MioGatto による数式グラウンディングデータセットの構築

朝倉卓人
東京大学
takuto@is.s.u-tokyo.ac.jp

宮尾祐介
東京大学
yusuke@is.s.u-tokyo.ac.jp

相澤彰子
国立情報学研究所
aizawa@nii.ac.jp

概要

計算機を用いた科学技術文書の読解には，数式内の 1 つ 1 つの記号の意味をグラウンディングできることが重要である．現実の科学技術文書においては数式内の記号の意味は必ずしも一定ではなく，同じ記号が複数の意味で用いられるため，グラウンディングの際には記号間の共参照関係も明らかにする必要がある．本研究では arXiv.org から選んだ 15 本の科学論文を対象に，数式内の記号の共参照関係を明示的にアノテーションしたコーパスの構築を行った．その結果，数式内の記号の曖昧性は狭い範囲に絞っても存在するほど複雑だが，共参照関係は高いアノテータ間一致率を保ちながらラベル付け可能であることが示された．

1 はじめに

論文や専門書などの科学技術文書に書かれた知識を，検索・数式処理・定理証明支援などの計算機を応用した諸技術を用いて十分に活用するためには，文書中の自然言語テキストのみならず数式についても解析を行うことが必須である．そして，自然言語中の数式解析を行うには，数式内に現れる 1 つ 1 つの文字や記号（トークン）がどのような意味で用いられているのかを明らかにする必要がある．著者らはこれまでに，この部分の処理を数式グラウンディングとして定式化・提案してきた [3, 4]．このタスクは各数式トークンに対してそれぞれ文脈に応じた説明を付与する説明アライメントタスク（図 1）と，まったく同じ意味で用いられているトークンとそうでないトークンを判別する共参照解析タスクの 2 つの性質を併せ持つ．

計算機による科学技術文書理解の達成には，数式グラウンディングの自動化が必要である．自動化手法の開発を目指し，著者らはまず実際にグラウンディングされたデータの観察・分析・学習・評価のため人の手によるグラウンディング結果のアノテーションを付したコーパス構築に取り組んだ．一般に共参照情報のアノテーションは特に高コストであるため [12] 著者らは専用のアノテーションツール MioGatto[1] を開発してデータ構築プロセスを効率化した [5]．その上で，実際に MioGatto を用いて 11 名の学生アノテータとともに計 15 本の科学論文について，それらの論文中の数式のすべての数式識別子出現に対するアノテーションを行った．

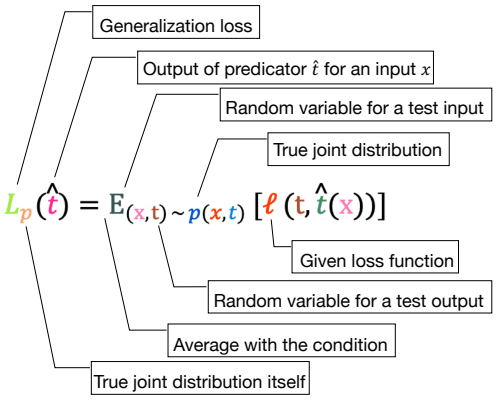

図 1 説明アライメント

本稿では，この数式グラウンディングデータセットのこれまでの構築手順を紹介するとともに，得られたアノテーション付きコーパスの概要と分析結果について報告する．構築したデータセットは SIGMathLing リポジトリ[2] にて配布している．

2 関連研究

文書中の数式理解に資することを目的とするデータセットは，複数提案されてきた．arXMLiv データセット [10] はプリプリントサーバ arXiv.org[3] に収録されている 150 万本以上の科学論文を，様々な研究用途の計算機プログラムで扱いやすいよう LaTeXML [11] を用いて XHTML 文書に変換した巨大な文書コーパスである．文書中の数式部分は

1) https://github.com/wtsnjp/MioGatto
2) https://sigmathling.kwarc.info/resources/grounding-dataset/
3) https://arxiv.org

LaTeXML の機能により機械的にプレゼンテーション MathML [6] に変換されているが，基本的に通常の LaTeX マークアップに含まれるのと同じ「見た目」に関する情報がエンコードされているに過ぎず，特にそれ以上に踏み込んだ情報は付加されてはいない．それでも数式を含む文書の貴重な言語資源として広く活用されており，MioGatto の入力形式もこの arXMLiv の XTHML の仕様 [9] に準じている．

