wish to study is the statistics of the space of mathematical expressions that occur in these texts, with a suitable weighting.

Rank	Author	Reference	Demand	Adoptions
1	Kreyszig	[8]	72%	67%
2	Greenberg	[9]	13%	14%
3	O'Neil	[10]	7%	8%
4	Jeffrey	[11]	5%	5%
5	Harman	[12]	2%	3%
6	Zill	[13]	1%	1%
7	Potter	[14]	1%	1%
8	Wylie	[15]	0%	1%

(Source 353 adoptions reported in TDIS.)

Table 1. Second year engineering texts

1	First Order ODEs
2	Second-Order Linear ODEs
3	Higher Order Linear ODEs
4	Systems of ODEs Phase Plane Qualitative Methods
5	Series Solutions of ODEs Special Functions
6	Laplace Transforms
7	Linear Algebra—Matrices,Vectors,Determinants,Lin.Systems
8	Linear Algebra—Matrix Eigenvalue Problems
9	Vector Differential Calculus—Grad Div Curl
10	Vector Integral Calculus—Integral Theorems
11	Fourier Series Integrals and Transforms
12	Partial Differential Equations PDEs
13	Complex Numbers and Functions
14	Complex Integration
15	Power Series Taylor Series
16	Laurent Series Residue Integration
17	Conformal Mapping
18	Complex Analysis and Potential Theory
19	Numerics in General
20	Numeric Linear Algebra
21	Numerics for ODEs and PDEs
22	Unconstrained Optimization Linear Programming
23	Graphs Combinatorial Optimization
24	Data Analysis Probability Theory
25	Mathematical Statistics

Table 2. Kreyszig table of contents

3 Methodology

Corpus Selection The first step in our approach was to identify the most popular textbooks in the area of second year engineering mathematics. Table 1 shows the US college and university bookstore sales for spring for 2006 to fall 2006. From this we see that three titles account for about 90% of the textbook use. We therefore build our model based on these three titles. The subjects covered in these three texts are shown in Tables 2, 3 and 4.

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.

1	Introduction to Differential Equations
2	Equations of First Order
3	Linear Differential Equations of Second Order and Higher
4	Power Series Solutions
5	Laplace Transform
6	Quantitative Methods—Numerical Solution of DEs
7	Qualitative Methods—Phase Plane and Nonlinear DEs
8	Systems of Linear Algebraic Equations—Gauss Elimination
9	Vector Space
10	Matrices and Linear Equations
11	The Eigenvalue Problem
12	Extension to Complex Case
13	Differential Calculus of Functions of Several Variables
14	Vectors in 3-Space
15	Curves Surfaces and Volumes
16	Scalar and Vector Field Theory
17	Fourier Series Integral Transform
18	Diffusion Equation
19	Wave Equation
20	Laplace Equation
21	Functions of a Complex Variable
22	Conformal Mapping
23	The Complex Integral Calculus
24	Taylor Laurent Series Residue Theorem

Table 3. Greenberg table of contents

1	ODEs—First Order Differential Equations
2	ODEs—Second Order Differential Equations
3	ODEs—The Laplace Transform
4	ODEs—Series Solutions
5	ODEs—Numerical Approximation of Solutions
6	Vectors and Linear Algebra—Vectors and Vector Spaces
7	Vectors and Linear Algebra—Matrices and Systems of Linear Equations
8	Vectors and Linear Algebra—Determinants
9	Vectors and Linear Algebra—Eigenvalues Diagonalization and Special Matrices
10	Systems of Linear Differential Equations
11	Qualitative Methods and Systems of Nonlinear Differential Equations
12	Vector Analysis—Vector Differential Calculus
13	Vector Analysis—Vector Integral Calculus
14	Fourier Series
15	The Fourier Integral and Fourier Transforms
16	Fourier Analysis—Special Functions Orthogonal Expansions and Wavelets
17	PDEs—The Wave Equation
18	PDEs—The Heat Equation
19	PDEs—The Potential Equation
20	Geometry and Arithmetic of Complex Numbers
21	Complex Analysis—Complex Functions
22	Complex Analysis—Complex Integration
23	Complex Analysis—Series Representations of Functions
24	Complex Analysis—Singularities and The Residue Theorem
25	Complex Analysis—Conformal Mappings
26	Counting and Probability
27	Statistics

Table 4. O'Neil table of contents

For the text by Kreyszig it was not possible to obtain sources. To obtain the mathematical expressions of the text in electronic form, we first scanned the entire book and used the Infty system 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 the text.

MathML Conversion Naive 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 the 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 `$a + {b c}$`. While it is true that a coarsening of the TeX layout tree would correspond to a coarsening of the mathematical expression tree, it is still in general necessary to regroup the TeX representation.

