The GF Mathematics Library

Jordi Saludes
Universitat Politècnica de Catalunya
Sistemes Avançats de Control
jordi.saludes@upc.edu

Sebastian Xambó
Universitat Politècnica de Catalunya
MA2, Edifici OMEGA, Barcelona (Spain)
sebastia.xambo@upc.edu

The aim of this paper is to present the *Mathematics Grammar Library*. The main points are the context in which it originated, its current design and functionality and the present development goals. We also comment on possible future applications in the area of artificial mathematics assistants.

1 Introduction

An archetypal meeting point for natural language processing and mathematics education is the realm of *word problems* ([1, 7]), a realm in which mechanised mathematics assistants (MMA) are expected to play an ever more prominent role in the years to come. The Mathematics Grammar Library (MGL) presented in this paper is not an MMA, but our working hypothesis is that it is a good basis on which to build useful MMA's for learning and teaching (cf. [14, 15] for some general clues on e-learning technologies).

In fact, we regard MGL as an enabling technology for *multilingual* dialog systems capable of helping students in learning how to solve word problems. This assessment is based on the MGL potential capabilities for dealing effectively with a mixture of text and mathematical expressions, and also for managing interactions with ancillary CAS systems.

To be more specific, the current general aim for MGL is to provide natural language services for mathematical constructs at the level of high school and college freshmen linear algebra and calculus. At the present stage, the concrete goal is to provide rendering of simple mathematical exercises in multiple languages (see [6] for a demo).

2 Origin

For a closer view of MGL, let us look briefly at its origins. The idea behind MGL was born, to a good extend, on reflecting about one of the key results of the WEBALT project (see [4, 2]). In summary, the unfolding of this reflection went as follows.

One of the aims of WEBALT was to produce a proof-of-concept platform for the creation of a multilingual repository of *simple* mathematical problems with guaranteed quality of the (machine) translations, in both linguistic and mathematical terms. The languages envisioned were Catalan, English, Finnish, French, Italian and Spanish. Of these, Finnish, with its great complexities, could not be raised to the same level of functionality as the others.

The WebALT prototype was successful and, as far as we know, that endeavour brought about the first application of the GF system ([10, 11]) for the multilingual translation of simple mathematical questions. The powerful GF scheme, based on the perfect interlocking of abstract and concrete grammars, was found to be a very sound choice, but the solution had several shortcommings that could not be addressed in that project. For the present purposes, the three that worried us the most are the following:

Submitted to:
SOS 2007

- The grammars did not work for later versions of GF (>2.9).
- The library was not modular with respect to semantic processing, and hence not easy to maintain.
- It included too few languages, especially as seen from an Europan perspective.

The springboard for the present library was the need to properly solve these problems, inasmuch as this was regarded as one of the most promissing prerequisites for all further advanced developments in machine processing of mathematical texts. Thus the main tasks were:

- To design a *modular* mathematics library structured according to the semantic standards (content dictionaries) of OPENMATH ([8]).
- To code it in the much more expressive GF 3.1 for the few languages mentioned above, and
- To write new code for a few additional languages (Bulgarian, Finnish, German, Romanian and Sweedish).

The first two points amount to a tidying of the original WEBALT programming methods. The third point represents not merely an addition of a few more languages, but a thourough testing of the methods and procedures enforced in the preceding steps. This testing is important in order to secure the rules for the inclusion of further languages and for a controlled uniform extension of the available grammars.

3 Some basic GF notions

For convenience, we include a few notions about the GF system that will ease the considerations about MGL in the next section. For a thourough reference, see [11].

Any GF application begins by specifying its *abstract syntax*. This syntax contains declarations of *categories* (the GF name for types) and *functions* (the GF name for constructor signatures) and has to capture the *semantic structure* of the application domain. For example, to let `Nat` stand for the type of natural numbers and `Prop` for propositions about natural numbers, the GF syntax is

```
cat Nat, Prop ;
```

That ‘zero is a natural number’ and that ‘the successor of any natural number is a natural number’ can be expressed as follows:

```
fun
   Zero : Nat ;
   Succ : Nat -> Nat ;
```

The signatures for ‘even number’ and ‘prime number’ can be captured with

```
fun
   Even, Prime : Nat -> Prop ;
```

Finally, we can abstract the logical ‘not’, ‘and’ and ‘or’ as follows:

```
fun
   Not : Prop -> Prop ;
   And, Or : Prop -> Prop -> Prop ;
```

In practical terms, these declarations would form the *body* of an *abstract module* that would have the form

```
abstract Arith = {<body>}
```

where `Arith` is the name of the module.

