

smutiling.sty: Multilinguality Support for \TeX

Michael Kohlhase
FAU Erlangen-Nürnberg
<http://kwarc.info/kohlhase>

Deyan Ginev
Authorea

November 16, 2017

Abstract

The `smutiling` package is part of the \TeX collection, a version of $\text{\TeX}/\text{\LaTeX}$ that allows to markup $\text{\TeX}/\text{\LaTeX}$ documents semantically without leaving the document format, essentially turning $\text{\TeX}/\text{\LaTeX}$ into a document format for mathematical knowledge management (MKM).

The `smutiling` package adds multilinguality support for \TeX , the idea is that multilingual modules in \TeX 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	3
1.1	\TeX Module Signatures	3
2	The User Interface	3
2.1	Multilingual Modules	4
2.2	Multilingual Definitions and Crossreferencing Terms	4
2.3	Multilingual Views	5
2.4	Mathematical Keywords	6
2.5	GF Metadata	6
3	Limitations	6
3.1	General <code>babel</code> Integration	6
3.2	Language-Specific Limitations	7
4	Implementation	8
4.1	Class Options	8
4.2	Signatures	8
4.3	Language Bindings	9
4.4	Multilingual Statements and Terms	9

4.5	GF Metadata	10
4.6	Miscellaneneous	10

1 Introduction

We have been using \TeX as the encoding for the Semantic Multilingual Glossary of Mathematics (SMGloM; see [IanJucKoh:sps14]). The SMGloM data model has been taxing the representational capabilities of \TeX with respect to multilingual support and verbalization definitions; see [Koh14], which we assume as background reading for this note.

1.1 \TeX Module Signatures

(monolingual) \TeX had the intuition that the symbol definitions ($\text{\textbackslash symdef}$ and $\text{\textbackslash symvariant}$) are interspersed with the text and we generate \TeX module signatures (SMS $\ast.\text{sms}$ files) from the \TeX files. The SMS duplicate “formal” information from the “narrative” \TeX 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 \TeX 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} ( $\bar{\phantom{x}}$ ) 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 \TeX

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 $\ast.\text{sms}$ 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 $\langle mod \rangle$ that the module signature file has the name $\langle mod \rangle.\text{tex}$ and the language bindings of language with the ISO 639 language specifier $\langle lang \rangle$ have the file name $\langle mod \rangle.\langle lang \rangle.\text{tex}$.¹

¹EDNOTE: 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 \LaTeX 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

\symi 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[$\langle keys \rangle$]{ $\langle name \rangle$ }** takes a symbol name $\langle name \rangle$ as an argument and reserves that name. The variant **\symi*[$\langle keys \rangle$]{ $\langle name \rangle$ }** declares $\langle name \rangle$ to be a primary symbol; see [Koh14] for a discussion. \LaTeX provides variants **\symii**, **\symiii**, **\symiv** – and their starred versions – for multi-part names. The key-value interface $\langle keys \rangle$ 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 Crossreferencing 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

²EdNOTE: @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 situation 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 on 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, 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 is syntactically optional to keep the parallelity to `*def*` `*tref*` it specifies the symbol via its name $\langle name \rangle$ and module name $\langle mod \rangle$ in a MMT URI $\langle mod \rangle ? \langle name \rangle$. 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` indicates the number of words in the actual language binding, not in the symbol name as in `\atref*`.

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]
    $\defeq{d(x,y)}{\|x-y\}}$ 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

For translations of the mathematical keywords, the `statements` and `sproofs` packages in `gTeX` define special language definition files, e.g. `statements-ngerman.ldf`.³⁴ There is currently only very limited support for this.

2.5 GF Metadata

Several `gTeX` 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 [**GFResourceGrammar:on**].

The verbalization-defining macros `\defi` and friends allow the `gfa` (GF apply) and `gfl` (GF linearization) keys. A definiendum of the form `\defi[gfa=mkN]{empty}{set}` generates the GF linearization `empty_set = mkN "empty set"`. `\defi[name=datum,gfl={mkN "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.

