Magdalena Wolska; Mihai Grigore
Symbol Declarations in Mathematical Writing

In: Petr Sojka (ed.): Towards a Digital Mathematics Library. Paris, France, July 7-8th, 2010. Masaryk University Press, Brno, Czech Republic, 2010. pp. 119--127.

Persistent URL: http://dml.cz/dmlcz/702580

Symbol Declarations in Mathematical Writing
A Corpus Study

Magdalena Wolska[1] and Mihai Grigore[2]

[1] Fachrichtung 4.7 Allgemeine Linguistik, Universität des Saarlandes
D-66041 Saarbrücken, Germany
magda@coli.uni-sb.de
[2] Computer Science, Jacobs University Bremen, D-28759 Bremen, Germany
m.grigore@jacobs-university.de

Abstract. We present three corpus-based studies on symbol declaration in mathematical writing. We focus on simple object denoting symbols which may be part of larger expressions. We look into whether the symbols are explicitly introduced into the discourse and whether the information on once interpreted symbols can be used to interpret structurally related symbols. Our goal is to support fine-grained semantic interpretation of simple and complex mathematical expressions. The results of our analysis empirically show the potential benefit of using larger discourse context in automated disambiguation of mathematical expressions.

Key words: mathematical discourse, disambiguation of mathematical expressions, corpus-based analysis

1 Motivation

Semantic search in mathematical documents, in order for it to account for their full mathematical content, must necessarily provide ways of searching through the symbolic expressions which are part of mathematical discourse. While dedicated approaches to formulae search do exist (see, for instance, [11,10] and references therein) they typically depend on semantically-oriented mark-up in their internal representation of mathematical expressions; be it OpenMath [5] or Content MathML [4]. Recent years have therefore seen increasing efforts towards improving automatic creation of machine-readable semantics-enriched mathematical documents [15].

Automatically inferring the semantics of a mathematical expression, both as a whole and of its constituent parts, is, however, a non-trivial task because of the infinite nature of the mathematical alphabet: new symbols may be invented, constructed from existing symbols, existing symbols may be typographically enriched to form new symbols, etc. All this possibly in a single document. There are of course certain conventions as to the usage of mathematical notation,

[1] Correspondence to: Magdalena Wolska, Universität des Saarlandes, Fachrichtung 4.7 Allgemeine Linguistik, Building C 7.2, Postfach 15 11 50, 66041 Saarbrücken, Germany; tel.: +49 681 302 4344

Petr Sojka (editor): DML 2010, Towards a Digital Mathematics Library, pp. 119–127.
 ISBN 978-80-210-5242-0

general and specific to mathematical sub-areas, as well as prescriptive rules on how to write mathematics (see, for instance, [9,12]) which mathematics' authors tend to follow, however, automated interpretation of arbitrary mathematical expressions remains a challenging task.

We performed a quantitative analysis of a subset of the arXMLiv collection [3] processed using LaTeXML [15], the state-of-the-art mathematical document processing architecture, and found out that approximately 41% of all the parsed mathematical symbols have not been interpreted by the LaTeXML grammar (2,842,813 out of a total number of 6,872,419 mathematical symbols); where by "not interpreted" we mean that the grammatical role attribute in the internal LaTeXML representation, the XMath `role`, has been set to `unknown`.

In our previous work [8], we showed that the local linguistic context, within which mathematical expressions are embedded, provides a good source of information for recognizing the denotation of mathematical expressions. Our approach, however, treats a mathematical term as a whole and attempts to identify an object type to which the entire term refers.

In this paper, we present three corpus-based studies which are meant to complement our previous work and constitute a step towards compositional semantic analysis of symbolic expressions. We now focus on simple object denoting symbols which may be part of larger expressions and ask, paraphrasing Knuth and colleagues, whether in actual mathematical papers "[a]ll variables [are] defined, at least informally, when they are first introduced" [9]. Certainly not all of them are: certain notational conventions are taken for granted, especially in academic scientific papers. They constitute part of what Clark calls *communal* (in this case, professional) *common ground* [6]. Our question of interest is rather "how much" of the notation is left implicit. More specifically, in the three studies described in this paper we were interested in the following questions:

1. To what extent are mathematical symbols systematically explicitly introduced into the discourse in mathematical scientific publications?
2. To what extent can symbol interpretation rely on larger local discourse context?
3. Can symbol interpretation be supported by an analysis of locally co-occurring symbolic expressions of similar structure?