We used our TeX to MathML [16] converter[17], described elsewhere [18], to resolve these difficulties, and performed our analysis on the resulting MathML expressions. The benefit of this approach was that the expressions treated were (for the most part) complete, well formed, and grouped appropriately. The difficulty with the approach was that not all the complexities of TeX were handled, and a small number of expressions were incorrectly translated. However, since we are interested in the most frequently occurring expressions, the incomplete handling of infrequently occurring expressions is not, in principle, a problem. The conversion process has been described in more detail elsewhere [1].

Analysis We grouped the chapters of each text into the general categories shown in Table 5 and analyzed the mathematical expressions for each subject/author combination, for each author with subjects combined (as given in the text), and for each subject with authors combined by weight.

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 _,, [,], <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.

4. Results

Single Symbols Table 6 shows the frequencies of the most commonly occurring symbols in the entire set of expressions. 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

- Ordinary Differential Equations
  (Kreyszig 1-6, Greenberg 1-7, O'Neil 1-5 & 10-11)
- Linear Algebra
  (Kreyszig 7-8, Greenberg 8-11 & 14, O'Neil 6-9)
- Vector Calculus
  (Kreyszig 9-10, Greenberg 16, O'Neil 12-13)
- Partial Differential Equations
  (Kreyszig 12, Greenberg 18-20, O'Neil 17-19
- Fourier Analysis
  (Kreyszig 11, Greenberg 17, O'Neil 14-16)
- Multivariable Calculus
  (Greenberg 13&15)
- Complex Analysis
  (Kreyszig 13-18, Greenberg 12&21-24, O'Neil 20-25)
- Numerical Analysis
  (Kreyszig 19-21)
- Linear Programming
  (Kreyszig 22)
- Graph Theory
  (Kreyszig 23)
- Probability and Statistics
  (Kreyszig 24-25, O'Neil 26-27)

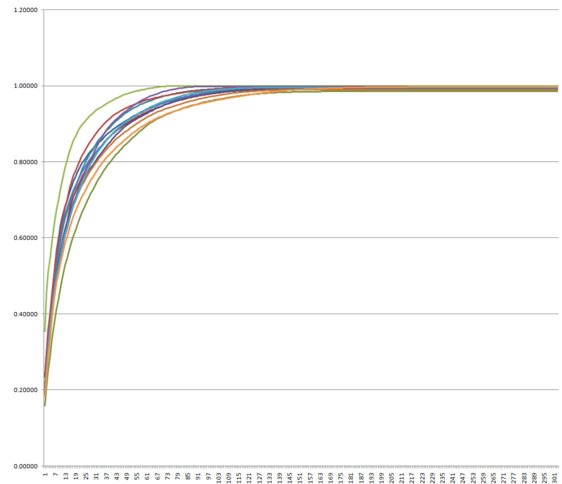
Table 5. Subject Groupings

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 and 467 044 and 391 602, respectively).

Tables 7 and 8 show the most commonly occurring symbols for the second year engineering versions of complex analysis and partial differential equations, respectively. We see that the curve of declining relative frequency of the most popular symbols 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.

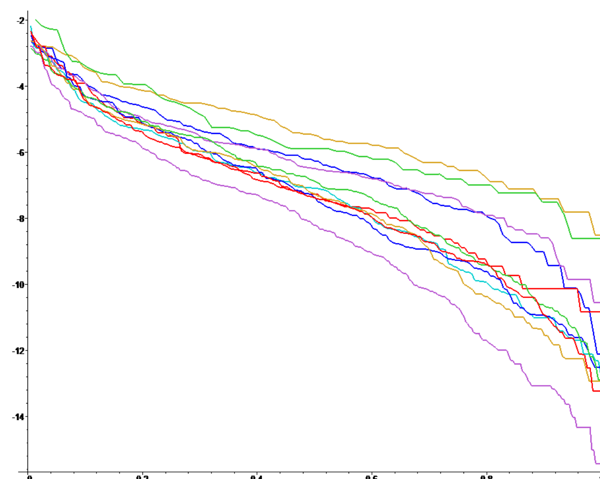
The cumulative frequency of symbols is shown in Figure 1 with one curve for each subject and one for the weighted combination. Figure 2 shows the same curves on a log plot, from which it is possible to see that the symbols follow an exponential distribution. Tables 9 and 10 show the most popular 2-grams and 5-grams respectively for the three authors of the selected corpus as well as from two comparison texts [19, 20]. 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 any n) was 479 388 for Kreyzig, 562 297 for Greenberg and 477 268 for 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).



X: Symbol rank number.
Y: Cumulative frequency.
(Each color represents a subject grouping of Table 5.)

