smultiling.sty: Multilinguality Support for sTeX

Michael Kohlhase, Deyan Ginev
Jacobs University, Bremen
http://kwarc.info/kohlhase

November 23, 2017

Abstract

The smultiling package is part of the sTeX collection, a version of TeX/LaTeX that allows to markup TeX/LaTeX documents semantically without leaving the document format, essentially turning TeX/LaTeX into a document format for mathematical knowledge management (MKM).

The smultiling package adds multilinguality support for sTeX, the idea is that multilingual modules in sTeX consist of a module signature together with multiple language bindings that inherit symbols from it, which also account for cross-language coordination.

Contents

1 Introduction

We have been using sTEX as the encoding for the Semantic Multilingual Glossary of Mathematics (SMGloM; see [Gin+16; SMG]). The SMGloM data model has been taxing the representational capabilities of sTEX with respect to multilingual support and verbalization definitions; see [Koh14], which we assume as background reading for this note.

1.1 sTEX Module Signatures

(Monolingual) sTEX had the intuition that the symbol definitions (`\symdef` and `\symvariant`) are interspersed with the text and we generate sTEX module signatures (SMS `*.sms` files) from the sTEX files. The SMS duplicate "formal" information from the "narrative" sTEX files. In the SMGloM, we extend this idea by making the the SMS primary objects that contain the language-independent part of the formal structure conveyed by the sTEX documents and there may be multiple narrative "language bindings" that are translations of each other – and as we do not want to duplicate the formal parts, those are inherited from the SMS rather than written down in the language binding itself. So instead of the traditional monolingual markup in Figure 1, we we now advocate the divided style in Figure 2.

```
\begin{module}[id=foo]
\symdef{bar}{BAR}
\begin{definition}[for=bar]
  A \defiii{big}{array}{raster} ($\bar$) is a\ldots, it is much
  bigger than a \defiii[sar]{small}{array}{raster}.
\end{definition}
\end{module}
```

Example 1: A module with definition in monolingual sTEX

We retain the old `module` environment as an intermediate stage. It is still useful for monolingual texts. Note that for files with a module, we still have to extract `*.sms` files. It is not completely clear yet, how to adapt the workflows. We clearly need a lmh or editor command that transfers an old-style module into a new-style signature/binding combo to prepare it for multilingual treatment.

2 The User Interface

`langfiles` The `smultiling` package accepts the `langfiles` option that specifies – for a module ⟨*mod*⟩ that the module signature file has the name ⟨*mod*⟩`.tex` and the language bindings of language with the ISO 639 language specifier ⟨*lang*⟩ have the file name ⟨*mod*⟩.⟨*lang*⟩`.tex`.[1]

EdN:1

[1] EdNote: implement other schemes, e.g. the onefile scheme.

```
\usepackage{multiling}
\begin{modsig}{foo}
  \symdef{bar}{BAR}
  \symi[gfc=N]{sar}
\end{modsig}

\begin{modnl}[creators=miko,primary]{foo}{en}
  \begin{definition}
    A \defiii[bar]{big}{array}{raster} ($\bar$) is a\ldots, it is much bigger
    than a \defiii[sar]{small}{array}{raster}.
  \end{definition}
\end{modnl}

\begin{modnl}[creators=miko]{foo}{de}
  \begin{definition}
    Ein \defiii[bar]{gro"ses}{Feld}{Raster} ($\bar$) ist ein\ldots, es
    ist viel gr"o"ser als ein \defiii[sar]{kleines}{Feld}{Raster}.
  \end{definition}
\end{modnl}
```

Example 2: Multilingual STEX for Figure 1.

2.1 Multilingual Modules

`modsig` There the `modsig` environment works exactly like the old `module` environment, only that the `id` attribute has moved into the required argument – anonymous module signatures do not make sense.