Outline: The paper is organised as follows: In Section 2 we describe our corpus-based methodology: first we briefly describe the data set we use, followed by the descriptons of our three study setups. In Section 3 we present quantitative results of our studies. We conclude with a discussion of the results in Section 4 and discuss further work in Section 5.

2 Method

We performed three corpus studies in order to investigate symbol declaration practices in mathematical scientific papers. In all the experiments we used

actual mathematical papers as they were originally published. The setup of our studies is outlined below.

2.1 Data and Preprocessing

The subsets of documents we used in the studies were randomly selected from a corpus of 1,000 mathematical publications from the arXMLiv collection, processed by the LaTeXML architecture [14,15]. arXMLiv is subset of the arXiv, an archive of electronic preprints of scientific papers in the fields of, among others, mathematics, statistics, physics, and quantitative biology [2]. That is, the documents we analyzed were advanced scientific contributions written by professional mathematicians.

The documents have been word- and sentence-tokenized. For the analysis of symbolic expressions, we used two mathematical expression markup formats: the XMath format, a LaTeXML internal representation, and the Presentation MathML format, a widely used W3C standard for rendering mathematical content on the Web [4,13].

2.2 Experiments

In the experiments presented here, we were interested in object-denoting terms of "simple" high-level structure. More specifically, as "simple" symbols we consider atomic identifiers and super- or sub-scripted atomic identifiers; we do not, however, analyse the expression(s) in the super-/sub-scripts. In the following sections, we will use the term *simple mathematical expression* to refer to this class of symbols. We extracted the expressions of interest by parsing the XMath and MathML representations.

The first study The purpose of the first experiment was to investigate mathematicians' practices as to explicitly declaring symbols in their scientific writing. We randomly selected 50 documents from the preprocessed collection and from each document we randomly extracted 10 simple mathematical expressions. Next, we manually checked whether among the first 5 occurrences of these expressions in the paper, the symbol is explicitly declared; i.e. we inspected 500 simple expressions (2,500 occurences).

In this and the following study, we considered two types of declarations: a symbol may be introduced in isolation, as in the fragment: "Let F be a *Hermitian vector bundle* over $W \ldots$", or embedded in a larger symbolic expression which additionally elaborates the properties of the object denoted by the symbol, as in: "Consider the cylinder $\mathsf{U} = M \times [-\epsilon, 0) \ldots$", where U is further qualified to a have certain property. We will refer to the former type of declaration as *unqualified* and to the latter as *elaborated* (the declared symbol is further qualified by the sub-expression within which it appears). The point of this distinction is

```
Let
<XMath>
    <XMApp>
        <XMTok role="SUBSCRIPTOP" scriptpos="post2"/>
        <XMTok role="UNKNOWN" font="italic">C</XMTok>
        <XMTok role="UNKNOWN" font="italic">i</XMTok>
    </XMApp>
</XMath>
be the closed convex hull in
<XMath>
    <XMTok role="UNKNOWN" font="italic">Y</XMTok>
</XMath>
of the tail end of the sequence.
```

Fig. 1. A fragment of LaTeXML XMath markup with elements of `unknown` roles

that the declarations of elaborated expressions require more sophistication in the process of their automated identification.[3]

The second study The second study was a more focused variant of the first study. This time we were interested in simple expressions which are embedded in complex expressions, in particular those simple expressions whose grammatical role has not been recognized by the LaTeXML process.

The grammatical role, specified in the `role` attribute of the XMath markup, captures the syntactic nature of a symbol, the "grammatical role" that the object plays in surrounding expressions. The role attribute is used in generating the presentation markup and it can also help drive the derivation of the semantics of an expression.

