♢ ♢ ♢

MLBIBTEX: *a New Implementation of* BIBTEX

JEAN-MICHEL HUFFLEN*

ABSTRACT. This paper describes MLBIBTEX, a new implementation of BIBTEX with multilingual features. We show how to use it as profitably as possible, and go thoroughly into compatibility between BIBTEX's current implementation and ours. Besides, the precise grammar of MLBIBTEX is given as an annex.

KEYWORDS: Bibliographies, multilingual features, LATEX, BIBTEX.

INTRODUCTION

THERE is increasing interest in multilingual word processors nowadays: some books may be composed of parts written in different languages, some documents have to be produced using several languages: for example, at least three languages for official documents of the EEC[1]. As word processors, TEX and LATEX are indisputably pioneers in this topic. From its first version, TEX [28] has provided commands to produce accents and other diacritical signs of European languages using the Latin alphabet. The ability for non-English words to be hyphenated has been improved by first MLTÊX (for 'Multilingual TÊX') [16] and then TEX's Version 3. When LATEX2ε [32] came out, the french [19] and german [41] packages strongly eased writing documents in French and German, even if these packages were *ad hoc* for one language only. 'Actual' multilinguism has been reached with the babel package [7], in the sense that this package processes all the languages it knows in a homogeneous way, without giving any privilege to a particular one. Therefore this package is especially suitable for mixing several languages within the same document. Besides, this package is now able to process some languages using a non-Latin alphabet (Greek, Russian, ... cf. [7, § 26 & 51]). Last but not least, the Ω and Λ projects [40] aim to develop 'super' TEX and LATEX engines, able to process all the world's languages[2], by using the Unicode standard encoding [45].

Now let us consider BIBTEX [38], the bibliography program associated with LATEX.

*Author's address: LIFC, University of Franche-Comté. 16, route de Gray. 25030 BESANÇON CEDEX. FRANCE. E-mail: hufflen@lifc.univ-fcomte.fr.

[1] **E**uropean **E**conomic **C**ommunity.

[2] In fact, some extensions of TEX were already able to process languages with right-to-left texts as well as languages with left-to-right texts: TEX-XETEX since 1987 [27], and ε-TEX [36].

```
@BOOK{gibson1988,
        AUTHOR = {William Gibson},
        TITLE = {Mona Lisa Overdrive},
        PUBLISHER = {Victor Gollancz, Ltd.},
        YEAR = 1988}
```

Figure 1: Example of a BIBTEX entry.

It offers some flexibility about foreign (that is, non-English) language support (cf. [21, § 13.8.2]), and the insertion of some slight multilingual features have been put into action: for example, the bibliography style files built by means of the makebst program (cf. [13] or [21, § 13.9]) can be interfaced with the babel package; another example is given by the Delphi BibStyles collection (cf. [21, § 13.8.2]). But it is obvious that BIBTEX's present version does not provide as many multilingual features as LATEX's.

Given these considerations, we started a new implementation in October 2000, so-called MLBIBTEX (for 'Multilingual BIBTEX'). The first version (1.1) is available now[3], and has already been introduced in [24], but very informally. For the EUROTEX conference, we propose a more precise description hereafter.

In this article, we do not pay particular attention to the typographical conventions ruling the layout of bibliographies, since MLBIBTEX produces the same outputs as BIBTEX from an 'aesthetic' point of view. Readers interested in this topic can consult some good typography manuals: for example, [10, § 10], [11, § 15.54–15.76], [22, p. 53–54]. French-speaking readers can also refer to [12, § 94] or [34, p. 31–36] about the conventions ruling French bibliographies, German-speaking readers can refer to [15], too. More official documents have been issued by the ANSI[4] [3], or *British Standards* [8, 9], or the French institute of standardisation[5] [1].

