The Structure of Mathematical Expressions

An arXiv Case Study

Deyan Ginev and Bruce R. Miller

National Institute of Standards and Techonology

March 29, 2012

Contents

Chapter 1

Introduction

In this study, we survey the notational diversity of present-day mathematical expressions, in order to uncover their linguistic phenomena. A practical motivation for this study is to provide a foundation for determining the boundary between syntactic and semantic phenomena in said expressions, from the perspective of language modeling. The ultimate goal of this project is to construct a grammar of mathematical expressions, which captures all relevant syntactic properties established in this study, and allows for the semantic analysis necessary to model and observe the semantic relationships.

1.1 Motivation

We want to enable machine-reading of formulas, in order to provide a variety of user-assistance services, such as semantic search, text-to-speech synthesis, semantic interactions (definition lookup), as well as computer algebra support ("evaluate subexpressions on demand") and ultimately computer verification ("does that proof step really hold?").[1] EdN:1

1.2 Related Resources

Notation census, beginnings of study are in Deyan's thesis, Naproche and FMathL have examples, but no real systematic study.[2] EdN:2

[1] EdNote: expand
[2] EdNote: expand

Chapter 2

Methods

2.1 Training Corpus

The primary corpus on which we base this investigation is the Cornel pre-print archive "arXiv"[3], consisting of over 700,000 articles in 37 scientific subfields. EdN:3

arXiv Sandbox

[4] EdN:4

As a secondary resource, we we will also consult entry-level literature on highschool mathematics, in order to exhibit basic phenomena, as well as to demonstrate phenomena apriori known to the authors.[5] EdN:5

2.2 Structural Annotation

As one of the goals of our study is to establish a first guess of an underspecified operator tree[6], any annotation must at its core mark up the applicative logical structure of the mathematical expression. This process will build up a formula tree, the collection of which can later be used as a gold standard for developing a grammatical model of the language of symbolic mathematics. EdN:6

[7] [8] EdN:7 EdN:8

[3] EdNote: cite here

[4] EdNote: Say that, on the arXiv front, we first start with the train sandbox from Deyan's thesis

[5] EdNote: Wikipedia? PEMDAS?

[6] EdNote: make sure the concepts are introduced and/or rephrase

[7] EdNote: I'm currently thinking of rendering the annotations as trees (tikz,pstricks...custom tree drawing package?), so that the annotator can proofread the annotations in an intuitive manner.

[8] EdNote: In the XHTML, I'm thinking of ContentMML+SVG rendering, all of this figured out by the binding, maybe a custom stylesheet?

Train1	Differential Geometry `http://arxmliv.kwarc.info/files/9609/dg-ga.9609012`
Train2	Quantum Physics `http://arxmliv.kwarc.info/files/0910/0910.5733/`
Train3	High Energy Physics - Theory `http://arxmliv.kwarc.info/files/9407/hep-th.9407125/`
Train4	Commutative Algebra `http://arxmliv.kwarc.info/files/0809/0809.4873/`
Train5	Statistics Theory `http://arxmliv.kwarc.info/files/0905/0905.1486/`
Train6	General Relativity and Quantum Cosmology `http://arxmliv.kwarc.info/files/0807/0807.2507/`
Train7	Cosmology and Extragalactic Astrophysics `http://arxmliv.kwarc.info/files/0908/0908.2548`
Train8	Exactly Solvable and Integrable Systems `http://arxmliv.kwarc.info/files/0905/0905.2033`
Train9	Geometric Topology `http://arxmliv.kwarc.info/files/0809/0809.4477`
Train10	Algebraic Geometry `http://arxmliv.kwarc.info/files/0704/0704.0537`

Table 2.1: Sandbox of Ten Random arXiv Papers from Diverse Scientific Subfields

2.3 Annotation Vocabulary

Another core goal is to discover and describe interesting linguistic phenomena that occur naturally in our corpus. Examples of what we consider "interesting" are phenomena that induce ambiguity, or legitimize what would typically be ungrammatical fragments. Cases of ambiguity are well-known to follow from semantic overloading of symbols, implicit argument scopes of operations or eliding syntax, leaving the reader with the task of guessing the "invisible" dynamics.Use of custom shorthands, however, as well as custom notations in general, expands the grammar of symbolic mathematics, often in completely non-standard ways that can only be grasped through a deep understanding of the document at hand.