Examples of role attributes which the LaTeXML parser does recognize include: `ATOM` (a general atomic subexpression), `APPLYOP` (an explicit infix application operator), `RELOP` (a relational operator), `ADDOP` (an addition operator), `INTOP` (an integral operator) [1]. Unrecognized symbols are assigned an `UNKNOWN` attribute by default, as illustrated in Figure 1.

In this study, we randomly extracted a subset of 100 mathematical documents from the collection. From each document we randomly selected 3 mathematical expressions whose XMath representations contained at least 3 simple expressions as defined above. For each of the resulting 300 expressions, we extracted all the simple sub-expressions tagged as `UNKNOWN` in the XMath representation and which *also* occurred independently in the discourse. For instance, for the expression $\rho =< \omega_i, \lambda >$, we would have extracted the following simple expressions should they be tagged as `UNKNOWN`: ρ, ω_i, and λ. Then we would check which of the resulting simple expressions occurred

[3] One could, for instance, assume that in the process of identifying declaration statements, it would be sufficient to consider as candidates only those sentences which contain unqualified simple object-denoting expressions. As we will show in the next section, this approach would miss a small percentage of instances.

Table 1. Results of the first study

Category		Occurrence	n (%)	N (%)
Explicitly declared	*unqualified*	1^{st}	290 (58%)	337 (67.4%)
		2^{nd}	15 (3.0%)	
		3^{rd}	11 (2.2%)	
		4^{th}	6 (1.2%)	
		5^{th}	2 (0.4%)	
	elaborated	—	13 (2.6%)	
Not explicitly declared				134 (26.7%)
Other				30 (5.9%)

in isolation in the document and select those for the analysis. We performed analogous manual analysis of the extracted instances as in the first study.

The third study Now, assuming that new symbols are systematically properly declared, in the third study, we were interested in finding out whether related symbols, which might not be individually introduced, indeed tend to be semantically related (that is, denote the same concept or different instances of the same concept). More specifically, we looked at simple terms based on the same main identifiers, i.e. sharing the same root/top-level node in the expression tree, and which are structurally similar modulo the structure of the subscript and superscipt terms. For instance, the following two expressions are structurally similar according to our criteria: ω_i and ω_{n-1}. By contrast, P_c^2 and A_n^k are not similar because they differ in the top-node operator.

For each pair of such expressions we verified whether the objects they denote are also semantically related if they occur in the same local discourse context. As discourse context we considered a section of a document; in the current study we ignored sub-section scopes. We randomly selected 25 mathematical documents and from each section of these documents we extracted all the pair-wise combinations of simple mathematical expressions which shared the same root symbol (same identifier) and either have the same surface structure or one expression is embedded in the other; 496 such pairs were extracted. Again, we analysed the extracted pairs manually as to whether each pair of expressions denotes the same concept in the context of the section scope.

The point of this study was to empirically verify whether the local discourse scope is a good indicator of semantic relatedness of structurally similar terms. Identification of structurally similar pairs could be used in document processing to construct sets of mathematical expressions which denote the same mathematical concept: n symbolic expressions would form a set if each of the possible C_n^2 pair-wise combinations fulfilled the above-mentioned conditions. Consider, for instance, the (unordered) pairs of simple mathematical expressions: (c, c_1), (c_2, c_1), and (c_2, c) which fulfill the criteria. They form a set $\{c, c_1, c_2\}$. Assuming that the expression c has been previously interpreted, for

instance, as a constant, the expressions c_1 and c_2 are likely to have the same mathematical interpretation.

3 Results

The results of the analyses are presented in Tables 1 through 3.

Table 1 shows the results of the symbol declaration study. The first column contains the categories. We present the absolute and percentage counts for the two subcategories of explicit declarations and for the location of the declaration (the occurrence number from the beginning of the document which is part of the symbol's declaration).