We assume that readers are familiar with the usual commands of LATEX and BIBTEX[6] and the way to use them in cooperation. However, we recall some points in order to make precise our terminology. A BIBTEX file (with the .bib suffix) contains a series of **entries** like that given in Figure 1. When BIBTEX runs, it builds a .bbl file that contains **references** for a LATEX document, using the `\bibitem` command, and according to a *bibliography style*. For example, the reference corresponding to Entry `gibson1988` (cf. Figure 1) and using the plain bibliography style—that is, references are labelled with numbers—will look like:

[1] William Gibson. *Mona Lisa Overdrive*. Victor Gollancz, Ltd., 1988.

after being processed by LATEX. There exist many other bibliography styles: most of them are described in [21, § 13.2], more in [25, § 4.3]—including some styles suitable for French bibliographies—and keys to design new bibliography styles are given in [37].

The sections of this article will successively explore first the multilingual exten-

[3] See `http://lifc.univ-fcomte.fr/PEOPLE/hufflen/texts/mlbibtex/mlbibtex/mlbibtex.html`.
[4] **A**merican **N**ational **S**tandards **I**nstitute.
[5] *AFNOR: **A**ssociation **F**rançaise de **NOR**malisation.*
[6] The basic reference is [32]. There are also [14] and [43] in French, [29, 30, 31] in German.

sions provided and processed by MLBIBTEX, then the ways of compatibility between BIBTEX's current implementation and ours. Finally we briefly describe our implementation and we conclude with what we plan for the future of MLBIBTEX. An annex sums up all our syntactic conventions.

The Extensions of mlbibtex

In this section, we present the extensions provided by MLBIBTEX. We attempt to detail them in an informal but precise way. Besides, we show that they meet actual requirements about multilingual bibliographies.

MLBIBTEX allows its users to specify *language changes* and *language switches*. Before we describe them, let us explain what 'the bibliographical entry's language' and 'the bibliographical reference's language' mean. In fact, there are two approaches for multilingual bibliographies.

- According to the first approach, the information related to a reference should be expressed using the language of the corresponding document. For example, the month of issue should be 'July' for a reference about a document written in English, '*juillet*' for a reference about a document written in French, '*Juli*' for a reference about a document written in German. Roughly speaking, the values to be put into the fields of BIBTEX entries are copied slavishly from what is printed on the document. From a 'philosophical' point of view, this convention proceeds from the idea that a reference is wholly suitable only for people reading the language this referred document is written in[7]. We have to add language information to each entry, in the sense that LATEX must be able to format the corresponding reference by using the typographical conventions of this language. Such language information is given by an additional MLBIBTEX field called `LANGUAGE`[8]. The value of this field defaults to `english`. In the following, this approach will be called **reference-dependent**.
- Given a printed work, the second approach consists of using its language for the information of all its bibliographical references, as far as possible. So, in comparison with the example illustrating the first approach, all the months of issue should be expressed in English if the work is written in English. In the same way, they should be expressed in French (resp. German) if the work is written in French (resp. German). However, if this approach is systematic, some information fields other than dates should be superseded by a translated form. For example, let us consider the entry given in Figure 2, concerning the novelisation of a film. This reference can be formatted as it is within a bibliography in English, but the `NOTE`

[7] Following this approach to extremes, the information related to a document written in a language using a non-Latin alphabet should be put using this alphabet's characters. For example, 'июль' for 'July' in Russian. But we can remark that if the title of such a document is transliterated into Roman letters, this operation needs to know the original language when the bibliographical entry is established.

[8] In fact, this 'new' field has already been used in conjunction with the mlbib package [35]. BIBTEX—like MLBIBTEX—allows end-users to add new fields, which are ignored by 'standard' bibliography styles.

```
@BOOK{bisson1995,
        AUTHOR = {Terry Bisson},
        TITLE = {{J}ohnny {M}nemonic},
        NOTE = {Based on a short story and screenplay by William Gibson},
        PUBLISHER = {Simon \&~Shuster, Inc.},
        YEAR = 1995}
```

Figure 2: A BIBTEX entry, but suitable for a bibliography in English.

field should be replaced by:

D'après une histoire et un scénario de William Gibson

within a bibliography in French. In the following, this approach will be called **document-dependent**.