`modnl` The `modnl` environment takes two arguments the first is the name of the module signature it provides language bindings for and the second the ISO 639 language specifier of the content language. We add the `primary` key `modnl`, which can specify the primary language binding (the one the others translate from; and which serves as the reference in case of translation conflicts).[2]

EdN:2

`\symi` `\symi*` `\symii` `\symiii` `\symiv` There is another difference in the multilingual encoding: All symbols are introduced in the module signature, either by a `\symdef` or the new `\symi` macro. `\symi[`⟨*keys*⟩`]{`⟨*name*⟩`}` takes a symbol name ⟨*name*⟩ as an argument and reserves that name. The variant `\symi*[`⟨*keys*⟩`]{`⟨*name*⟩`}` declares ⟨*name*⟩ to be a primary symbol; see [Koh14] for a discussion. STEX provides variants `\symii`, `\symiii`, and `\symiv` – and their starred versions – for multi-part names. The key-value interface ⟨*keys*⟩ does not have any effect on the LATEX rendering, it can be used to embed metadata. See for instance Subsection 2.5.

2.2 Multilingual Definitions and Cross-referencing Terms

We do not need a new infrastructure for defining mathematical concepts, only the realization that symbols are language-independent. So we can use symbols for the coordination of corresponding verbalizations. As the example in Figure 2 already

[2] EdNote: @Hang: This needs to be implemented in LaTeXML

shows, we can just specify the symbol name in the optional argument of the `\defi` macro to establish that the language bindings provide different verbalizations of the same symbol.

For multilingual term references the situtation is more complex: For single-word verbalizations we could use `\atrefi` for language bindings. Say we have introduced a symbol `foo` in English by `\defi{foo}` and in German by `\defi[foo]{Foo}`. Then we can indeed reference it via `\trefi{foo}` and `\atrefi{Foo}{foo}`. But one the one hand this blurs the distinction between translation and "linguistic variants" and on the other hand does not scale to multi-word compounds as `bar` in Figure 2, which we would have to reference as `\atrefiii{gro"ses Feld Raster}{bar}`. To avoid this,

`\mtref` the `smultiling` package provides the new macros `\mtrefi`, `\mtrefii`, and `\mtrefiii` for multilingual references. Using this, we can reference `bar` as `\mtrefiii[?bar]{gro"ses}{Feld}{Raster}`, where we use the (up to three) mandatory arguments to segment the lexical constituents.

The first argument it syntactically optional to keep the parallelism to `\*def*` `\*tref*` it specifies the symbol via its name ⟨*name*⟩ and module name ⟨*mod*⟩ in a MMT URI ⟨*mod*⟩`?`⟨*name*⟩. Note that MMT URIs can be relative:

1. `foo?bar` denotes the symbol `bar` from module `foo`
2. `foo` the module `foo` (the symbol name is induced from the remaining arguments of `\mtref*`)
3. `?bar` specifies symbol `bar` from the current module

Note that the number suffix `i`/`ii`/`iii`/`iv` indicates the number of words in the actual language binding, not in the symbol name as in `\atref*`.

Finally note that hyperlinks on term references only have information on the underlying symbol and module names – i.e. signature information – and we need to cross-reference into the language bindings. To do this, we need to know the base language of the document. To ensure basic functionality we set this to `en`

`\sTeXlanguage` and provide the `\sTeXlanguage` macro to set it.

2.3 Multilingual Views

Views receive a similar treatment as modules in the `smultiling` package. A multilingual view consists of

`viewsig` 1. a **view signature** marked up with the `viewsig` environment. This takes three required arguments: a view name, the source module, and the target module. The optional first argument is for metadata (`display`, `title`, `creators`, and `contributors`) and load information (`loadfrom` and `loadto`) and

`viewnl` 2. multiple **language bindings** marked up by the `viewnl` environment, which takes two required arguments: the view name and the language specifier. The optional first key/value argument takes the same keys as `viewsig` except the last two.

```
\begin{viewsig}[creators=miko]{norm-metric}{metric-space}{norm}
  \vassign{base-set}{base-set}
  \fassign{x,y}{\metric{x,y}}{\norm{x-y}}
\end{viewsig}
```

