Locating Mathematical Definitions in a Document

Akashdeep Bansal*[1], Pawan Kumar†[1], Volker Sorge‡[2], and M Balakrishnan§[1]

[1]Indian Institute of Technology Delhi, India
[2]University of Birmingham, United Kingdom

Abstract

At present STEM education beyond middle school is inaccessible to visually impaired students in countries like India. A few exceptions can be traced to enormous efforts of the parents in working with these students to make STEM text material accessible. Access to equations, tables, charts and figures are the key bottleneck. With the increased penetration of screen reading software, effective audio rendering of equations can significantly help in making many of the e-texts accessible. Unfortunately, linear syntactic rendering of equations can create considerable cognitive load even for relatively simple equations. In this work we plan to address an important aspect for making mathematical subjects accessible. It is related to contextual rendering of equations based on the subject of the document as well as local definitions. We are working on a classifier-cum-locator which can identify the mathematical definitions in a document. This will help in adapting audio rendering of equations based on their contextual semantics.

1 Introduction and Motivation

Despite significant advances in assistive technologies, a large section of visually impaired students, especially in developing countries like India, are not able to pursue STEM subjects in senior school years and college. Inaccessible STEM content including textbooks contribute in a major way to this challenge. Clearly when significant employment opportunities depend upon STEM education, this has a huge impact on their employability and integration into society.

Access to equations is a critical requirement for STEM. Audio rendering and tactile Braille are the two main modalities used by persons with visual impairment for accessing equations. Audio is the preferred modality due to ease and cost of production, possibility of digital dissemination as well as having no need for specially trained instructors to teach Braille/Tactile Graphics. MathJax [5, 6], MathPlayer [12, 8], ChattyInfty [9, 18], etc. are some of the common tools used for accessing equations using audio. All these solutions provide syntactical audio rendering with optimized use of lexical and prosodic cues. None of these solutions have the capability of adapting the rendering on the basis of the contextual semantics. For example, A^T can be "the transpose of the matrix A" or "A to the power T" depending on the context whether A is a matrix or a variable, respectively. Current systems will always read it as "A superscript T", which is a syntactical rendering.

Lack of contextual rendering adds cognitive load and leads to the steepness in the learning curve for users using audio rendering. Another key contributor is linearization of the long and complex equations to render it through a linear interface such as audio [3]. We are already working on a solution for effective delivery of long and complex equations [4]. To achieve this we first convert eBorn PDFs into ePub [14, 13] and based on a complexity metric decompose equations dynamically when rendering.

In this work, we are focusing on contextually improved voicing of equations by locating and interpreting mathematical definitions in documents. Mathematical definitions are those phrases that assign mathematical

*akashdeep@cse.iitd.ac.in
†pawan.kumar@cse.iitd.ac.in
‡v.sorge@cs.bham.ac.uk
§mbala@cse.iitd.ac.in

properties to symbols used in subsequent text. For example, "Let $F : R^n \to R^n$ be a C^1-vector field" defines F to be a C^1-vector field. Unfortunately, definitions are only rarely given in such obvious manner and meaning is often assigned to symbols in a much more subtle way. Consequently, previous research [19, 15, 2] has demonstrated that machine learning techniques are more helpful than simple pattern matching techniques. In our work we aim to train a classifier for definitions based on a large ground truth set. In order to construct the latter we have assembled an annotation tool that presents the concordance analysis of a corpus of pre-processed documents which can be manually annotated to extract valid mathematical definitions. In this abstract we present the basic principles of the analysis and the functioning of the annotation tool.

2 Semantic Context Analysis

The true meaning of mathematical symbols can only rarely be deduced by their occurrence in a single formula alone. And although many commonly used mathematical symbols often have widely understood semantics, their meaning in context of a particular mathematical text might differ considerably. For example, while in non-associative algebra structures like rings use addition and multiplication symbols, their meaning must not be confused with their counter parts in traditional arithmetic.

Lack of contextual semantic leads to two types of ambiguities —

1. **Lexical ambiguity** — A symbol can have different meaning on the basis of the context. e.g., δ can be mapped to KroneckerDelta, DiracDelta, DiscreteDelta, or δ.