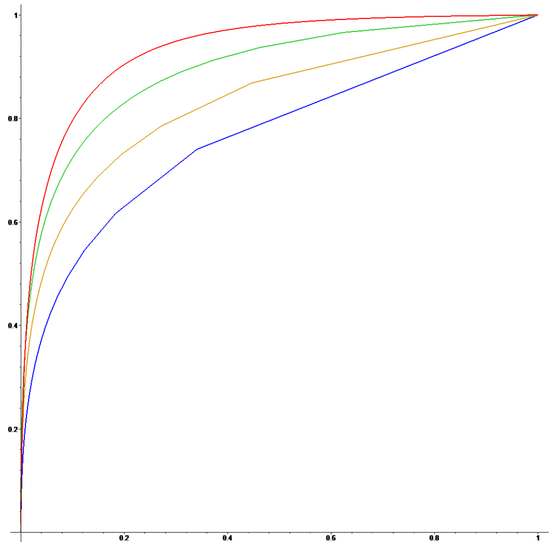
Figure 1. Cumulative symbol freq. by subject



X: Symbol rank number.
Y: Log frequency.
(Each color represents a subject grouping of Table 5.)

Figure 2. Log frequencies

Figure 3 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 rank-



From top to bottom, curves count 2-, 3-, 4- and 5-grams for entire corpus.

Figure 3. n-gram cumulative frequency

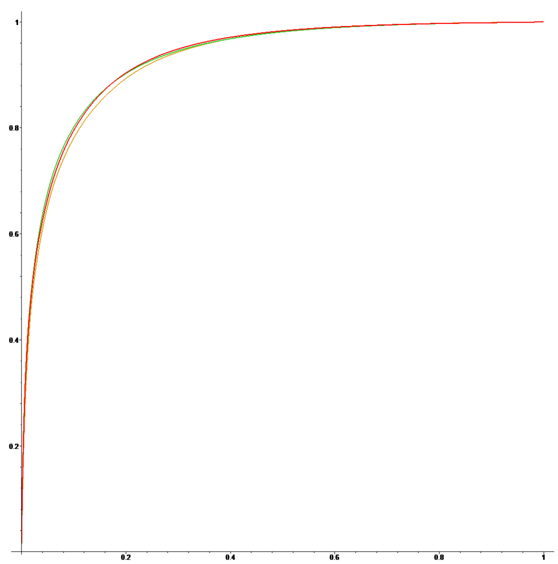


Figure 4. Bigram cumul. frequency per author

ing of the particular n-grams is different for the each author, the cumulative n-gram frequency curves are virtually identical for each author. Figure 4 shows the cumulative frequency of bigrams, ordered by popularity, for the three authors.

5 Conclusions

Earlier work had shown that statistical analysis of mathematical research documents could produce n-grams that improve mathematical handwriting recognition rates for high-level mathematics.

We have have therefore been motivated to define a more elementary corpus of mathematics that would be more widely applicable and analyze its statistical structure. We have selected second year engineering mathematics as taught in North America as the subject and have analyzed the expressions that occur in the textbooks are adopted in more than 90% of the classes. We are then able to produce statistics weighted by the popularity of the textbooks, thus modeling the set of expressions that are used in practice.

Analyzing the population of symbols and n-grams that occur in these texts, 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 all of the frequently occurring items. Thus applications, even those for portable devices, could use these statistics to guide their recognition.

Future work will explore how well this performs in practice for elementary engineering mathematics.

Acknowledgments

The author would like to thank Michael Greenberg, Peter O'Neil, Prentice-Hall and Thomas-Nelson for the use of their materials. We also thank Robert Lopez and Maplesoft for the use of additional materials. We thank Jeliazko Polihronov for his assistance in gathering the data and Elena Smirnova for her work on the n-gram analysis software. This work was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada, Microsoft and Maplesoft.

References

[1] Clare M. So and Stephen M. Watt, Determining Empirical Properties of Mathematical Expression Use, pp. 361-375, Proc. Fourth International Conference on Mathematical Knowledge Management, (MKM 2005), July 15-17 2005, Bremen Germany, Springer Verlag LNCS 3863.

[2] ArXiv e-Print Archive. `http://arxiv.org`

[3] Elena Smirnova and Stephen M. Watt, Combining Prediction and Recognition to Improve On-Line Mathematical Character Recognition (submitted). Available as ORCCA Technical Report TR-06-06. `http://www.orcca.on.ca/TechReports/2006/TR-06-06.html`

[4] Elena Smirnova and Stephen M. Watt, A pen-based mathematical environment "Mathink" (submitted). Available as ORCCA Technical Report TR-06-05. `http://www.orcca.on.ca/TechReports/2006/TR-06-05.html`

[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, Ed. C.Vanoirbeek, C.Roisin, E. Munson, 2003, pp.95-104

[6] S.Uchida, A.Nomura, and M.Suzuki, Quantitative analysis of mathematical documents, International Journal on Document Analysis and Recognition, Vol.7, Issue.4, pp.211-218. (September 2005)

[7] U.Garain and B.B Chaudhuri, A corpus for OCR research on mathematical expressions, International Journal on Document Analysis and Recognition, Vol.7, Issue.4, pp.241-259. (September 2005)

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

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

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

[11] Alan Jeffrey, *Advanced Engineering Mathematics*, 2^{nd} *edition*, Academic Press 2002.