Views have language bindings just as modules do, in our case, we have

```
\begin{viewnl}[creators=miko]{norm-metric}{en}
  \obligation{metric-space}{obl.norm-metric.en}
  \begin{assertion}[type=obligation,id=obl.norm-metric.en]
    $\defeq{d(x,y)}{\norm{x-y}}$ is a \trefii[metric-space]{distance}{function}
  \end{assertion}
  \begin{sproof}[for=obl.norm-metric.en]
    {we prove the three conditions for a distance function:}
    ...
  \end{sproof}
\end{viewnl}
```

2.4 Mathematical Keywords

EdN:3 EdN:4 For translations of the mathematical keywords, the `statements` and `sproofs` packages in sTEX define special language definition files, e.g. `statements-ngerman.ldf`.[3][4] There is currently only very limited support for this.

2.5 GF Metadata

Several sTEX macros and environments allow keys for syntactical information about the objects declared.

`gfc` The symbol-declaring macros `\symi` and friends as well as `\symdef` allow gfc key allows to specify the grammatical category in terms of the Resource Grammar of the Grammatical Framework [GFR].

The verbalization-defining macros `\defi` and friends allow the `gfa` (GF apply) and `gfl` (GF linearization) keys.

A definiendum of the form `\defii[gfa=mkN]{empty}{set}` generates the GF linearization `empty_set = mkN "empty set"`. Some what less conveniently, `\defii[name=datum,gfl={mkN "Datum", "Daten"}{Datum}` can be used if the GF linearization is more complex than simply applying a "make command" to the verbalization.

3 Limitations

We list the limitations of the `smultiling` package.

[3] EdNote: say more about this

[4] EdNote: There is the translator package which belongs to beamer, maybe we should switch to that.

3.1 General `babel` Integration

There is currently no integration with the `babel` package that handles language-specific aspects in LaTeX. In particular, selecting the right language must be done manually. In particular, the example from Figure **??** would really have the form given in Figure 3 – see the `\usepackage[usenglish,ngerman]{babel}` in line 2, and the `\selectlanguage` statements in lines 6 and 13.

```
\usepackage{multiling}
\usepackage[usenglish,ngerman]{babel}% babel support
\begin{modsig}{foo}
  \symdef{bar}{BAR}
  \symi{sar}
\end{modsig}
\selectlanguage{english}% english version follows
\begin{modnl}[creators=miko,primary]{foo}{en}
  \begin{definition}
    A \defiii[bar]{big}{array}{raster} ($\bar$) is a\ldots, it is much bigger
    than a \defiii[sar]{small}{array}{raster}.
  \end{definition}
\end{modnl}
\selectlanguage{german}% german umlauts please
\begin{modnl}[creators=miko]{foo}{de}
  \begin{definition}
    Ein \defiii[bar]{gro"ses}{Feld}{Raster} ($\bar$) ist ein\ldots, es
    ist viel gr"o"ser als ein \defiii[sar]{kleines}{Feld}{Raster}.
  \end{definition}
\end{modnl}
```

Example 3: Multilingual sTeX with `babel`

For the `langfiles` setup, which assumes that module signatures and language bindings are in separate files, `babel` integration can be simplified by providing a language-specific preamble file with `\usepackage{`⟨*language*⟩`}{babel}` which is pre-pended to all language binding files when formatted. This preamble can also contain the other language-specific packages (e.g. for font encodings, etc.).

3.2 PDF links on term references are language-dependent

Given the `langfiles` mode, we need the intended language to generate PDF links on term references. But we cannot infer this for top-level "papers" (we do in the language bindings). So it has to be specified via `\sTexlanguage`, and we do not really had a way to check that it is. Unfortunately, the only place it would be natural to do so is in `\mod@component`, but the `\PackageError` there had to be commented out, since it leads to serious errors. Thus we set the language to `en` by default, which is sub-optimal. Maybe there is a way to infer the document language from the babel settings.