2. **Characteristical ambiguity** -– An expression can have different interpretation based on the characteristics of a symbol, even after having the same lexical meaning. e.g., w^{-1} can be an inverse function or $1/w$ depending on whether the symbol w is a function or a variable, respectively. Here, even though the lexical meaning of w and -1 is same in both the scenarios. Still, the expression w^{-1} has a completely different interpretation based on whether w is a function or a variable. In the former case,w^{-1} means the inverse function, in the latter w^{-1} means "1 divided by w

We can roughly divide the sources from which a given symbol or expression can get its definition in a document into three scenarios:

- **Scenario 1**: The symbol is never defined within the document, but is well understood in general or in the subject domain the document belongs (e.g., the equality sign or the normal subgroup sign in Group theory). Note here there is some implicit assumption about the audio rendering system already having default meaning of symbols for every domain.

- **Scenario 2**: The symbol is defined once within the document and is expected to carry that meaning throughout the document.

- **Scenario 3**: The same symbol is defined multiple times within the document and is expected to carry that meaning only within the scope of those definitions. In this case, it is also necessary to determine the scope of each definition and thus tailor the rendering accordingly.

Scenario (1) requires the identification of subjects domains having different semantics and creation of default symbol mappings for each domain. To some extent, this exists for certain topics, e.g., Wikipage [1] provides a detailed list of definitions of mathematical symbols in a small number of different subjects. While reading a document dealing with probabilities, reading $P(A)$ as "Probability of event A" is critical for understanding the equation instead of "P OpenParen A CloseParen". One could naively enforce an interpretation of P as probability or allow for explicit selection of the domain. But even if the subject domain is explicitly known to be probability theory, a default interpretation for P as probability might not necessarily be correct, as P could for example also denote a polynomial or a partition. Therefore, a dynamic approach should be more robust.

In our work, we are therefore focusing on the semantic information, which can be learned from the document itself; no prior domain knowledge is required. Hence, this work covers the scenarios (2) and (3) in the above list.

In the past, there have been few attempts in this area: it has been shown previously [17, 19, 10, 15, 2] that the consideration of surrounding text can relatively improve the performance of semantic disambiguation in comparison with a mere expression analysis [16, 7, 11]. But it has also been demonstrated [19] that machine learning based approaches can provide better results in comparison to simple pattern matching based methods. We therefore follow a machine learning approach for inferring mathematical definitions automatically from the document using a classifier-cum-locator. In particular we aim to extend work from [2] to train a classifier on the basis of a large ground truth set of definitions obtained from a concordance analysis, following a methodology that is outlined by these steps:

- **Identifying mathematical symbols definitions**: This requires the identification of the concordances which define the definitions of the mathematical symbols in the document. We achieved it by creating an efficient annotation tool (described in section 4). This is helping us in effectively annotating a large corpus to create the ground truth.
- **Classifier**: Developing a machine learning based classification algorithm. This will be trained on the ground truth created using the above annotation tool. This will help us in identifying whether the given concordance contains the mathematical definition or not. This is described in section 6.
- **Identifying the scope of these definitions**: To distinguish between the above mentioned scenarios (2) and (3), it needs to be inferred whether the given definition's scope is local or global. We plan to achieve it by checking whether a particular expression/symbol has multiple different definitions in the given document and associating the occurrence of that particular expression/symbol to an immediately preceding definition. By default, the expression/symbol would be mapped to the lexical meaning. After working on the scenario (1), it can be mapped to the default meaning associated with the subject.

3 Concordance Analysis

Our machine learning approach is based on a concordance analysis, that is on a enumeration of all expressions in question together with the context in which they occur. In our case, we are working with mathematical expressions and as context we take up to five words or expressions before and after the principal expression.

Table 1 contains a number of examples of concordances, where the principal mathematical expression is marked in red. Note, that single mathematical expressions are counted as a single element of the concordance, regardless of the number of symbols they contain. Moreover, while the principal is usually in the center, this can be broken up by paragraph making concordance pre- or postfixes shorter than five expressions/words.

Obviously not every concordance constitutes a definition of a mathematical expression. Table 1 contains examples of both together with an explanation why or why not a concordance can be seen as a definition. As a consequence it is necessary to build a ground truth by manually tagging of concordances as to whether they contain a definition or not.

4 Annotation Tool and Ground Truth

To generate the ground truth we need to annotate the concordances manually by several independent annotators. Although there are a number of tools for concordance analysis and corpus annotations available, none can adequately deal with documents containing considerable amount of mathematics. We have consequently developed an annotation tool with the objective to provide an efficient way for annotating the concordances having and not having a valid mathematical definition.