科学技術文書の数式部分について，各トークンに説明を付与したアノテーション付きコーパスはいくつか提案されている．NTCIR-10 では Math Pilot 共通タスクの一部として自然言語テキスト内にあるトークンの定義を抽出する Math Understanding サブタスクが提案され，その開発・評価のために arXMLiv データセットに含まれる XHTML 文書内の数式に人手で説明が付与されたデータセットが提供された [1]．また類似のタスクは数式内の 1 つ 1 つの識別子に対して説明を付与する MathAlign タスクとしても定式化されており，同じく arXiv.org に収録されている 116 本の論文の中から，合わせて 584 の識別子に説明を与えたデータセットも存在する [2]．

ある程度以上の長さのある現実の科学技術文書では，しばしば数式トークンは 1 つの文書内においても複数の意味で用いられる [3, 4]．例えば機械学習分野の教科書 PRML [7] の第 1 章では，太字の $\mathbf{y}$ は同じ章の中で少なくとも 4 つの意味で用いられている（表 1）．そのため文書中の数式理解には，こうした同一文書内のトークン間の共参照関係を明らかにすることが必要だが，数式トークンの共参照関係が明示的にラベル付けされたデータセットは知られていない．本研究では，本文にある程度の長さがあり数式も豊富なものを中心として 15 本の科学論文を選定し，それらに現れる合計 12,352 個の数式識別子出現すべてについて，各論文内での共参照関係を明示的にアノテーションしたデータセットを構築した．

表 1　PRML [7] 第 1 章における $\mathbf{y}$ の曖昧性

本文のテキスト	$\mathbf{y}$ の意味
…得られるのは関数 $\mathbf{y}(\mathbf{x})$ である…	画像を入力とする関数
…出力ベクトル $\mathbf{y}$ が出力される…	$\mathbf{y}(\mathbf{x})$ の出力ベクトル
2 つの確率変数ベクトル $\mathbf{x}$ と $\mathbf{y}$ に…	確率変数ベクトル
…同時分布 $p(\mathbf{x}, \mathbf{y})$ を考えよう．	$\mathbf{x}$ に対応する値

3　アノテーションの目的と方法

数式グラウンディングの自動化には，手法の構築と評価のためにデータセットが必須である．統計モデルを構築して自動化を実現するには一般に大量の教師データが必要になる．最初はルールベースでデータを増やす場合にも，ルールの検討には実際の文書における数式トークンの使われ方を観察する必要があり，やはりある程度のデータが要ることに変わりはない．またいずれにしても評価用のデータは必要である．本研究では，グラウンディング自動化の第一歩として実際の科学論文を対象に次の 2 種類の情報をすべて人手でアノテーションした（図 2）．

数学概念　数式中のトークンが参照する概念．実際のアノテーションデータとしては，単純な説明 (description) に加えて，数学的な型やアリティ，制約条件などの付加的な属性を付与することができる．

グラウンディング情報源　人間がグラウンディングを行う際に，その根拠として利用できるテキストスパン．数学的に定義や宣言にあたる箇所が該当する．例えば図 2 の最初の f は実数関数にグラウンディングされるが，その根拠となる情報源は直前の “a real-valued function” である．

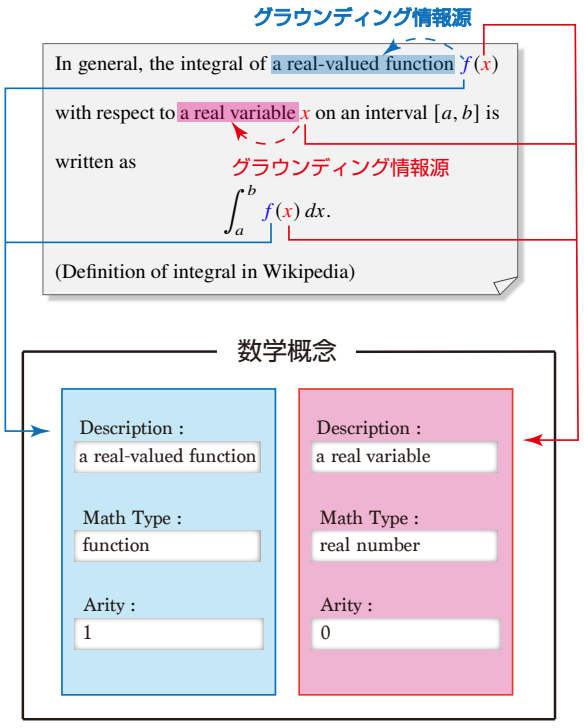

図 2　**アノテーションした情報**：数式中の各トークンをそれぞれの参照する数学概念に紐付ける．その際の根拠となるテキストスパンをグラウンディング情報源と呼ぶ．例文テキストは英語版 Wikipedia[4] より．