3.3 Language-Specific Limitations

Some languages have more problems than others

Turkish makes `=` an active character (to give better spacing); this interacts unfavourably with the `keyval` package which needs `=` as key/value separator (and gives it a different category code). Therefore we need to prohibit this by restricting the `shorthands` option: use `\usepackage[turkish,shorthands=:!]{babel}`.

EdN:5 **Chinese** needs special fonts and `xelatex`[5].

[5] EdNote: get Jinbo to document this

4 Implementation

4.1 Class Options

```
⟨*sty⟩
\newif\if@smultiling@mh@\@smultiling@mh@false
\DeclareOption{mh}{\@smultiling@mh@true}
\newif\if@langfiles\@langfilesfalse
\DeclareOption{langfiles}{\@langfilestrue}
\DeclareOption*{\PassOptionsToPackage{\CurrentOption}{modules}}
\ProcessOptions
```

We load the packages referenced here.

```
\if@smultiling@mh@\RequirePackage{smultiling-mh}\fi
\RequirePackage{etoolbox}
\RequirePackage{structview}
```

4.2 Signatures

`modsig` The `modsig` environment is just a layer over the `module` environment. We also redefine macros that may occur in module signatures so that they do not create markup. Finally, we set the flag `\mod@`⟨*mod*⟩`@multiling` to `true`.

```
\newenvironment{modsig}[2][]{\def\@test{#1}%
\ifx\@test\@empty\begin{module}[id=#2]\else\begin{module}[id=#2,#1]\fi%
\expandafter\gdef\csname mod@#2@multiling\endcsname{true}
\ignorespacesandpars}
{\end{module}\ignorespacesandparsafterend}
```

`\mod@component` We redefine the macro from the `modules` package that computes the module component identifier for external links on term references. If `\mod@`⟨*mod*⟩`@multiling` is `true`, then we make the component identifier `.`⟨*lang*⟩, which can be customized by the next macro below.

```
\renewcommand\mod@component[1]{%
\expandafter\ifx\csname mod@#1@multiling\endcsname\@true%
\@ifundefined{smultiling@language}{}
% for some reason this error message bombs big time; so we leave it out.
% {\PackageError{smultiling}%
%   {No document language specified for term reference links}
%   {use \protect\sTeXlanguage to specify it!}}
{.\smultiling@language}%
\fi}
```

`\sTeXlanguage` This macro sets the internal flag `\smultiling@language`, we set the default to `en`, since otherwise hyper-references on term references do not work.

```
\newcommand\sTeXlanguage[1]{\def\smultiling@language{#1}}
\sTeXlanguage{en}
```

`viewsig` The `viewsig` environment is just a layer over the `view` environment with the keys suitably adapted.

```
\newenvironment{viewsig}[4][]{\def\@test{#1}\ifx\@test\@empty%
```

```
  \begin{view}[id=#2,ext=tex]{#3}{#4}\else\begin{view}[id=#2,#1,ext=tex]{#3}{#4}\fi%
  \ignorespacesandpars}
  {\end{view}\ignorespacesandparsafterend}
```