About 67% of simple mathematical expressions were explicitly introduced in the discourse. In most cases the first occurrence of a symbol is within a declaration, however, as can be seen from the fourth column, 'n (%)', in some rare cases the declaration does not come till the fourth and fifth occurrence. It appears that for this study, extracting the first five occurrences was a good choice, with only two out of 336 instances being declared as far as the fifth occurrence from the first mention. Moreover, in most cases symbols in declarations do not appear as part of a larger expression (only about 3% of occurrences were elaborated by means of a symbolic expresson). In 6% cases we encountered processing errors or were not able to distinguish how an object was declared.

Now, the results of the second study, Table 2, shows that about 72% of simple sub-terms of complex expressions, which were not recognized by LaTeXML have been explicitly introduced in the discourse. The declaration of most of these, again, appears together with the first occurrence of the expression, and, again unqualified declarations of these were more frequent. The remaining 27% of unknown symbols were not declared in the documents, so assigning them a role automatically based on the discourse context would perhaps require sophisticated inferences based on the context of the other occurrences.

Finally, Table 3 shows the results of semantic relatedness of locally occurring structurally similar expressions. Indeed, in most cases, 89%, structurally similar expressions which share the root identifier are also semantically related. We were unable to relate the expressions in 5% of the cases.

4 Discussion

The results of the study show that mathematicians do indeed tend to explicitly introduce object-denoting symbols which they use in their writings. While it is somewhat surprising that symbol declarations occur past the first mention of a symbol (that is, symbols are used before they have been introduced) overall, the context of the first mention accounts for the majority of symbol declarations.

The findings of the first and the second study also indicate that the global discourse context is a good starting point in an automated interpretation (and

Table 2. Results of the second study

Category		Occurrence	n (%)	N (%)
Explicitly declared	*unqualified*	1^{st}	331 (53.5%)	449 (72.5%)
		2^{nd}	22 (3.5%)	
		3^{rd}	23 (3.7%)	
		4^{th}	7 (1.1%)	
		5^{th}	20 (3.2%)	
	elaborated	–	46 (7.4%)	
Not explicitly declared				170 (27.5%)

Table 3. Results of the third study

Category	N (%)
Same concept	441 (88.9%)
Different concept	28 (5.6%)
Not classifed	27(5.4%)

disambiguation) of symbolic expressions in mathematical scientific documents. From a point of view of computational processing of mathematical discourse, this means that if the linguistic context in which a symbol appears can be parsed and interpreted (in particular, the first-mention context) then the intended usage of the symbol at hand, i.e. the symbol's meaning, can be recognized. Interpretation recovered in this way would, in turn, help complete the information in the (semantic) mark-up of mathematical expressions.

Now, the last study shows that the structural similarity of mathematical expressions and their disourse proximity can be exploited in propagating the interpretation of mathematical symbols. That is, assuming the a set of structurally similar expressions can be identified in a local discourse context and we can find the interpretation of one of them (for instance, using methods such as those proposed in [8]) then the interpretation of the related symbols can be with a large likelihood assumed to be the same. This can be seen as analogous to the "one sense per discourse" tendency in well-written prose (see [7]).

5 Conclusion and Further Work

In this paper, we presented the design and the results of three corpus-based studies on mathematical symbols in scientific papers, which were concerned with explicit declarations of symbols' denotations. The results of the studies empirically motivate methods of automated disambiguation of mathematical expressions based on the discourse context in which the symbols appear. While the data set we used was not large, the preliminary results we obtained are encouraging and suggest the need for comprehensive incremental interpretation as the methodology for semantic processing of mathematical documents. We are

planning to implement the results of the studies as part of a larger architecture for mathematical expression disambiguation.

We are planning a number of follow-up studies: A natural continuation of the presented experiments would be to investigate the way symbolic mathematical expressions are declared, from the linguistic point of view. That is, to study the languagage of symbol declarations in mathematical discourse. While a number of lexico-syntactic patterns for symbol declarations can be anticipated based on general familiarity with mathematical writing (the obvious being "`Let SYMBOL be` *a mathematical concept-denoting term*") given the size of the arXMLiv corpus we should be able to discover a variety of verbalizations.