3.1 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.

³EDNOTE: say more about this

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

```

\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 \LaTeX 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 Language-Specific Limitations

Some languages have more problems than others

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

Chinese needs special fonts and `xelatex`⁵.

EdN:5

⁵EdNOTE: get Jinbo to document this

4 Implementation

4.1 Class Options

```

1 \*sty)
2 \newif\if@smultiling@mh@\@smultiling@mh@false
3 \DeclareOption{mh}{\@smultiling@mh@true}
4 \newif\if@langfiles@\@langfiles@false
5 \DeclareOption{langfiles}{\@langfiles@true}
6 \DeclareOption*{\PassOptionsToPackage{\CurrentOption}{modules}}
7 \ProcessOptions

```

We load the packages referenced here.

```

8 \if@smultiling@mh\RequirePackage{smultiling-mh}\fi
9 \RequirePackage{etoolbox}
10 \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.

```

11 \newenvironment{modsig}[2] [] {%
12 \def\@test{#1}\ifx\@test\@empty\begin{module}[id=#2]\else\begin{module}[id=#2,#1]\fi\ignorespaces
13 \end{module}\ignorespacesandparsafterend}

```

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

```

14 \newenvironment{viewsig}[4] [] {\def\@test{#1}\ifx\@test\@empty%
15 \begin{view}[id=#2,ext=tex]{#3}{#4}\else\begin{view}[id=#2,#1,ext=tex]{#3}{#4}\fi%
16 \ignorespacesandpars}
17 {\end{view}\ignorespacesandparsafterend}

```

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

```

18 \newcommand\sympi{\@ifstar\@sympi@star\@sympi}
19 \newcommand\@sympi[2] [] {\metasetkeys{sympi}{#1}%
20 \if@importing\else\par\noindent Symbol: \textsf{#2}\fi\ignorespacesandpars}
21 \newcommand\@sympi@star[2] [] {\metasetkeys{sympi}{#1}%
22 \if@importing\else\par\noindent Primary Symbol: \textsf{#2}\fi\ignorespacesandpars}
23 \newcommand\symii{\@ifstar\@symii@star\@symii}
24 \newcommand\@symii[3] [] {\metasetkeys{sympi}{#1}%
25 \if@importing\else\par\noindent Symbol: \textsf{#2-#3}\fi\ignorespacesandpars}
26 \newcommand\@symii@star[3] [] {\metasetkeys{sympi}{#1}%
27 \if@importing\else\par\noindent Primary Symbol: \textsf{#2-#3}\fi\ignorespacesandpars}
28 \newcommand\symiii{\@ifstar\@symiii@star\@symiii}
29 \newcommand\@symiii[4] [] {\metasetkeys{sympi}{#1}%
30 \if@importing\else\par\noindent Symbol: \textsf{#2-#3-#4}\fi\ignorespacesandpars}
31 \newcommand\@symiii@star[4] [] {\metasetkeys{sympi}{#1}%

```



```

32 \if@importing\else\par\noindent Primary Symbol: \textsf{#2-#3-#4}\fi\ignorespacesandpars}
33 \newcommand\symiv{\@ifstar\@symiv@star\@symiv}
34 \newcommand\@symiv[5] [] {\metasetkeys{symi}{#1}%
35 \if@importing\else\par\noindent Symbol: \textsf{#2-#3-#4-#5}\fi\ignorespacesandpars}
36 \newcommand\@symiv@star[5] [] {\metasetkeys{symi}{#1}%
37 \if@importing\else\par\noindent Primary Symbol: \textsf{#2-#3-#4-#5}\fi\ignorespacesandpars}

```

4.3 Language Bindings

`modnl:`*

```

38 \addmetakey{modnl}{load}
39 \addmetakey*{modnl}{title}
40 \addmetakey*{modnl}{creators}
41 \addmetakey*{modnl}{contributors}
42 \addmetakey{modnl}{srccite}
43 \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.