[12] Thomas L. Harman, James Dabney, Norman J. Richert, *Advanced Engineering Mathematics with MATLAB*, 2^{nd} *edition*, Thomson-Engineering 2000.

[13] Dennis G. Zill, Michale R. Cullen, *Advanced Engineering Mathematics*, 3^{rd} *edition*, Jones and Bartlett 2006.

[14] Merle C. Potter, *Advanced Engineering Mathematics*, 3^{th} *edition*, Oxford University Press 2005.

[15] C. Ray Wylie, *Advanced Engineering Mathematics*, 6^{th} *edition*, McGraw-Hill 1995.

[16] David Carlisle, Patrick Ion, Robert Miner, Nico Poppelier, Editors. Mathematical Markup Language (MathML) Version 2.0 (Second Edition). W3C Recommendation. `http://www.w3.org/TR/2003/REC-MathML2-20031021/`. October 21, 2003.

[17] Ontario Research Centre for Computer Algebra. On-line TeX to MathML translator. `http://www.orcca.on.ca/MathML/texmml/textomml.html`

[18] Stephen M. Watt. Implicit Mathematical Semantics in Conversion between TEX and MathML, TUGBoat, Vol 23, No 1 (2002)

[19] R. Lopez, *Advanced Engineering Mathematics wtih Maple*, Maplesoft 2005.

[20] J. Weiner, *The Mathematics Survival Kit*, Maplesoft 2006.

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
$'$	1.24019	4427	7757	4749
a	1.21198	4627	6305	3390
5	1.14952	4771	3030	3674
i	0.91795	3451	4251	3940
u	0.91478	3392	5740	2121
c	0.89854	3638	3096	2727
s	0.87843	3539	3742	1756
d	0.84576	2761	6929	3460
e	0.84518	3010	4819	4019
$\vert$	0.81767	3270	2962	2691
π	0.76648	2913	2710	4243
/	0.75086	2849	3730	2557
6	0.72635	2981	1648	3088
k	0.71945	2892	2217	2807
]	0.70123	2698	3114	2558
[	0.70104	2697	3110	2565
m	0.64372	2712	2033	1114
8	0.55862	2374	963	1977
r	0.55619	2008	3348	2085
b	0.54474	2080	2731	1678
9	0.49895	2144	741	1698
sin	0.46307	1704	2190	2310
v	0.45379	1679	2863	1067
j	0.44919	1783	1565	1716
7	0.44045	1818	890	1930
$\cdots$	0.43961	1918	957	723
cos	0.42582	1667	1409	1988
∂	0.41059	1363	2621	2578
C	0.40904	1470	2918	906
A	0.40878	1517	2092	1660
$<$	0.39223	1334	2919	1487
p	0.38261	1249	2972	1815
$\leq$	0.38081	1534	1016	1693
$\int$	0.37320	1216	2585	2276
w	0.35636	1505	985	793
∞	0.34904	1093	2796	2019
A	0.34528	1294	1989	927
F	0.34459	1396	1402	708
L	0.32925	1097	2217	1848
λ	0.31876	1210	1817	722
h	0.29846	1195	1176	822
θ	0.27871	926	2266	995
T	0.27346	1078	1282	619
R	0.26417	1033	1114	878
P	0.26299	1021	1041	1057
D	0.24927	780	2531	629

(Top 65 out of 305)