From our point of view, the choice between these two approaches does not have to be made by the designer of a bibliography program like BIBTEX. This choice proceeds from personal considerations, or requirements from an editor or a publisher. That is why we think that a multilingual bibliography program associated with LATEX should be able to put both these two approaches into action. To emphasise this point, let us consider a book like an anthology or proceedings, with several chapters written in several languages. If each chapter ends with its own bibliography and if the publisher requires that the bibliography of a chapter must be written in the chapter's language, a reference—for example, a reference to Entry `bisson1995` above—cited in two chapters written in English and French will appear differently within the two bibliographies. In this case, we show that it should be possible for a bibliography program to be adapted to several languages used within a document, as well as it should be possible for an entry to be used to generate different references according to the language chosen.

There exist some LATEX2ϵ packages for each of these two approaches. About the reference-dependent approach, the mlbib package [35] uses a `LANGUAGE` field and allows each item of a bibiliography to be printed according to suitable typography. About the document-dependent approach, the oxford package [4] allows users to choose a language for the whole of the bibliography. But these implementations are partial: only basic typographical rules (rules for spacing in French, for example) and some keywords (month names, for example) are taken into account. The same fact holds about bibliography styles built with the makebst program (cf. [13] or [21, § 13.9]) and interfaced with the babel package. If 'standard' BIBTEX is used, there is no 'actual' multilinguism in entries, unless users fill in some fields by using commands originating from the multilingual packages of LATEX2ϵ.

So, here are the precise definitions related to our terminology.

◇ A **language identifier** is a non-ambiguous prefix of:
- an option of the babel package,
- or a multilingual package name such as french or german[9].

[9]This choice of a non-ambiguous prefix allows a language identifier to get access to several ways to process a language. For example, a language identifier set to `french` works with the frenchb option [18]

```
@BOOK{king1982f,
        AUTHOR = {Stephen~Edwin King},
        TITLE = {The Running Man},
        NOTE = {[Written as] * english
                [Sous le pseudonyme de] * french
                [Unter der Pseudonym] * german
                Richard Bachman},
        PUBLISHER = {New American Library},
        YEAR = 1982,
        MONTH = may,
        LANGUAGE = english}
```

Figure 3: Example of a multilingual entry.

- ◇ The **entry's language** is given by the `LANGUAGE` field[10].
- ◇ The **reference's language** is given:
  - – either by the `LANGUAGE` field if each item of the bibliography should be expressed in its own language (reference-dependent approach),
  - – or by the language in which the document is written if this language is to be used for the whole of the bibliography (document-dependent approach).

As mentioned above, this convention about the reference's language allows us to put both possible approaches for multilingual bibliographies into action.

Language switches
There are two kinds of language switches, with and without a default language. They are used for information about what must be put, possibly in another language, and for details that can be given in a particular language, but can be omitted if no translation is available.

SWITCH WITH A DEFAULT LANGUAGE It is expressed by the following syntax:

$$\texttt{[}\mathit{string}_0\texttt{]}\ \texttt{*}\ \mathit{idf}_0\ \texttt{[}\mathit{string}_1\texttt{]}\ \texttt{*}\ \mathit{idf}_1\ \ldots\ \texttt{[}\mathit{string}_n\texttt{]}\ \texttt{*}\ \mathit{idf}_n \qquad (1)$$

where string_0, string_1, …, string_n ($n \in \mathbb{N}$) are strings of characters[11] and idf_0, idf_1, …, idf_n are pairwise-different language identifiers. MLBIBTEX processes it as follows:

- ◇ if there exists i ($0 \leq i \leq n$) such that the reference's language is equal to idf_i, Expression (1) yields string_i;
- ◇ otherwise, MLBIBTEX puts the string associated with the value of the `LANGUAGE` field: if such a value does not exist, Expression (1) is replaced by an empty string.

of the babel package as well as the french package [19]. Readers who are interested in a comparative study between these two ways to write French documents can consult [23, § 2].