4 The library

MGL is a collection of GF modules. The categories used in these modules are in correspondence with all possible combinations of **Variable** and **Value** with the mathematical types **Number**, **Set**, **Tensor** and **Function**. Thus the category `VarNum` denotes a numeric variable like x, while `ValSet` denotes an actual set like “the domain of the natural logarithm”. The distinction between variables and values allows us to type-check productions like lambda abstractions that require a variable as the first argument. Obviously variables can be promoted to values when needed.

Other categories stand for propositions, geometric constructions and indices.

The library is organised in a matrix-like form, with an horizontal axis ranging over the targeted natural languages. At the moment these are: Bulgarian, Catalan, English, Finnish, French, German, Italian, Romanian, Spanish and Swedish.

The vertical axis is for complexity and contains from bottom to top, three layers:

1. *Ground*. It deals with literals, indices and variables.
2. *OpenMath*. It is modelled after the OM *Content Dictionaries* (CD’s). Each CD defines a collection of mathematical objects. This is a *de facto* standard for mathematical semantics and for each *Content Dictionary* there is a corresponding module in this layer. The CD’s are collected by the *OpenMath consortium* [8].
3. *Operations*. This layer takes care of simple mathematical exercises. These appear in drilling materials and usually begin with directives such as ‘Compute’, ‘Find’, ‘Prove’, ‘Give an example of’, . . .

The following tree is an example belonging to the *OpenMath* layer:

```
mkProp
(lt_num
  (abs (plus (BaseValNum (Var2Num x) (Var2Num y))))
  (plus (BaseValNum (abs (Var2Num x)) (abs (Var2Num y)))))
```

When linearized with the Spanish concrete grammar, it yields

> El valor absoluto de la suma de x y de y es menor que la suma del valor absoluto de x y del valor absoluto de y

Similarly, the tree

```
DoSelectFromN
  (Var2Num y)
  (domain (inverse tanh))
  (mkProp
    (gt_num
    (At cosh (Var2Num y))
     pi))
```

gives, when linearized with the English concrete grammar:

> Select y from the domain of the inverse of the hyperbolic tangent such that the hyperbolic cosine of y is greater than pi.

5 Conclusions and further work

After a first step in which the main concern was tidying and modularizing the WEBALT prototype for simple mathematics exercises for the languages Catalan, English, French, Italian and Spanish, we have extended it, in a second step, with the languages Bulgarian, Finnish, German, Romanian and Sweedish (Finnish was considered in the first step, but it had to be worked out from scratch). In this note we have described the GF library produced so far for multilingual mathematical text processing, which we call MGL (Mathematics Grammar Library). We have also indicated how it originated in the WEBALT projecte, its relation to GF, and its present functionality.

Further work has three main lines:

- Addition of new languages, like Danish, Dutch, Norwegian, Polish, Portuguese, Russian, ... This is a continuation of the first two steps referred to above and our assessment is that it can be done reliably with the methods and procedures established so far. To some extent, the library modules for a new language can be generated automatically up to a point in which the remaining work corresponds to natives in that language.
- Describing a controlled procedure for the uniform and reliable extension of the grammars according to new semantic needs. This is an important step that is being researched from several angles. One important point is to ascertain when a piece of mathematical text requires functionalities (categories, constructors, operations) not covered yet.
- Advancing in the use of MGL for the production of ever more sophisticated artificial mathematics assistants. This is also the focus of current research that includes a collaboration with statistical machine translation methods, as in principle they can suggest grammatical structures out of a corpus of mathematical sentences.

Acknowledgements

We are grateful for the generous help of many people and institutions: The WEBALT and MOLTO projects

WebALT

MOLTO steering committee

Katerina Bohmová, Alba Hierro, Ramona Enache,

Francesc Massanés, Thomas Hallgren, Inari Listenmaa, Krasimir Angelov

Barcelona MT team: Cristina España-Bonet, Lluís Márquez, Xavier Carreras

word problem: a type of textbook problem designed to help students apply abstract mathematical concepts to ‘real-world’ situations. Significant background information on the problem is presented as text rather than in mathematical notation.

> A farm raises two kinds of animals: ducks and rabbits. In total there are 60 animals and 140 legs. How many ducks and rabbits are there in the farm?
>
> John is David’s father. He is three times as old as his son, but in 10 years, he will be twice as old. How old are they now?