Table 6. Top Symbols: All subjects combined

Symbol	Weighted Freq. (%)	Symbol Counts		
		Kreyszig	Greenberg	O'Neil
z	11.28007	5740	4155	4670
$=$	6.19577	3052	2879	2566
$)$	5.76133	2664	2761	4145
$($	5.75744	2661	2761	4152
1	5.59297	2790	2559	2006
2	5.21226	2520	2669	2223
$-$	4.02399	1912	2075	2058
0	3.88584	1934	1609	1756
$+$	3.71409	1793	1845	1719
i	2.95919	1358	1609	1888
n	2.94910	1504	1016	1315
$\vert$	2.78406	1381	1120	1282
x	2.45995	1125	1621	1086
f	1.98821	926	1004	1262
$,$	1.69837	839	842	579
y	1.60176	759	903	699
π	1.30730	631	537	815
C	1.18192	570	855	56
3	1.13346	527	683	524
d	1.10683	477	824	634
$/$	1.09869	519	619	498
e	1.08463	507	585	599
a	0.95106	383	893	497
4	0.85898	421	382	411
w	0.84605	406	344	560
∞	0.75755	352	459	348
u	0.72164	331	476	307
$<$	0.68370	309	511	230
s	0.63924	338	245	107
t	0.62265	248	397	704
θ	0.56642	273	255	316
r	0.55408	281	190	266
$\leq$	0.54975	279	134	363
$*$	0.52955	308	39	85
R	0.52740	261	285	131
c	0.51782	226	369	296
D	0.51152	186	572	317
$\cos$	0.50425	247	201	286
$.$	0.49176	154	350	949
$\int$	0.48189	210	324	315
$\cdots$	0.48000	263	141	59
$\sin$	0.47708	220	235	338
v	0.44584	178	438	214
b	0.42329	196	215	279
$'$	0.40738	162	379	242
F	0.39388	220	90	50
m	0.38962	205	100	168
$\sum$	0.37173	183	140	216
5	0.37171	179	204	141
Φ	0.36030	215	5	38
Δ	0.34065	176	183	5
$\rightarrow$	0.32410	135	230	258
$\oint$	0.32389	149	191	175
$\blacksquare$	0.29125	156	—	234
6	0.28148	139	93	183
k	0.26354	122	122	196
L	0.25968	132	93	114
$\square$	0.25574	108	207	140
$[$	0.25324	100	247	137
$]$	0.25234	100	247	131
∂	0.24631	106	145	214
$>$	0.24236	126	79	90
$!$	0.20472	101	100	74
$\neq$	0.19061	92	82	112
ϵ	0.18347	98	81	—
$\ln$	0.18341	100	42	50

(Top 65 out of 194.)

Table 7. Top Symbols: Complex Analysis

Symbol	Weighted Freq. (%)	Symbol Counts		
		Kreyszig	Greenberg	O'Neil
$=$	7.22187	1362	3661	2625
x	7.04832	1289	4080	2874
$($	6.44756	1125	3923	3745
$)$	6.43967	1123	3922	3751
2	5.54914	1064	2378	2187
0	4.82981	827	3394	2562
u	4.28608	822	2355	949
1	3.35806	607	2117	1306
n	3.33931	659	1175	1200
t	3.10607	594	2205	1859
y	2.63211	480	1442	1224
$-$	2.38819	419	1485	1283
$+$	2.02753	354	1521	764
$,$	2.00038	308	2050	1031
c	1.68841	334	645	514
r	1.67920	325	817	446
π	1.66333	317	593	875
f	1.38602	260	722	512
∂	1.32067	245	285	1130
m	1.13741	249	165	114
L	1.08369	184	748	634
w	0.97867	217	153	12
$\sin$	0.97415	184	412	463
$.$	0.93572	136	649	1129
d	0.88630	171	301	434
$/$	0.85220	165	418	220
s	0.83542	164	457	91
F	0.79309	161	277	173
∞	0.73690	117	589	525
θ	0.70563	130	396	281
4	0.69761	132	318	293
3	0.66666	120	419	274
G	0.65305	145	83	30
B	0.63865	138	143	39
k	0.63278	109	420	354
5	0.61589	110	445	193
p	0.60337	130	127	58
e	0.58116	100	423	275
a	0.57791	102	441	179
φ	0.55596	120	93	82
$\cos$	0.55098	108	112	333
A	0.53950	106	244	129
$'$	0.51395	74	546	369
$\int$	0.50346	92	164	388
α	0.48979	82	513	68
R	0.43566	86	130	189
v	0.43559	92	144	15
λ	0.41367	88	41	134
$<$	0.40432	73	950	470
i	0.39696	77	227	53
b	0.38346	61	375	173
T	0.32036	54	163	280
8	0.29625	61	63	100
Δ	0.28748	48	259	103
$\sum$	0.26555	49	127	135
z	0.26496	39	248	216
$\cdots$	0.25696	56	60	12
$]$	0.25482	36	231	257
$[$	0.25441	36	230	256
C	0.24920	49	135	27
$*$	0.24767	55	33	9
ω	0.24534	16	172	797
g	0.24349	41	164	157
$\leq$	0.24063	36	69	403
J	0.23427	44	132	71
W	0.22762	51	28	—

(Top 65 out of 193.)