For this, we took around 100 XML documents from various domains of STEM. In these documents, the tool first identifies the mathematical entities (maybe a symbol or expression). As in general on the web, we found documents which are not appropriately tagged. Hence, the tool searches for various tags such as MathML, bold, and italic, and also look for the LaTeX expressions. The tool searches for bold and italic because many of the times authors tagged the mathematical entities by bold/italic to give the same visual appearance. The average number of entities identified in the documents are 1052.644, having max of 3046 and min of 252 and standard deviation of 586.2051.

Concordance	**Definition (Yes/No)**	**Explanation**
We present, in dimension $n \geq 2$, a survey of samples to:	Yes	n is a dimension.
ensuring that an equilibrium point x^* is a local attractor is	Yes	x^* is a point.
$F : R^n \to R^n$ satisfying that for any $x \in R^n$, all the eigenvalues of $JF(x)$	Yes	x represents eigenvalues.
$n < 2$ and non-injective polynomial maps $f : R^n \to R^n$ with $[0, \infty) \cap Spec(Jf(x)) = 0$, for all $x \in R^n$".	Yes	f is a non-injective polynomial map.
that the C^1-vector fields $F : R^n \to R^n$ satisfying that for any $x \in R^n$	Yes	F is a C^1-vector field.
Let us recall that the C^1-vector fields $F : R^n \to R^n$ satisfying that for any $x \in R^n$	No	This is not a definition of C^1.
$JF(x)$, the Jacobian matrix of F at x, has negative real	No	The definition of F is not given; even though you can infer that F is a function, this is not explicit in this concordance.
restricted to the invariant plane $z = 0$ is a center. Moreover, perturbing	No	Even though $z = 0$ is a plane is defined in this concordance. Still, to infer that z is a dimension, you requires the prior domain knowledge. That's why it is not a valid mathematical definition. Here, we are only looking for the concordances which has the proper definition of the symbol within the concordance.
construct polynomial maps $F = \lambda I + H : R^3 \to R^3$ with JH nilpotent, such that the WMYC	No	This is not a definition of JH.
the MYC is true when $n \leq 2$ and false when $n \geq 3$ (see	No	This is not a definition of n.

Table 1: Some sample concordances containing and not containing valid mathematical definitions. The principal mathematical expression is marked with red color in above concordances.

Corresponding to each identified entity, five previous and five next words are highlighted (as shown in Figure 1). This whole highlighted part, including the central mathematical entity is referred to as concordance in this paper. The person who is annotating the concordances needs to press $\rightarrow$ (right arrow) or $\leftarrow$ (left arrow) to mark the highlighted concordance as a valid or not a valid mathematical definition, respectively. To provide the assurance, on pressing the right-arrow/left-arrow the background colour gets changed to blue or red, respectively (as shown in Figure 2 and 3, respectively.). The tool also provides an option to mark error in concordance identification/highlighting by pressing the key "e" (Figure 4). The navigation functionality was provided through up/down arrow keys to go to the previous/next concordance, respectively. The tool also automatically goes to the next concordance after 1 second of marking the currently highlighted concordance.

This tool allows us to create a large amount of ground truth efficiently. The average annotation speed observed is that a person can annotate 400 concordances within 30 minutes. Also, the same file is annotated by multiple people having a strong background in STEM, to remove any personal bias.

5 Data Collection

The present analysis is based on nine files, all of them have been annotated by three different annotators. The same file was annotated by multiple annotators to avoid the individual biases. Final label is decided based on the max approach, detailed rules are mentioned in table 2. e.g., If two annotators have given

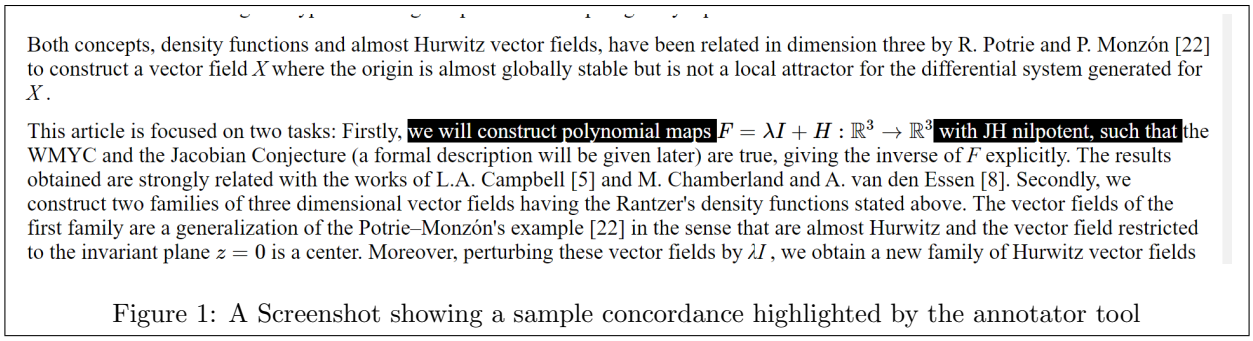

Figure 1: A Screenshot showing a sample concordance highlighted by the annotator tool