4）　https://en.wikipedia.org/wiki/Integral

各数式トークンの出現に直接説明をアノテーションする代わりに，文書ごとにアノテータが作成する数学概念の一覧（**数学概念辞書**と呼ぶ）に定義される概念 ID をラベル付けすることで，共参照関係を明示したデータセットを構築することができる．すなわち，同じ概念 ID と紐づく出現同士は共参照関係にある一方で，異なる ID と紐づくものは共参照関係にない．こうしたアノテーションを実現するために，実際のアノテーションにおいては著者らが開発した専用のアノテーションツール MioGatto を用いた（図 3）．MioGatto は (1) 数学概念辞書の作成，(2) 各出現への数学概念 ID の紐付け，(3) 各出現に対応するグラウンディング情報源のスパン位置のアノテーションを GUI 操作のみで素早く行うことができるように設計されている [5]．

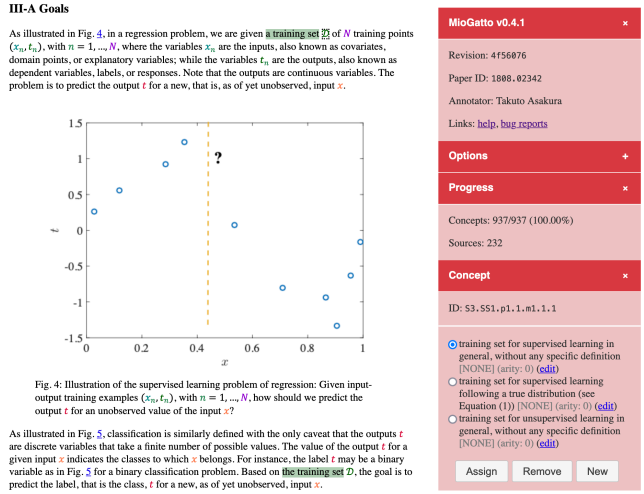

図 3　MioGatto の操作画面．画面の左側にアノテーション対象の論文の本文があり，右側のサイドバーに MioGatto の提供する情報やアノテーション操作に必要なボタン類が配置されている．

データセット構築にあたっては，arXiv.org に LaTeX 文書ソース付きで収録されている英文論文の中から，ある程度以上の数式が用いられているものを選んでアノテーションを行った．こうした英文で数式を含む科学論文を正しく読解してアノテーションを行うには，それぞれの論文の分野について専門的な知識を要することから，様々な分野の専門知識を持つ学生（主に大学院生と学部生）の中から協力者を募集した（付録 A）．参加アノテータにそれぞれの背景知識・興味に合った論文を arXiv.org 収録論文の中から選定してもらい，選ばれた論文を MioGatto を用いてアノテーションするための前処理を行った．この前処理には論文著者によるオリジナルの LaTeX 文書を LaTeXML で XML に変換する作業と，著者の誤った数式マークアップを修正する作業（付録 B）が含まれる．各アノテータには MioGatto の使用法を説明するとともに，実際のアノテーションに必要な論文 XHTML データ，アノテーションデータの雛形，手順書[5] を提供し，同手順書にしたがってアノテーション作業を実施してもらった．アノテーションの結果得られたデータは著者らが内容を確認の上，その後の分析を行った．

今回アノテーションの対象としたのは，選定した論文内で用いられているすべての数式の，すべての数式識別子の出現である．ここで数式識別子とは数式トークンの一種で，変数・関数・定数を表す単一の文字（x や θ）または短い名前（sin など）のことである．数式内には識別子以外にも演算子（+ など）や数値といったトークンもあるが，アノテーション対象が多くなり過ぎることを防ぐため今回は識別子に対象を絞っている．またグラウンディング情報源については，論文内でアノテータが発見できたものすべてをアノテーションした．1 つの概念に紐付く情報源が論文内に複数箇所ある場合や，ある概念に紐付く情報源が 1 つもない場合もあるため，ラベル付けする情報源の数には制約を設けなかった．