Table 8. Top Symbols: PDEs

Kreyszig		Greenberg		O'Neil		Lopez [19]		MSKit [20]	
Freq (%)	**Sequence**	**Freq (%)**	**Sequence**	**Freq (%)**	**Sequence**	**Freq (%)**	**Sequence**	**Freq (%)**	**Sequence**
0.015609	$1\langle/\mathtt{sub}\rangle$	0.013729	$(x$	0.013652	$)=$	0.015275	$\langle\mathtt{sup}\rangle 2$	0.026046	$(x$
0.013716	$\langle\mathtt{sub}\rangle 1$	0.012302	$1\langle/\mathtt{sub}\rangle$	0.013640	$\langle\mathtt{sup}\rangle 2$	0.015171	$2\langle/\mathtt{sup}\rangle$	0.019772	$x)$
0.012866	$2\langle/\mathtt{sup}\rangle$	0.011704	$2\langle/\mathtt{sup}\rangle$	0.013500	$2\langle/\mathtt{sup}\rangle$	0.012549	$(x$	0.018647	$2\langle/\mathtt{sup}\rangle$
0.012828	$\langle\mathtt{sup}\rangle 2$	0.011643	$\langle\mathtt{sup}\rangle 2$	0.012630	$(x$	0.009457	$)=$	0.017966	$\langle\mathtt{sup}\rangle 2$
0.011231	$2\langle/\mathtt{sub}\rangle$	0.011210	$)=$	0.008977	$t)$	0.009434	00	0.015542	$x\langle\mathtt{sup}\rangle$
0.011127	$\langle\mathtt{sub}\rangle 2$	0.010881	$\langle\mathtt{sub}\rangle 1$	0.008486	$x)$	0.009044	-1	0.013704	$)=$
0.009607	$)=$	0.008806	$=0$	0.008406	$1\langle/\mathtt{sub}\rangle$	0.008534	$x)$	0.010710	$x+$
0.009482	$(x$	0.008434	$x)$	0.008301	-1	0.007400	$1\langle\mathtt{frac/}\rangle$	0.010583	$n($
0.009255	$x\langle\mathtt{sub}\rangle$	0.007672	$2\langle/\mathtt{sub}\rangle$	0.007969	$e\langle\mathtt{sup}\rangle$	0.007261	$t)$	0.009933	-1
0.008517	$\langle/\mathtt{sub}\rangle =$	0.007556	$e\langle\mathtt{sup}\rangle$	0.007835	$0\langle/\mathtt{sub}\rangle$	0.007216	$1\langle/\mathtt{sub}\rangle$	0.009696	$x-$
0.007745	$0\langle/\mathtt{sub}\rangle$	0.007504	$\langle\mathtt{sub}\rangle 2$	0.007278	$(t$	0.006365	$\langle/\mathtt{sup}\rangle +$	0.008967	$\langle/\mathtt{sup}\rangle +$
0.007711	$=0$	0.006287	$0\langle/\mathtt{sub}\rangle$	0.007161	$\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle$	0.005821	$\langle\mathtt{sub}\rangle 1$	0.008650	$y=$
0.007060	$\langle\mathtt{sub}\rangle 0$	0.006233	$x\langle\mathtt{sub}\rangle$	0.006839	$\langle\mathtt{sub}\rangle 0$	0.005767	$0 :=$	0.008618	$x=$
0.007030	$y\langle\mathtt{sub}\rangle$	0.006224	$\langle/\mathtt{sub}\rangle =$	0.006631	$\langle\mathtt{sub}\rangle n$	0.005740	$(t$	0.008365	$\langle\mathtt{root/}\rangle 2$
0.006699	-1	0.006182	$\langle\mathtt{sub}\rangle n$	0.006592	$1\langle\mathtt{frac/}\rangle$	0.005608	$x\langle\mathtt{sup}\rangle$	0.008349	$1\langle\mathtt{frac/}\rangle$
0.006676	$=1$	0.006182	$\langle\mathtt{sup}\rangle\hat{}$	0.006258	$=0$	0.005419	$e\langle\mathtt{sup}\rangle$	0.008333	dx
0.006362	$0.$	0.006182	$\hat{}\langle/\mathtt{sup}\rangle$	0.006033	$\langle\mathtt{sub}\rangle 1$	0.005344	$< root/ > 2$	0.007002	$+1$
0.006349	$\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle$	0.006015	-1	0.005804	$\langle/\mathtt{sup}\rangle +$	0.005287	$\langle\mathtt{sub}\rangle k$	0.006939	$2x$
0.006187	$\langle\mathtt{sub}\rangle n$	0.005707	$\langle\mathtt{sub}\rangle 0$	0.005798	$f($	0.005178	$\langle\mathtt{frac/}\rangle 2$	0.006876	$f($
0.005891	$x)$	0.005471	$\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle$	0.005506	$=1$	0.005061	$f($	0.006464	$3\langle/\mathtt{sup}\rangle$
0.005847	$\langle\mathtt{sup}\rangle\prime$	0.005457	$\langle/\mathtt{sub}\rangle($	0.005479	$x\langle\mathtt{sup}\rangle$	0.005007	10	0.006432	$=1$
0.005831	$\prime\langle/\mathtt{sup}\rangle$	0.005333	$t)$	0.005372	$n($	0.004949	$2\langle/\mathtt{sub}\rangle$	0.005877	$\langle/\mathtt{sup}\rangle -$
0.005706	$e\langle\mathtt{sup}\rangle$	0.004978	$n\langle/\mathtt{sub}\rangle$	0.005139	$\langle/\mathtt{sup}\rangle($	0.004941	$=1$	0.005656	$\langle/\mathtt{sub}\rangle($
0.005546	$\langle/\mathtt{sup}\rangle +$	0.004962	in	0.005133	$2\langle/\mathtt{sub}\rangle$	0.004806	$\langle\mathtt{sub}\rangle 2$	0.005640	$)\langle\mathtt{sup}\rangle$
0.005454	$1\langle\mathtt{frac/}\rangle$	0.004903	$\langle/\mathtt{sup}\rangle +$	0.005077	$y\langle\mathtt{sup}\rangle$	0.004792	$=0$	0.005513	$1)$