As multiple interesting observations can be made for a single large mathematical formula, it is natural to annotate multiple relevant subexpressions. More concretely, for each phenomenon of interest, we annotate the greatest common subtree (GCT) of all participating subtrees. In case we find a long-

Property	Keywords
Fixity	prefix, infix, postfix, superfix, subfix, circumfix, transfix, nofix[1]
Role (Symbols)	separator, modifier, relation, operator, metarelation
Role (Objects)	factor, term, statement, variable, constant, modified
Role (Structure)	tuple, sequence, expression, shorthand, notation
Composition	invisible, atom, complex, chained
Shallow Semantics	type, function, constructor, other
Linguistic	ellipsis, metonymy, ambiguity, vagueness, anaphora
Math Practices	framing

Table 2.2: Keyword Vocabulary for Syntactic Properties

range relationship in a large formula, the annotation would hence be placed on the formula root.

The annotations can be utilized for different purposes - browsing by specific phenomena, syntactic feature or lemma, training a classifier, etc. Thus, we take a compositional, standardized approach to providing labels from a fixed vocabulary for the relevant ontological classes of structural properties.

Chapter 3

A Study of Mathematical Syntax

3.1 Basics

Foundations

[9] [10] [11] EdN:9 EdN:10 EdN:11

High School

[12] [13] EdN:12 EdN:13

3.2 Discrete math

Set Theoretic Notations

[14] [15] EdN:14 EdN:15

Logical Operators

[16] EdN:16

[9] EdNote: arithmetic, grouping fences and equality
[10] EdNote: basic relations and orderings
[11] EdNote: arithmetic and algebraic sequences?
[12] EdNote: geometry here, otherwise a separate geometry subsection
[13] EdNote: trigonometry, complex and rational numbers
[14] EdNote: elementhood, inclusions, set constructors, overloaded arith ops
[15] EdNote: also maps : domains -¿ codomains, xRy notations
[16] EdNote: classic logic, HOL, type theories

Combinatorics

EdN:17
EdN:18

[17] [18]

Number Theory

EdN:19
EdN:20
EdN:21
EdN:22

[19] [20] [21] [22]

Graph Theory

EdN:23
EdN:24
EdN:25

[23] [24] [25]

Algebra

EdN:26
EdN:27
EdN:28
EdN:29

[26] [27] [28] [29]

Functions Theory

EdN:30

[30]

3.3 Continuous math

Calculus

EdN:31

[31]

Probability

EdN:32
EdN:33

[32] [33]

---

[17]EdNote: Infinite sums
[18]EdNote: binomials, combinations, permutations,
[19]EdNote: modulo modifiers
[20]EdNote: tuples
[21]EdNote: divisibility notations $a \mid b$ and b/a
[22]EdNote: DLMF sneaky notations
[23]EdNote: edge and vertex notations
[24]EdNote: incidence and adjacency notations
[25]EdNote: Wiki is very nice: `http://en.wikipedia.org/wiki/Glossary\_of\_graph\_theory`
[26]EdNote: vectors
[27]EdNote: maps and complements
[28]EdNote: groups
[29]EdNote: lattices
[30]EdNote: talk about associativity of application and composition, ";" and "$\circ$" as notation variants, discuss complex examples
[31]EdNote: differentials, integrals, limits, remember brownian motion integral notations!
[32]EdNote: Bayes formula with multiple denotations of P
[33]EdNote: Various conditional and joint probability notations

Interval Notation and Arithmetic

[34] EdN:34

Topology

[35] EdN:35

3.4 Other fields

Quantum Physics

[36] [37] ⋮ EdN:36 EdN:37

[34] EdNote: introduce interval notations, then move to interval arithmetic
[35] EdNote: manifold constructors and notations
[36] EdNote: Bra-ket notation
[37] EdNote: computer science, biology, chemistry...

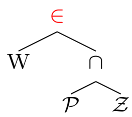
	Expression	Denotation	Annotation
1.	$W \in \mathcal{P} \cap \mathcal{Z}$	set membership	
	Discussion: set ops precede set relations, [Train1]		

Table 3.1: Set Theory Notations, Part 1

Chapter 4

Discussion

Chapter 5

Conclusion