[10] Let us recall that it defaults to the `english` value.

[11] We follow the convention originating from BIBTEX, that is, a group surrounded by braces ('`{...}`') is viewed as a single character. We will go thoroughly into this point in Subsection *Using square brackets as syntactic tokens*.

```
@BOOK{king1990,
        AUTHOR = {Stephen~Edwin King},
        TITLE = {The Stand},
        NOTE = {[The Complete and Uncut Edition.] * english
                [Version int\'{e}grale.] * french
                [Abridged version issued in 1978] * english
                [Version abr\'{e}g\'{e}e parue en 1978] * french
                [Abgek\"{u}rzt Auffassung im Jahre 1978 erschienen] * german},
        PUBLISHER = {Doubleday \&~C\textsuperscript{o}},
        ADDRESS = {New-York},
        YEAR = 1990,
        LANGUAGE = english}
```

Figure 4: Example with two language switches.

An example is given in Figure 3. When using the plain bibliography style, MLBIBTEX will produce the following references, after processing by LATEX:

◇ when the reference's language is English:

[1] Stephen Edwin King. *The Running Man.* New American Library, May 1982. Written as Richard Bachman.

◇ when it is French:

[1] Stephen Edwin King. *The Running Man.* New American Library, mai 1982. Sous le pseudonyme de Richard Bachman.

◇ when it is German:

[1] Stephen Edwin King. *The Running Man.* New American Library, Mai 1982. Unter der Pseudonym Richard Bachman.

As an example of using default values, if MLBIBTEX is asked for a Russian-speaking reference, the information in English is put since we did not specify any string for the Russian language, and the value of the `LANGUAGE` field is `english`. In fact, only the month name—given by the abbreviation `may` that BIBTEX and MLBIBTEX know—is printed in Russian:

[1] Stephen Edwin King. *The Running Man.* New American Library, май 1982. Writing as Richard Bachman.

'`* idf`' where `idf` is the value of the `LANGUAGE` field can be omitted. The example given in Figure 3 about the `NOTE` field could be abridged as follows:

```
NOTE = {[Written as]
        [Sous le pseudonyme de] * french
        [Unter der Pseudonym] * german
        Richard Bachman}
```

A language switch ends:

◇ either at the end of a MLBIBTEX field or before a common part, not surrounded by square brackets: this second convention holds for the example above, the common part being '`Richard Bachman`';

```
@BOOK{king1978f,
        AUTHOR = {Stephen~Edwin King},
        TITLE = {Night Shift},
        NOTE = {[Collection of 20 short stories] * english
                [Recueil de 20~nouvelles] * french
                [Titre de la traduction fran\c{c}aise : Danse macabre] * french
                [Titel der deutschen \"{U}bersetzung: Nachtschicht] * german},
        PUBLISHER = {Doubleday \&~C\textsuperscript{o}},
        YEAR = 1978.
        LANGUAGE = english}
```

Figure 5: Two language switches, too.

◇ when a language identifier is repeated, in which case another language switch begins. For example, look at Figure 4. There are two switches, the first specifying a choice between English and French, the second a more complete choice among English, French and German, the default language being English in both cases. If this entry is used to produce a German-speaking reference, that will result in using the default value for the first switch and an *ad hoc* value for the second:

> [1] Stephen Edwin King. *The Stand*. Doubleday & C°, New-York, 1990. The Complete and Uncut Edition. Abgekürzt Auffassung im Jahre 1978 erschienen.