Table 9. Most Popular Bigrams

Kreyszig		Greenberg		O'Neil		Lopez		MSKit	
Freq (%)	**Sequence**	**Freq (%)**	**Sequence**	**Freq (%)**	**Sequence**	**Freq (%)**	**Sequence**	**Freq (%)**	**Sequence**
0.001049	(x,y)	0.002046	$e\langle\mathtt{sup}\rangle\hat{}\langle/\mathtt{sup}\rangle\langle\mathtt{sub}\rangle$	0.001519	(x,y)	0.002845	00000	0.004420	$lim\langle\mathtt{sub}\rangle x$
0.000951	$y\langle\mathtt{sup}\rangle\prime\prime\langle/\mathtt{sup}\rangle$	0.001415	$\int\langle\mathtt{sub}\rangle 0\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle$	0.001488	$\int\langle\mathtt{sub}\rangle 0\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle$	0.001361	(x,y)	0.004055	$im\langle\mathtt{sub}\rangle x \rightarrow$
0.000816	$x\langle\mathtt{sub}\rangle 1\langle/\mathtt{sub}\rangle +$	0.001295	(x,y)	0.001020	$0\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle\infty\langle/\mathtt{sup}\rangle$	0.001202	$x\langle\mathtt{sup}\rangle 2\langle/\mathtt{sup}\rangle +$	0.003200	$x\langle\mathtt{sup}\rangle 2\langle/\mathtt{sup}\rangle +$
0.000812	$f(x)=$	0.000774	$0\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle\infty\langle/\mathtt{sup}\rangle$	0.001001	$\sum\langle\mathtt{sub}\rangle n=1$	0.001045	$\int\langle\mathtt{sub}\rangle 0\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle$	0.002851	$dy\langle\mathtt{frac/}\rangle dx$
0.000803	$\int\langle\mathtt{sub}\rangle 0\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle$	0.000770	$x\langle\mathtt{sup}\rangle 2\langle/\mathtt{sup}\rangle +$	0.000999	$\langle\mathtt{sub}\rangle n=1\langle/\mathtt{sub}\rangle$	0.000959	$f(x)=$	0.001996	$f(x)=$
0.000728	$0\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle\infty\langle/\mathtt{sup}\rangle$	0.000711	(x,t)	0.000995	$n=1\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle$	0.000736	$x\langle\mathtt{sup}\rangle 2\langle/\mathtt{sup}\rangle -$	0.001996	$sin(x$
0.000722	$\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle\prime\langle/\mathtt{sup}\rangle =$	0.000699	$1\langle/\mathtt{sub}\rangle,\ldots,$	0.000935	$1\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle\infty\langle/\mathtt{sup}\rangle$	0.000703	$(x,y,$	0.001901	$x\langle\mathtt{sup}\rangle 2\langle/\mathtt{sup}\rangle -$
0.000718	$x\langle\mathtt{sup}\rangle 2\langle/\mathtt{sup}\rangle +$	0.000674	$\langle\mathtt{sub}\rangle 1\langle/\mathtt{sub}\rangle,\ldots$	0.000927	$=1\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle\infty$	0.000663	$\langle\mathtt{sup}\rangle 2\langle/\mathtt{sup}\rangle +1$	0.001885	$in(x)$
0.000706	$\langle\mathtt{sup}\rangle\prime\prime\langle/\mathtt{sup}\rangle +$	0.000615	$y(x)=$	0.000898	$)sin($	0.000647	$,\ldots,$	0.001536	$2x\langle\mathtt{sup}\rangle 2\langle/\mathtt{sup}\rangle$
0.000635	$-z\langle\mathtt{sub}\rangle 0\langle/\mathtt{sub}\rangle$	0.000601	$\langle\mathtt{sub}\rangle 0\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle\infty$	0.000859	$x\langle\mathtt{sup}\rangle 2\langle/\mathtt{sup}\rangle +$	0.000644	$f(x,y$	0.001410	$cos(x$
0.000599	.....	0.000571	$\langle/\mathtt{sub}\rangle(x)=$	0.000839	$(-1)\langle\mathtt{sup}\rangle$	0.000643	$+y\langle\mathtt{sup}\rangle 2\langle/\mathtt{sup}\rangle$	0.001330	$os(x)$
0.000566	$\langle\mathtt{sub}\rangle 1\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle\prime$	0.000566	$i\langle\mathtt{sup}\rangle\hat{}\langle/\mathtt{sup}\rangle +$	0.000820	$sin(n$	0.000609	x,y,z	0.001314	$x\langle\mathtt{sup}\rangle 3\langle/\mathtt{sup}\rangle +$