4　構築済みデータセットの分析

NLP・数理論理学・代数学・天文学などの分野の科学論文合わせて 15 本について，すべての数式識別子出現に対する手動アノテーションを完了した（表 2）．データセット全体では合計 12,354 個の識別子出現があり，そのすべてに数学概念が割り当てられた．またグラウンディング情報源にあたるテキストスパンは計 938 個収集できた．各論文について識別子の「種類」数を，対応する概念辞書の「概念」数で割ると，識別子の種類ごとに平均いくつの意味で用いられているかが算出できるが，その平均値は全体を通して 2.09 であった．また識別子の種類ごとに使われる回数が異なるので，出現数を考慮して加重平均したものが表 2 の「平均候補」数である．これは実際のアノテーション時に，各出現に対して数学概念を割り当てる際にアノテータが目にした選択肢の数の平均値に相当する．そのため「平均候補」の数が大きいほど識別子の曖昧性の度合いが高く，アノテーションの難易度も高かったと解釈できる．

5)　https://github.com/wtsnjp/MioGatto/wiki/Annotator's-Guide

表 2　アノテーション結果.「番号」は説明のための便宜的な論文 ID,「単語」は本文の単語数,「種類」は識別子の種類数,「出現」は識別子の出現数,「概念」は概念辞書に登録された項目数,「平均候補」は辞書項目数の識別子出現回数に応じた加重平均,「情報源」はグラウンディング情報源の数．具体的な文献情報は付録 A に掲載する．

番号	単語	種類	出現	概念	平均候補	情報源
1	10976	40	937	104	6.4	232
2	4267	42	266	73	2.6	30
3	3563	38	433	79	2.5	34
4	3567	46	1648	64	1.9	30
5	13154	141	4629	424	5.2	180
6	2881	25	162	30	2.7	12
7	5543	31	203	47	2.6	36
8	4613	23	217	27	1.1	28
9	6255	34	510	74	2.7	27
10	5415	73	1175	167	3.3	60
11	4451	33	237	61	2.9	34
12	4261	31	186	39	1.7	25
13	2257	23	124	27	1.2	18
14	10032	59	1064	129	4.2	97
15	4863	41	561	73	2.3	95
合計	86098	680	**12352**	1418	—	**938**

4.1　アノテータ間一致率

本研究でアノテーション対象としたのは専門性の高い科学論文であるので，同一の文書に対してアノテーション可能な人員を複数名確保するのは容易でない．しかしアノテーションの正確性・再現性を確認するため，論文 1 については計 5 名のアノテータで独立にアノテーションを実施し，アノテータ間一致率の算出を行った（表 3）．概念辞書の作成はアノテータ A が担い，他はその辞書を用いて概念の割当や情報源のアノテーションを行った．アノテータにより作業精度に多少ばらつきがあるが，数学概念の一致率・Cohen の κ 値 [8] を算出すると十分に高い値となっている．またグラウンディング情報源も高い頻度で重複しており，人間が情報源とみなすテキストスパンはよく一致することがわかった．

4.2　スコープの切替位置と情報源

ある数式識別子の出現に対して割り当てられた数学概念が，同じ識別子種の直前の出現に割り当てられた概念と異なる場合，その 2 つの識別子出現の間に**スコープの切替**があると言う．数式グラウンディングを行うには，1 つの文書内のすべてのスコープの切替位置を特定する必要があり，これが自動化において最も挑戦的な部分である．今回構築したデータセットでは，論文 15 本の中で合計 2,378 回のスコープ切替があった．このうち 1 つのセクションの中でスコープ切替が起こっているものは 2,129 回 (89.5%) であったことから，文書内の単一のセクションという狭い範囲に絞ってみても数式識別子には曖昧性があることがわかった．

表 3　アノテータの役割と一致率. 上 3 行に各アノテータの役割，中 2 行に数学概念の一致率，下 2 行に情報源の数を示す．一致率・重複率は対アノテータ A の値である．

アノテータ	A	B	C	D	E
辞書作成	✓				
概念割当	✓	✓	✓	✓	✓
情報源	✓			✓	✓
一致率 (%)	—	96.5	87.4	92.1	84.2
κ 値[6]	—	0.94	0.80	0.87	0.75
情報源の数	232	—	—	249	257
└ 重複率 (%)	—	—	—	80.3	93.4