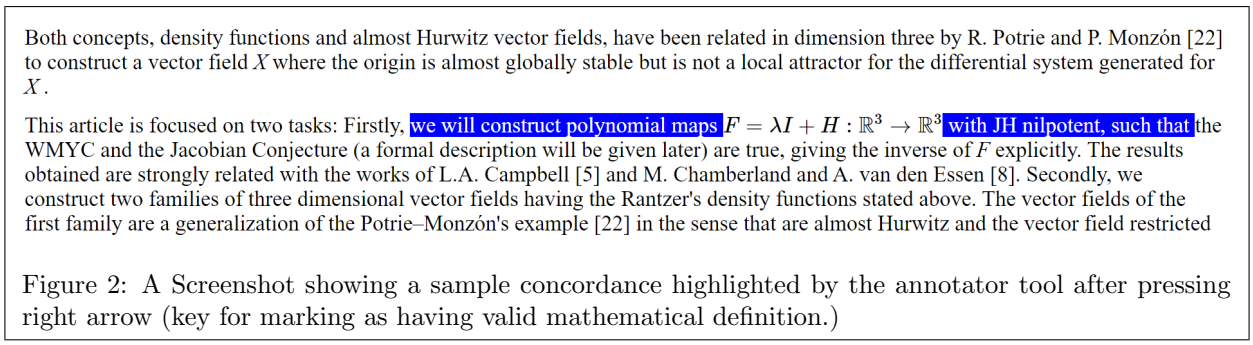

Figure 2: A Screenshot showing a sample concordance highlighted by the annotator tool after pressing right arrow (key for marking as having valid mathematical definition.)

label 1, whereas, one annotator has given the label 0, then the final label will be 1. Here, 0 represents "doesn't contain a valid mathematical definition", 1 represents "contains a valid mathematical definition", and -1 represents "an error". The average number of concordances in these files are 619. The total number of concordances having final label as 0 and 1 are mentioned in table 3. We have split this collection into training and test set in 80:20 ratio.

Label 1	**Label 2**	**Label 3**	**Final Label**
0	0	0	0
0	0	1	0
0	1	1	1
1	1	1	1
-1	0	0	0
-1	0	1	-1
-1	1	1	1
-1	-1	0	-1
-1	-1	1	-1

Table 2: Rules for final label based on the labels from different annotators. Here, 0 represents "doesn't contain a valid mathematical definition", 1 represents "contains a valid mathematical definition", and -1 represents "an error".

Label	**Total Number**
1	892
0	4199

Table 3: The total number of concordances having final label as 0 and 1.

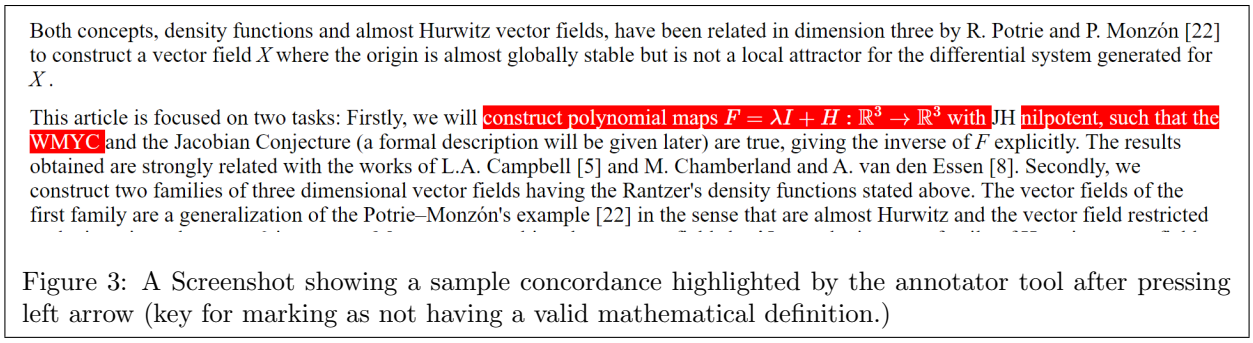

Figure 3: A Screenshot showing a sample concordance highlighted by the annotator tool after pressing left arrow (key for marking as not having a valid mathematical definition.)