In order to make language switches more readable, we recommend end-users to begin each of them with the default value. For example, look at Figure 5: the `NOTE` field contains two language switches: the first specifying a choice between English and French, defaulting to English, the second a choice between French and German, without default. Even if MLBIBTEX is able to process the `NOTE` field from Figure 5 without any trouble, we think that it should be put down using this more readable form:

```
NOTE = {[Collection of 20~short stories]
        [Recueil de 20~nouvelles] * french
        []
        [Titre de la traduction fran\c{c}aise : Danse macabre] * french
        [Titel der deutschen \"{U}bersetzung: Nachtschicht] * german}
```

A language switch with a default language may occur anywhere except:

◇ within a `LANGUAGE` field, which *must* be a language identifier if defined,

◇ or within `CROSSREF` or `KEY` fields.

In MLBIBTEX's Version 1.1, a language switch with a default language cannot occur within `AUTHOR` or `EDITOR` fields, either. We plan to implement this last feature in Version 1.2, because it can be useful for the transliteration of person names originating from a language using a non-Latin alphabet. Since this transliteration is generally phonetic, it may be different from one language to another, as shown by the following example:

```
AUTHOR = {[Арам Ильич Хачатурян] * russian
```

```
@BOOK{gibson1986,
        AUTHOR = {William Gibson},
        TITLE = {Burning Chrome and Other Stories},
        NOTE = {[Titre de la traduction fran\c{c}aise :
                 \emph{Grav\'{e} sur chrome}] ! french
                [Titel der deutschen \"{U}bersetzung: \emph{Cyberspace}] ! german},
        PUBLISHER = {Victor Gollancz, Ltd.},
        YEAR = 1986,
        LANGUAGE = english}
```

Figure 6: Example of language switch without default language.

```
          [Aram Il'yich Khachaturian] * english
          [Aram Ilyich Khatchatourian] * french
          [Aram Iljitsch Chatschaturjan] * german},
...
LANGUAGE = ...
```

It is for this reason—related to the possible use of other alphabets—that we allow language switches within fields such as:

ADDRESS	JOURNAL	SCHOOL
BOOKTITLE	ORGANIZATION	SERIES
INSTITUTION	PUBLISHER	TITLE

In the same way, the date may be expressed using different calendars, which is why MLBIBTEX's future versions will probably allow language switches with a default language within fields such as MONTH or YEAR.

SWITCH WITHOUT DEFAULT LANGUAGE It is expressed by the following syntax:

$$\texttt{[}\mathit{string}_0\texttt{] ! }\mathit{idf}_0\ \texttt{[}\mathit{string}_1\texttt{] ! }\mathit{idf}_1\ \ldots\ \texttt{[}\mathit{string}_n\texttt{] ! }\mathit{idf}_n \qquad (2)$$

where n, string_0, string_1, ..., string_n, idf_0, idf_1, ..., idf_n have the same meaning as in Expression (1). MLBIBTEX processes it as follows:

- ◇ it behaves exactly like a language switch with '*' if there exists i ($0 \leq i \leq n$) such that the reference's language is equal to idf_i, that is, Expression (2) yields string_i;
- ◇ otherwise, Expression (2) is replaced by an empty string, and a warning message is emitted by MLBIBTEX.

An example is given in Figure 6. When using the plain bibliography style, MLBIBTEX will produce the following references, after processing by LATEX:

- ◇ when the reference's language is French:

  [1] William Gibson. *Burning Chrome and Other Stories*. Victor Gollancz, Ltd., 1986. Titre de la traduction française : *Gravé sur chrome*.

- ◇ when it is German:

```
@BOOK{king1981i,
        AUTHOR = {Stephen~Edwin King},
        TITLE = {[Danse macabre] : french},
        PUBLISHER = {Everest House},
        YEAR = 1981,
        MONTH = jul,
        LANGUAGE = english}
```

Figure 7: Example of a language change.

[1] William Gibson. *Burning Chrome and Other Stories*. Victor Gollancz, Ltd., 1986. Titel der deutschen Übersetzung: *Cyberspace*.

◇ and an empty note will be put otherwise:

[1] William Gibson. *Burning Chrome and Other Stories*. Victor Gollancz, Ltd., 1986.