またグラウンディング情報源とされたテキストスパンと，それぞれと紐付けられている数式識別子の出現との位置関係についても分析を行った．アノテーションされた計 938 個のグラウンディング情報源のうち 76.5%にあたる 718 個が数式識別子よりも先行する位置に存在していた．また各情報源とそれと紐付く数式識別子出現（最も距離が近いもの）の距離を，間に挟まる単語の数で計数すると，平均して 14.7 単語であった．ただし，情報源・識別子間の距離はばらつきが大きく，距離の中央値はすべての論文で 0〜4 単語の範囲にあり，典型的な情報源は対応する識別子出現の直前数単語以内に存在する．

5　結論と今後の展望

本研究では多様な分野の科学論文 15 本について，グラウンディング情報を人手でアノテーションしたデータセットを構築した．その中ではすべての識別子出現に説明といくつかの付加情報がラベル付けされるとともに，各論文内で識別子間の共参照関係が明示的になっている．さらにこうしたデータセットは必ずしも言語資源構築を専門としないアノテータによる作業でも構築可能であることを示した．

今後は概念辞書の作成のみを人手で行い，論文中の識別子出現に対してその概念辞書のうちの適当な項目を自動で割り当てる半自動化を行う．これにより効率よく提案データセットを量的に拡張し，ひいてはグラウンディング全体の自動化を達成する．

6)　識別子の種類ごとの出現頻度に応じた加重平均（参考値）．

謝辞

本研究は JST ACT-X (JPMJAX2002) の支援を受けて実施しました．日頃より実りある議論をしている Michael Kohlhase 教授と André Greiner-Petter 氏に感謝します．また MioGatto の改善を手伝ってくださった石井太河氏と，本研究に参加いただいたすべてのアノテータの方々にも感謝します．

参考文献

[1] Akiko Aizawa, Michael Kohlhase, Iadh Ounis. “NTCIR-10 Math Pilot Task Overview.” *Proceedings of NTCIR*-10. 2013.

[2] Maria Alexeeva, Rebecca Sharp, Marco A. Valenzuela-Escárcega and Kadowaki, Jennifer Kadowaki, Adarsh Pyarelal, Clayton Morrison. “MathAlign: Linking Formula Identifiers to their Contextual Natural Language Descriptions”. *Proceedings of the* 12*th Language Resources and Evaluation Conference* (LREC 2020).

[3] 朝倉卓人，André Greiner-Petter，相澤彰子，宮尾祐介．『数学概念への数式グラウンディングのためのデータセット』．言語処理学会第 26 回年次大会予稿集 (NLP2020).

[4] Takuto Asakura, André Greiner-Petter, Akiko Aizawa, Yusuke Miyao. “Towards Grounding of Formulae.” In *Proceedings of First Workshop on Scholarly Document Processing* (SDP 2020).

[5] Takuto Asakura, Yusuke Miyao, Akiko Aizawa, Michael Kohlhase. “MioGatto: A Math Identifier-oriented Grounding Annotation Tool.” In 13*th MathUI Workshop at* 14*th Conference on Intelligent Computer Mathematics* (MathUI 2021).

[6] Ron Ausbrooks et al. *Mathematical Markup Language* (*MathML*) 3.0 *Specification*. 2014. https://www.w3.org/TR/MathML3/.

[7] Christopher M Bishop. *Pattern Recognition and Machine Learning*. 2006.

[8] Jacob Cohen. “A coefficient of agreement for nominal scales.” *Educational and Psychological Measurement*. 1960.

[9] Deyan Ginev, Heinrich Stamerjohanns, Bruce R. Miller, and Michael Kohlhase. ‘The LaTeXML Daemon: Editable Math on the Collaborative Web.” *Intelligent Computer Mathematics*. 2011.

[10] Deyan Ginev. *arXMLiv*:2020 *dataset, an HTML5 conversion of arXiv.org*. SIGMathLing. 2020. https://sigmathling.kwarc.info/resources/arxmliv-dataset-2020/.

[11] Bruce Miller. LaTeXML The Manual—A LaTeX to XML/HTML/MathML Converter, Version 0.8.3. 2018. https://dlmf.nist.gov/LaTeXML/.

[12] Bruno Oberle. “SACR: A Drag-and-Drop Based Tool for Coreference Annotation.” In *Proceedings of the* 11*th International Conference on Language Resources and Evaluation* (LREC 2018).