0.000553	$z\langle\mathtt{sub}\rangle 0\langle/\mathtt{sub}\rangle)$	0.000562	$(0)=0$	0.000721	$,\ldots,$	0.000604	$,y,z)$	0.001235	$\langle\mathtt{sup}\rangle 2\langle/\mathtt{sup}\rangle +1$
0.000551	$y\langle\mathtt{sub}\rangle 1\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle$	0.000561	$f(x,y$	0.000700	(x,t)	0.000588	$2x\langle\mathtt{sup}\rangle 2\langle/\mathtt{sup}\rangle$	0.001219	$log\langle\mathtt{sub}\rangle a$
0.000545	$y(0)=$	0.000522	$\langle\mathtt{sub}\rangle 1\langle/\mathtt{sub}\rangle(x$	0.000678	$-1)\langle\mathtt{sup}\rangle n$	0.000585	$\langle\mathtt{sup}\rangle 2\langle/\mathtt{sup}\rangle +y$	0.001219	$og\langle\mathtt{sub}\rangle a\langle/\mathtt{sub}\rangle$
0.000543	$1\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle\prime\langle/\mathtt{sup}\rangle$	0.000522	$1\langle/\mathtt{sub}\rangle(x)$	0.000651	$(x,y,$	0.000579	$)sin($	0.001172	$y\langle\mathtt{frac/}\rangle dx=$
0.000512	$\langle\mathtt{sub}\rangle 2\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle\prime$	0.000515	$1\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle\infty\langle/\mathtt{sup}\rangle$	0.000643	$-\infty\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle\infty$	0.000576	$\langle/\mathtt{sup}\rangle cos($	0.001156	$< root/ > 2x\langle\mathtt{sup}\rangle 2$
0.000508	$z-z\langle\mathtt{sub}\rangle 0$	0.000513	x,y,z	0.000643	$\infty\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle\infty\langle/\mathtt{sup}\rangle$	0.000569	$\langle/\mathtt{sup}\rangle sin($	0.001156	$(x\langle\mathtt{sup}\rangle 2\langle/\mathtt{sup}\rangle$
0.000501	$,y\langle\mathtt{sub}\rangle 2\langle/\mathtt{sub}\rangle$	0.000505	$f(x)=$	0.000641	$z\langle\mathtt{sub}\rangle 0\langle/\mathtt{sub}\rangle)$	0.000552	$0\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle\infty\langle/\mathtt{sup}\rangle$	0.001156	$du\langle\mathtt{frac/}\rangle dx$
0.000497	$y\langle\mathtt{sub}\rangle 2\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle$	0.000494	$(x,y,$	0.000620	$\langle\mathtt{sub}\rangle -\infty\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle$	0.000516	$2\langle/\mathtt{sup}\rangle +y\langle\mathtt{sup}\rangle$	0.001156	$g\langle\mathtt{sub}\rangle a\langle/\mathtt{sub}\rangle($
0.000495	$\langle\mathtt{sub}\rangle n+1\langle/\mathtt{sub}\rangle$	0.000444	$\langle\mathtt{sup}\rangle\hat{}\langle/\mathtt{sup}\rangle\langle\mathtt{sub}\rangle\theta$	0.000620	$\langle/\mathtt{sup}\rangle sin($	0.000500	$sin(t$	0.001093	$=log\langle\mathtt{sub}\rangle$
0.000489	$2\langle/\mathtt{sub}\rangle\langle\mathtt{sup}\rangle\prime\langle/\mathtt{sup}\rangle$	0.000444	$\hat{}\langle/\mathtt{sup}\rangle\langle\mathtt{sub}\rangle\theta\langle/\mathtt{sub}\rangle$	0.000613	$\int\langle\mathtt{sub}\rangle -\infty\langle/\mathtt{sub}\rangle$	0.000499	$\langle/\mathtt{sup}\rangle +y\langle\mathtt{sup}\rangle 2$	0.001093	$ln(x)$
0.000487	$x\langle\mathtt{sub}\rangle 2\langle/\mathtt{sub}\rangle =$	0.000438	$\langle\mathtt{sub}\rangle n\langle/\mathtt{sub}\rangle(x$	0.000611	$\langle/\mathtt{sup}\rangle cos($	0.000478	$cos(t$	0.001013	$\langle\mathtt{sup}\rangle 2\langle/\mathtt{sup}\rangle(x$
0.000485	$f(x,y$	0.000438	$sinn\pi$	0.000599	$)cos($	0.000476	$\langle\mathtt{sub}\rangle k\langle/\mathtt{sub}\rangle(x$	0.001013	$y=f(x$

Table 10. Most Popular 5-grams