Another motivating experience. In 1991 A. Deledicq started in France the *Kangourou sans frontières*, a mathematical contest for secondary education students that P. O’Holloran had been running in Australia for thousands of students. In the following years, it was introduced progressively, with similar names, in many other countries around the world. The contest has four levels, according to age (13 to 16). Each level consists of thirty questions, the allowed time is 75 minutes and calculators are not allowed. At present it is taken by hundreds of thousands of students.

For our research, one interesting aspect is that it is inheretly a multilingual undertaking. The problems are proposed by all national mathematical societies that have adopted the contest and are chosen by a large international committee composed of delegates from those societies. Once chosen, the problems are translated by each society, so that in principle there exists a very large body of mathematical questions with human translations into more than forty languages. Unfortunately, the national societies tend to adapt the questions to their curricula and this causes that in many instances we do not have a completely faithful translation.

Here are two examples that hopefully convey a flavour of the kind of questions that are proposed.

> (Proposed by Russia) What is the least number of digits you need to delete in the number 12323314 in order to find a palindrome?
>
> A) 1 B) 2 C) 3 D) 4 E) 5

> (Proposed by Ukraine) We have three boxes: one is white, another is red and the last is black. One is empty, one has an apple, and another has a chocolate bar. Find what color is the box that has the chocolate bar if we know that the chocolate is in the white box or in red box and the apple is neither in the white box nor in the black box.
>
> A) White B) Red C) Black D) Red or Black E) It is impossible to know

References

[1] *Word problem (mathematics education).* `http://en.wikipedia.org/wiki/Word_problem_(mathematics_education)`.

[2] O. Caprotti (2006): *WebALT! Deliver Mathematics Everywhere.* In: *Proceedings of SITE 2006.*

[3] O. Caprotti, W. Ng'ang'a & M. Seppälä (2005): *Multilingual technology for teaching mathematics.* In: *Proceedings of the International Conference on Engineering Education, Instructional Technology, Assessment, and E-learning (EIAE 05).*

[4] O. Caprotti & M. Seppälä (2006): *Multilingual Delivery of Online Tests in Mathematics.* In: *Proceedings of Online Educa Berlin 2006.*

[5] L. Carlson, J. Saludes & A. Strotmann (2005): *State of the art in multilingual and multicultural creation of digital mathematical content.* Available at `http://webalt.math.helsinki.fi/content/e16/e301/e305/Deliverable1.2_eng.pdf`.

[6] Thomas Hallgren & Jordi Saludes (2010): *Math bar online.* Available at `http://www.grammaticalframework.org/demos/minibar/mathbar.html`.

[7] B Greer L Verschffel & E De Corte (2000): *Making sense of word problems.* Taylor & Francis.

[8] Paul Libbrecht (2010): *OpenMath.* Available at `http://www.openmath.org/`.

[9] W. Ng'ang'a (2006): *Multilingual content development for eLearning in Africa.* In: *eLearning Africa: 1st Pan-African Conference on ICT for Development, Education and Training.*

[10] Aarne Ranta (2010): *Grammatical Framework.* Available at `http://www.grammaticalframework.org/`.

[11] Aarne Ranta (2011): *Grammatical Framework: Programming with Multilingual Grammars.* CSLI Publications, Stanford.

[12] A. Strotmann, W. Ng'ang'a & O. Caprotti (2005): *Multilingual Access to Mathematical Exercise Problems.* In M. Dobreva & J. Engelen, editors: *Electronic Proceedings of the Internet Accessible Mathematical Computation Workshop. ISSAC 2005*, Chinese Academy of Sciences, Bejing, China.

[13] A. Strotmann & M. Seppälä (2005): *Web Advanced Learning Technologies for Multilingual Mathematics Teaching Support*. In M. Dobreva & J. Engelen, editors: *ELPUB2005. From Author to Reader: Challenges for the Digital Content Chain. Proceedings of the 9th ICCC International Conference on Electronic Publishing*, Peeters Publishing, Leuven-Heverlee (Belgium).

[14] S Xambó, H Bass, G Bolanos, R Seiler & M Seppälä (2006): *E-Learning Mathematics*. In: *Proceedings of the ICM-2006 (Volume III)*, European Mathematical Society, pp. 1743–1768.

[15] S Xambó, O Caprotti & M Seppälä (2008): *Toward autonomous learners of mathematics*. In E M Rocha J M Borwein & J F Rodrigues, editors: *Communicating Mathematics in the Age of Digital Libraries*, A K Peters, pp. 239–252.