```

44 \newenvironment{modnl}[3] [] {\metasetkeys{modnl}{#1}%
45 \def\@test{#1}\ifx\@test\@empty\begin{module}[id=#2.#3]\else\begin{module}[id=#2.#3,#1]\fi%
46 \if@langfiles\importmodule[load=#2,ext=tex]{#2}\else
47 \ifx\modnl@load\@empty\importmodule{#2}\else\importmodule[ext=tex,load=\modnl@load]{#2}\fi%
48 \fi%
49 \ignorespacesandpars}
50 {\end{module}\ignorespacesandparsafterend}

```

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

```

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

```

4.4 Multilingual Statements and Terms

`\mtref` we first first define an auxiliary conditional `\@instring` that checks if `?` 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 `?`.

```

56 \def\@instring#1#2{TT\fi\begin{group}\edef\x{\end{group}\noexpand\in@{#1}{#2}}\x\ifin@}
57 \def\@mtref#1?#2\relax{\@mtref{#1}{#2}}
58 \newcommand\@mtref[3]{\def\@cd{#1}\def\@name{#2}%
59 \ifx\@cd\@empty%
60 \ifx\@name\@empty\termref[] {#3}\else\termref[name=\@name]{#3}\fi%
61 \else%

```

⁶EDNOTE: 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)`!

```

62 \ifx\@name\empty\termref[cd=\@cd]{#3}\else\termref[cd=\@cd,name=\@name]{#3}\fi%
63 \fi}
64 \newcommand\mtref[2][]{\if\@instring{?}{#1}\@mtref #1\relax{#2}\else\termref[cd=#1]{#2}\fi}

\mtrefi*
65 \newcommand\mtrefi[2][]{\if\@instring{?}{#1}\@mtref #1\relax{#2}%
66 \else\termref[cd=#1]{#2}\fi}
67 \newcommand\mtrefis[2][]{\mtrefi[#1]{#2s}}
68 \newcommand\Mtrefi[2][]{\if\@instring{?}{#1}\@mtref #1\relax{\capitalize{#2}}%
69 \else\termref[cd=#1]{\capitalize{#2}}\fi}
70 \newcommand\Mtrefis[2][]{\Mtrefi[#1]{#2s}}
71 \newcommand\mtrefii[3][]{\mtrefi[#1]{#2 #3}}
72 \newcommand\mtrefiis[3][]{\mtrefi[#1]{#2 #3s}}
73 \newcommand\Mtrefii[3][]{\Mtrefi[#1]{#2 #3a}}
74 \newcommand\Mtrefiis[3][]{\Mtrefi[#1]{#2 #3s}}
75 \newcommand\mtrefiii[4][]{\mtrefi[#1]{#2 #3 #4}}
76 \newcommand\Mtrefiis[4][]{\Mtrefi[#1]{#2 #3 #4s}}
77 \newcommand\Mtrefiii[4][]{\Mtrefi[#1]{#2 #3 #4}}
78 \newcommand\mtrefiis[4][]{\mtrefi[#1]{#2 #3 #4s}}
79 \newcommand\mtrefiv[5][]{\mtrefi[#1]{#2 #3 #4 #5}}
80 \newcommand\mtrefivs[5][]{\mtrefi[#1]{#2 #3 #4 #5s}}
81 \newcommand\Mtrefiv[5][]{\Mtrefi[#1]{#2 #3 #4 #5}}
82 \newcommand\Mtrefivs[5][]{\Mtrefi[#1]{#2 #3 #4 #5s}}

```

4.5 GF Metadata

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

```

83 \addmetakey{symi}{gfc}
84 \addmetakey{symdef}{gfc}%

```

gfa/l

```

85 \addmetakey{definiendum}{gfa}
86 \addmetakey{definiendum}{gfl}

```

4.6 Miscellaneneous

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

\ttl

```

87 \newcommand\ttl[1]{\red{TTL: #1}}
88 \</sty>

```

References

- [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>.