`\@sym*` has a starred form for primary symbols. The key/value interface has no effect on the LaTeX side. We read the to check whether only allowed ones are used. We also add the `gfc` key to `\symdef`, since that makes a

```
\newcommand\symi{\@ifstar\@symi@star\@symi}
\newcommand\@symi[2][]{\metasetkeys{symi}{#1}%
  \if@importing\else\par\noindent Symbol: \textsf{#2}\fi\ignorespacesandpars}
\newcommand\@symi@star[2][]{\metasetkeys{symi}{#1}%
  \if@importing\else\par\noindent Primary Symbol: \textsf{#2}\fi\ignorespacesandpars}
\newcommand\symii{\@ifstar\@symii@star\@symii}
\newcommand\@symii[3][]{\metasetkeys{symi}{#1}%
  \if@importing\else\par\noindent Symbol: \textsf{#2-#3}\fi\ignorespacesandpars}
\newcommand\@symii@star[3][]{\metasetkeys{symi}{#1}%
  \if@importing\else\par\noindent Primary Symbol: \textsf{#2-#3}\fi\ignorespacesandpars}
\newcommand\symiii{\@ifstar\@symiii@star\@symiii}
\newcommand\@symiii[4][]{\metasetkeys{symi}{#1}%
  \if@importing\else\par\noindent Symbol: \textsf{#2-#3-#4}\fi\ignorespacesandpars}
\newcommand\@symiii@star[4][]{\metasetkeys{symi}{#1}%
  \if@importing\else\par\noindent Primary Symbol: \textsf{#2-#3-#4}\fi\ignorespacesandpars}
\newcommand\symiv{\@ifstar\@symiv@star\@symiv}
\newcommand\@symiv[5][]{\metasetkeys{symi}{#1}%
  \if@importing\else\par\noindent Symbol: \textsf{#2-#3-#4-#5}\fi\ignorespacesandpars}
\newcommand\@symiv@star[5][]{\metasetkeys{symi}{#1}%
  \if@importing\else\par\noindent Primary Symbol: \textsf{#2-#3-#4-#5}\fi\ignorespacesandpars}
```

4.3 Language Bindings

`modnl:*`

```
\addmetakey{modnl}{load}
\addmetakey*{modnl}{title}
\addmetakey*{modnl}{creators}
\addmetakey*{modnl}{contributors}
\addmetakey{modnl}{srccite}
\addmetakey{modnl}{primary}[yes]
```

`modnl` The `modnl` environment is just a layer over the `module` environment and the `\importmodule` macro with the keys and language suitably adapted.

```
\newenvironment{modnl}[3][]{\metasetkeys{modnl}{#1}%
  \def\@test{#1}\ifx\@test\@empty\begin{module}[id=#2.#3]\else\begin{module}[id=#2.#3,#1]\fi%
  \def\smultiling@language{#3}%
  \if@langfiles\importmodule[load=#2,ext=tex]{#2}\else
  \ifx\modnl@load\@empty\importmodule{#2}\else\importmodule[ext=tex,load=\modnl@load]{#2}\fi%
  \fi%
  \ignorespacesandpars}
{\end{module}\ignorespacesandparsafterend}
```

viewnl The `viewnl` environment is just a layer over the `view` environment with the keys and language suitably adapted.[6]

EdN:6

```
\newenvironment{viewnl}[5][]{\def\@test{#1}\ifx\@test\@empty%
  \begin{view}[id=#2.#3,ext=tex]{#4}{#5}\else%
  \begin{view}[id=#2.#3,#1,ext=tex]{#4}{#5}\fi%
  \ignorespacesandpars}
  {\end{view}\ignorespacesandparsafterend}
```

4.4 Multilingual Statements and Terms

\mtref we first first define an auxiliary conditional `\@instring` that checks of `?` is in the first argument. `\mtrefi` uses it, if there is one, it just calls `\termref`, otherwise it calls `\@mtrefi`, which assembles the `\termref` after splitting at the `?`.

```
\def\@instring#1#2{TT\fi\begingroup\edef\x{\endgroup\noexpand\in@{#1}{#2}}\x\ifin@}
\def\@mtref#1?#2\relax{\@@mtref{#1}{#2}}
\newcommand\@@mtref[3]{\def\@@cd{#1}\def\@@name{#2}%
\ifx\@@cd\@empty%
\ifx\@@name\@empty\termref[]{#3}\else\termref[name=\@@name]{#3}\fi%
\else%
\ifx\@@name\@empty\termref[cd=\@@cd]{#3}\else\termref[cd=\@@cd,name=\@@name]{#3}\fi%
\fi}
\newcommand\mtref[2][]{\if\@instring{?}{#1}\@mtref #1\relax{#2}\else\termref[cd=#1]{#2}\fi}
```