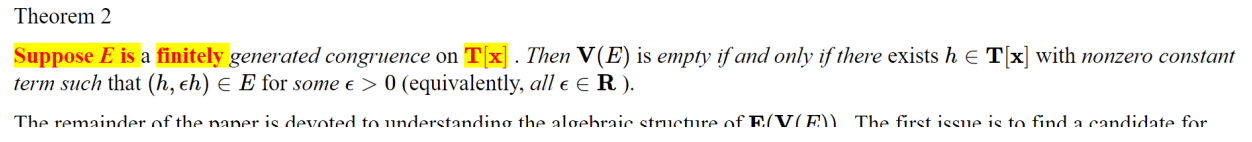

Figure 4: A Screenshot showing a concordance for an erroneously identified element.

6 Classifier

As a baseline classifier, we have considered Naive Bayes and SVM. The concordance pre-processing requires removing punctuations and stopwords, using NLTK library, and replacing mathematical symbols/expressions with keyword "MATH". Table 4 shows some sample concordances after replacing mathematical symbols/expressions by the keyword "MATH". Each concordance is vectorized using term frequency and inverse document frequency, abbreviated as tf-dif. We have used uni-grams as well as bi-grams and taken logarithm of their term-frequency. The SVM, from library sickit-learn, is trained with early-stopping and uses 20% of the training data as validation set. We have used the same tf-idf based feature vector for SVM as well. As evident from the table 5 and 6, SVM provides better f1-score and makes for a better choice for a baseline classifier. In Future experiment on classifier, we will try to bring in semantics of the context window using deep-learning based models to learn embeddings for the concordances that might help improve the accuracy and f1-score.

7 Conclusion and Future Work

The developed annotation tool is found to be very useful in quickly annotating the files for generating ground-truth. Even with limited annotations, using a classifier we are able to achieve an accuracy of 85% and f1-score of 0.70. SVM with early-stopping is found to be a better choice for baseline classifier.

Further, We are exploring the possibility of more meaningful replacement of mathematical symbols/expressions for tokenization. We also want to explore how we can use these definitions for semantic disambiguation of the mathematical symbols. We are also analyzing whether the current choice of five previous and next words is "appropriate and effective" or it needs to be changed. Finally, we want to identify each mathematical definition's scope to determine whether this is a local or global definition.

Acknowledgement

We would like to thank Mark Lee for his valuable inputs. This project is funded under the SPARC (Scheme for Promotion of Academic Research Collaboration), Ministry of Education, Govt. of India. MathJax work was supported in part by Simons Foundation Grant, No.514521.

Concordance	Tokens
We present, in dimension $n \geq 2$, a survey of samples to:	We present, in dimension MATH , a survey of samples to:
ensuring that an equilibrium point x^* is a local attractor is	ensuring that an equilibrium point MATH is a local attractor is
$F : R^n \to R^n$ satisfying that for any $x \in R^n$, all the eigenvalues of $JF(x)$	MATH satisfying that for any MATH , all the eigenvalues of MATH
$n < 2$ and non-injective polynomial maps $f : R^n \to R^n$ with $[0, \infty) \cap Spec(Jf(x)) = 0$, for all $x \in R^n$".	MATH and non-injective polynomial maps MATH with MATH, for all MATH".
that the C^1-vector fields $F : R^n \to R^n$ satisfying that for any $x \in R^n$	that the MATH -vector fields MATH satisfying that for any MATH
Let us recall that the C^1-vector fields $F : R^n \to R^n$ satisfying that for any $x \in R^n$	Let us recall that the MATH -vector fields MATH satisfying that for any MATH
$JF(x)$, the Jacobian matrix of F at x, has negative real	MATH, the Jacobian matrix of MATH at MATH, has negative real
restricted to the invariant plane $z = 0$ is a center. Moreover, perturbing	restricted to the invariant plane MATH is a center. Moreover, perturbing
construct polynomial maps $F = \lambda I + H : R^3 \to R^3$ with JH nilpotent, such that the WMYC	construct polynomial maps MATH with MATH nilpotent, such that the WMYC
the MYC is true when $n \leq 2$ and false when $n \geq 3$ (see	the MYC is true when MATH and false when MATH (see

Table 4: Tokenized form of sample concordances after replacing mathematical symbols/expressions with the keyword "MATH".