付録 A　アノテータの募集方法とアノテーション対象の論文

数式を含む英文科学論文の読解を伴うアノテーションタスクは一般のクラウドソーシングでは実現が難しいため，本研究に参加するアノテータは，ソーシャル・ネットワーキング・サービスも活用し[7]広く理工系分野を専攻する学生の中から募集し，直接謝金を支払って作業に従事してもらった．実際のアノテーションに従事した学生は，分野としては自然言語処理を専門とする者が最多だが，他にも数理論理学・代数学・物理学・天文学などを専攻する者もいた．参加アノテータの身分としては大学院生（修士課程）が最も多く，大学院生（博士課程）や学部生，さらにはそれよりも若い参加者もいた．アノテーション対象の論文は arXiv.org に LaTeX 文書ソース付きで収録されているものの中から，参加アノテータの専門と論文中の数式の数を考慮して選定した（表 A）．

表 A　アノテーション対象論文．論文の通し番号は本文の表 2 の番号と対応している．

番号	著者	タイトル	arXiv ID	arXiv カテゴリ
1	Osvaldo Simeone	A Very Brief Introduction to Machine Learning With Applications to Communication Systems	1808.02342	cs.IT
2	Tsung-Hsien Wen et al.	Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems	1508.01745	cs.CL
3	Qian Chen et al.	Enhanced LSTM for Natural Language Inference	1609.06038	cs.CL
4	Joseph Singleton	A Logic of Expertise	2107.10832	cs.LO
5	Edward Frenkel	Recent Advances in the Langlands Program	math0303074	math.AG
6	Laura Aina et al.	Putting words in context: LSTM language models and lexical ambiguity	1906.05149	cs.CL
7	Jian Guan et al.	A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation	2001.05139	cs.CL
8	Richard Antonello et al.	Selecting Informative Contexts Improves Language Model Finetuning	2005.00175	cs.CL
9	Jinhua Zhu et al.	Incorporating BERT into Neural Machine Translation	2002.06823	cs.CL
10	Xuan-Phi Nguyen et al.	Tree-structured Attention with Hierarchical Accumulation	2002.08046	cs.CL
11	Jiangang Bai et al.	Syntax-BERT: Improving Pre-trained Transformers with Syntax Trees	2103.04350	cs.CL
12	Zenan Xu et al.	Syntax-Enhanced Pre-trained Model	2012.14116	cs.CL
13	Yangyifan Xu et al.	Bilingual Mutual Information Based Adaptive Training for Neural Machine Translation	2105.12523	cs.CL
14	Daisuke Taniguchi et al.	Effective temperatures of red supergiants estimated from line-depth ratios of iron lines in the YJ bands, 0.97–1.32 micron	2012.07856	astro-ph.SR
15	Daisuke Taniguchi et al.	Pressure-induced two-step spin crossover in double-layered elastic model	1708.02771	cond-mat.mtrl-sci

付録 B　前処理における数式マークアップの修正

MioGatto は単純に入力 XHTML 文書のタグ情報を参照することでアノテーション対象となるトークナイズされた識別子を認識している．入力 XHTML 文書のタグ構造は LaTeXML で変換する前の論文著者による LaTeX ソースの記述のされ方によって決定されているため，大元の LaTeX ソースの数式マークアップが誤っていると正しく識別子を認識することができない．そのため，このようなマークアップミスは前処理において修正を行った．この過程で修正された誤ったマークアップのほとんどは大きく 3 種類に分類できる．第一は数式でない部分に対して（単に強調などでイタリック体の出力を得る目的で）数式モードが使用されているケースである．第二は逆に数式であるはずの部分がテキストとしてマークアップされているケースである．例えば数式内で FNN のような関数名を使用する際に `\text{FNN}` と記述されていると，その部分は数式識別子ではなくテキストと扱われてしまう．第三は本来はひとまとまりの識別子が，1 文字ごとの識別子の積であるものとしてマークアップされているケースである．例えば正弦関数は LaTeX では `$\sin$` と記述される必要があるが，これが単に `$sin$` と記述されていると s, i, n はそれぞれ別々の変数として認識され，全体としては 3 変数の積となってしまう．実際に LaTeXML の出力では各変数間に不可視の乗算記号 (Invisible Times, `U+2062`) が挿入された状態となってしまい，意図された数式とは明確に異なるものとなる．

7）`https://twitter.com/wtsnjp/status/1410154165288902658`