\mtrefi*

```
\newcommand\mtrefi[2][]{\if\@instring{?}{#1}\@mtref #1\relax{#2}%
\else\termref[cd=#1]{#2}\fi}
\newcommand\mtrefis[2][]{\mtrefi[#1]{#2s}}
\newcommand\Mtrefi[2][]{\if\@instring{?}{#1}\@mtref #1\relax{\capitalize{#2}}%
\else\termref[cd=#1]{\capitalize{#2}}\fi}
\newcommand\Mtrefis[2][]{\Mtrefi[#1]{#2s}}
\newcommand\mtrefii[3][]{\mtrefi[#1]{#2 #3}}
\newcommand\mtrefiis[3][]{\mtrefi[#1]{#2 #3s}}
\newcommand\Mtrefii[3][]{\Mtrefi[#1]{#2 #3a}}
\newcommand\Mtrefiis[3][]{\Mtrefi[#1]{#2 #3s}}
\newcommand\mtrefiii[4][]{\mtrefi[#1]{#2 #3 #4}}
\newcommand\Mtrefiiis[4][]{\Mtrefi[#1]{#2 #3 #4s}}
\newcommand\Mtrefiii[4][]{\Mtrefi[#1]{#2 #3 #4}}
\newcommand\mtrefiiis[4][]{\mtrefi[#1]{#2 #3 #4s}}
\newcommand\mtrefiv[5][]{\mtrefi[#1]{#2 #3 #4 #5}}
\newcommand\mtrefivs[5][]{\mtrefi[#1]{#2 #3 #4 #5s}}
\newcommand\Mtrefiv[5][]{\Mtrefi[#1]{#2 #3 #4 #5}}
\newcommand\Mtrefivs[5][]{\Mtrefi[#1]{#2 #3 #4 #5s}}
```

4.5 GF Metadata

gfc We add the `gfc` key to various symbol declaration macros.

[6] EdNote: MK: we have to do something about the if@langfiles situation here. But this is non-trivial, since we do not know the current path, to which we could append .⟨*lang*⟩!

```
\addmetakey{symi}{gfc}
\addmetakey{symdef}{gfc}%
```

gfa/l

```
\addmetakey{definiendum}{gfa}
\addmetakey{definiendum}{gfl}
```

4.6 Miscellaneneous

the `\ttl` macro (to-translate) is used to mark untranslated stuff. We need a better LaTeXMLtreatment of this eventually that is integrated with MathHub.info.

`\ttl`

```
\newcommand\ttl[1]{\red{TTL: #1}}
```

⟨/sty⟩

Change History

v0.1
General: First Version 1
v0.2
General: Adding a key-value argument to \symi and friends for GF metadata 1

References

[GFR] B. Bringert, T. Hallgren, and A. Ranta. *GF Resource Grammar Library: Synopsis*. URL: http://www.grammaticalframework.org/lib/doc/synopsis.html (visited on 09/27/2017).

[Gin+16] Deyan Ginev et al. "The SMGloM Project and System. Towards a Terminology and Ontology for Mathematics". In: *Mathematical Software - ICMS 2016 - 5th International Congress*. Ed. by Gert-Martin Greuel et al. Vol. 9725. LNCS. Springer, 2016. DOI: 10.1007/978-3-319-42432-3. URL: http://kwarc.info/kohlhase/papers/icms16-smglom.pdf.

[Koh14] Michael Kohlhase. "A Data Model and Encoding for a Semantic, Multilingual Terminology of Mathematics". In: *Intelligent Computer Mathematics*. Conferences on Intelligent Computer Mathematics. (Coimbra, Portugal, July 7–11, 2014). Ed. by Stephan Watt et al. LNCS 8543. Springer, 2014, pp. 169–183. ISBN: 978-3-319-08433-6. URL: http://kwarc.info/kohlhase/papers/cicm14-smglom.pdf.

[SMG] *SMGloM Glossary*. URL: http://mathhub.info/mh/glossary (visited on 04/21/2014).