	Naive Bayes	**SVM**
Accuracy	0.841	0.852
f1-score	0.652	**0.702**

Table 5: Accuracy and f1-score of base classifier based on the Naive Bayes and SVM.

	Naive Bayes		**SVM**	
	1	**0**	**1**	**0**
1	53	126	72	107
0	36	804	43	797

Table 6: Confusion Matrices

References

[1] List of mathematical symbols by subject. `https://en.wikipedia.org/wiki/List_of_mathematical_symbols_by_subject`, 2020.

[2] R. Almomen. *Context classification for improved semantic understanding of mathematical formulae*. PhD thesis, University of Birmingham, 2018.

[3] A. Bansal, M. Balakrishnan, and V. Sorge. Comprehensive accessibility of equations by visually impaired. *ACM SIGACCESS Accessibility and Computing*, (126):1–1, 2020.

[4] A. Bansal, M. Balakrishnan, and V. Sorge. Evaluating cognitive complexity of algebraic equations. In *Journal on Technology & Persons with Disabilities*, page (to appear). CSUN, 2021.

[5] D. Cervone, P. Krautzberger, and V. Sorge. New accessibility features in mathjax. In *31th Annual International Technology and Persons with Disabilities Conference Scientific/Research Proceedings*. California State University, Northridge., 2016.

[6] D. Cervone, P. Krautzberger, and V. Sorge. Towards universal rendering in mathjax. In *Proceedings of the 13th Web for All Conference*, page 4. ACM, 2016.

[7] I. A. Doush, F. Alkhateeb, and E. Al Maghayreh. Towards meaningful mathematical expressions in e-learning. In *Proceedings of the 1st International Conference on Intelligent Semantic Web-Services and Applications*, pages 1–5, 2010.

[8] L. Frankel, B. Brownstein, and N. Soiffer. Navigable, customizable tts for algebra. In *28th Annual International Technology and Persons with Disabilities Conference Scientific/Research Proceedings*. California State University, Northridge, 2014.

[9] T. Kanahori and M. Suzuki. Scientific pdf document reader with simple interface for visually impaired people. In *International Conference on Computers for Handicapped Persons*, pages 48–52. Springer, 2006.

[10] M.-Q. Nghiem, G. Y. Kristianto, G. Topić, and A. Aizawa. A hybrid approach for semantic enrichment of mathml mathematical expressions. In *International Conference on Intelligent Computer Mathematics*, pages 278–287. Springer, 2013.

[11] M.-Q. Nghiem, G. Yoko, Y. Matsubayashi, and A. Aizawa. Towards mathematical expression understanding.

[12] N. Soiffer. Mathplayer v2.1: web-based math accessibility. In *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, pages 257–258. ACM, 2007.

[13] V. Sorge, M. Balakrishnan, N. Jadhav, A. Bansal, and H. Garg. Rule based mathematics extraction and layout analysis of pdf documents. In *Proceedings of the 4th International Workshop on Digitization and E-Inclusion in Mathematics and Science (DEIMS)*, 2021.

[14] V. Sorge, A. Bansal, N. M. Jadhav, H. Garg, A. Verma, and M. Balakrishnan. Towards generating web-accessible stem documents from pdf. In *Proceedings of the 17th International Web for All Conference*, pages 1–5, 2020.

[15] Y. Stathopoulos and S. Teufel. Mathematical information retrieval based on type embeddings and query expansion. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2344–2355, 2016.

[16] J. Stuber and M. Van den Brand. Extracting mathematical semantics from LaTeX documents. In *International Workshop on Principles and Practice of Semantic Web Reasoning*, pages 160–173. Springer, 2003.

[17] M. Wolska and M. Grigore. Symbol declarations in mathematical writing. 2010.

[18] K. Yamaguchi, T. Komada, F. Kawane, and M. Suzuki. New features in math accessibility with infty software. In *International Conference on Computers for Handicapped Persons*, pages 892–899. Springer, 2008.

[19] K. Yokoi, M.-Q. Nghiem, Y. Matsubayashi, and A. Aizawa. Contextual analysis of mathematical expressions for advanced mathematical search. *Polibits*, (43):81–86, 2011.