Another natural follow-up direction which we are planning to pursue, is to look in more details into the set of symbols of which we have not found explicit declarations in the documents. Is there systematism to what symbols tend to be left unexplained, for their interpretation can be assumed as obvious? It is common knowledge that there are certain notational conventions in the usage of symbols, in mathematics in general and within sub-areas of mathematics (e.g. the use of mnemonics), can we automatically recognize these conventions based on corpus analysis focused on symbol declarations? Finally, aside from the knowledge of notational conventions, what other kinds of knowledge would be required to find automatically the interpretations of the remaining undeclared symbolic expressions in mathematical scientific documents?

Acknowledgments We would like to thank Deyan Ginev of Jacobs University Bremen without whose many preprocessing scripts it would not have been possible to conduct this study at ease. We would also like to thank the four anonymous reviewers for their helpful comments.

References

1. LaTeXML Manual. `http://dlmf.nist.gov/LaTeXML/manual/`, Retrieved June 2010.
2. `arXiv.org` e-Print archive, Retrieved June 2010. `http://www.arxiv.org`.
3. arXMLiv Project. `http://arxmliv.kwarc.info/`, Retrieved April 2010.
4. Ron Ausbrooks, Stephen Buswell David Carlisle, Giorgi Chavchanidze, Stéphane Dalmas, Stan Devitt, Angel Diaz, Sam Dooley, Roger Hunter, Patrick Ion, Michael Kohlhase, Azzeddine Lazrek, Paul Libbrecht, Bruce Miller, Robert Miner, Murray Sargent, Bruce Smith, Neil Soiffer, Robert Sutor, and Stephen Watt. Mathematical Markup Language (MathML) version 3.0. W3C Working Draft of 24. September 2009, World Wide Web Consortium, 2009.
5. Stephen Buswell, Olga Caprotti, David P. Carlisle, Michael C. Dewar, Marc Gaetano, and Michael Kohlhase. The Open Math standard, version 2.0. Technical report, The Open Math Society, 2004.
6. Herbert H. Clark. *Arenas of Language Use*. University Of Chicago Press, 1993.
7. William A. Gale, Kenneth W. Church, and David Yarowsky. One sense per discourse. In *Proceedings of the HLT-91 Workshop on Speech and Natural Language*, pages 233–237, 1992.

8. Mihai Grigore, Magdalena Wolska, and Michael Kohlhase. Towards context-based disambiguation of mathematical expressions. In *The Joint Conference of ASCM 2009 and MACIS 2009*, volume 22 of *Math-for-Industry, COE Lecture Note*, pages 262–271, 2009.
9. Donald Ervin Knuth, Tracy Larrabee, and Paul M. Roberts. *Mathematical writing*. The Mathematican Association of America, 1989.
10. Michael Kohlhase, Ştefan Anca, Constantin Jucovschi, Alberto González Palomo, and Ioan A. Şucan. MathWebSearch 0.4, A Semantic Search Engine for Mathematics. `http://search.mathweb.org/index.xhtml` (Retrieved April 2010), 2008.
11. Michael Kohlhase and Ioan Şucan. A search engine for mathematical formulae. In: Tetsuo Ida, Jacques Calmet, and Dongming Wang, editors, *Proceedings of Artificial Intelligence and Symbolic Computation, AISC 2006*, number 4120 in LNAI, pages 241–253. Springer Verlag, 2006.
12. Steven George Krantz. *A primer of mathematical writing*. The Americal Mathematical Society, 1997.
13. W3c math home. `http://www.w3.org/Math/`, Retrieved April 2010.
14. Bruce Miller. `LaTeXML`: A LaTeX to XML converter. Web Manual at `http://dlmf.nist.gov/LaTeXML/`, seen April 2010.
15. Heinrich Stamerjohanns, Michael Kohlhase, Deyan Ginev, Catalin David, and Bruce Miller. Transforming large collections of scientific publications to XML. *Mathematics in Computer Science*, 3(3):299–307, 2010.