A language switch without default language may occur anywhere except within following fields:

AUTHOR	JOURNAL	TITLE
BOOKTITLE	KEY	YEAR
CROSSREF	LANGUAGE	
EDITOR	MONTH	

Language change
It is expressed by the following syntax:

$$\texttt{[}\textit{string}\texttt{]}\ \texttt{:}\ \textit{idf} \tag{3}$$

where *string* is a string of characters, and *idf* a language identifier. It expresses conformity to other typographical conventions and can be used to hyphenate foreign words: here 'foreign' means 'belonging to a language different from the value of the LANGUAGE field'. For example, the language change given within the TITLE field in Figure 7 ensures that the title, using French words for an American book, will be hyphenated correctly if need be.

A language change may occur anywhere except within the following fields:

CROSSREF	LANGUAGE	YEAR
KEY	MONTH	

In MLBIBTEX's Version 1.1, a language change cannot occur within AUTHOR or EDITOR fields either, but we plan to implement this feature in Version 1.2.

Using square brackets as syntactic tokens
In MLBIBTEX, the use of square brackets does not interfere with the different meanings of braces. It is known that in BIBTEX (cf. [21, § 13.5.2]), the two following field specifications:

```
{The Eyes of the Dragon}     "The Eyes of the Dragon"
```
(4)

are equivalent. Using braces to surround the whole of a field—like in the expression at the left—is needed when this field contains the double-quote character:

```
{"For the Love of Barbara Allen"}
```

In fact, a double-quote character, as part as a field's value, must be surrounded by braces. The following value:

```
"{"For the Love of Barbara Allen"}"
```

is correct. Anyway, these two field specifications are equivalent, too:

```
{{The Eyes of the Dragon}}   "{The Eyes of the Dragon}"          (5)
```

but Expressions (4) make possible for this title to be non-capitalised, whereas Expressions (5) tells BIBTEX to consider the string `The Eyes of the Dragon` as it is[12]. In the same way, the following specifications are equivalent in MLBIBTEX:

```
{[Firestarter] [Charlie] * french [Feuerkind] * german}
"[Firestarter] [Charlie] * french [Feuerkind] * german"          (6)
```

and so are the following four:

```
{{[Firestarter] [Charlie] * french [Feuerkind] * german}}
{[{Firestarter}] [{Charlie}] * french [{Feuerkind}] * german}
"{[Firestarter] [Charlie] * french [Feuerkind] * german}"
"[{Firestarter}] [{Charlie}] * french [{Feuerkind}] * german"    (7)
```

but Expressions (6)—resp. (7)—are analogous to Expressions (4)—resp. (5)—with respect to possible capitalisation.

Square brackets must be balanced, unless they are put in math mode—that is, enclosed between two '`$`' characters—in which case they do not have any meaning for MLBIBTEX. A double-quote character belonging to a field's value must be surrounded by braces or square brackets. If square brackets are used like in this example:

```
TITLE = {The Girl Who [Loved] Tom Gordon},
```

MLBIBTEX believes that they surround the default part of a language switch with '`*`', the '`*`' character followed by the default language being omitted. So the result will be:

The Girl Who Loved Tom Gordon

in any case.

If users wish to insert them within the value of a field[13], the right square bracket must be followed by the empty string:

```
NOTE = {Brilliant\ldots\ A delight to read [\ldots] {} A true original.
        (\emph{Sunday Times})}
```

[12]This rule does not hold if the enclosed opening brace is followed by a TEX or LATEX command. For example, '`{{\relax The Eyes of the Dragon}}`' produces the same outputs as one of Expressions (4): cf. [21, § 15.5.2] and [28, Ch. 24] about the `\relax` command. This way, MLBIBTEX behaves exactly like BIBTEX.

[13]Let us recall that in typography, square brackets are used to enclose editorial interpretations, corrections, explanations, ... cf. [11, § 5.128–5.132]. '[...]' means that some words have been skipped.

The same convention holds if users wish to print square brackets followed by '`*`', '`!`' or '`:`' characters:

```
{Read [Trails in Darkness] {} * by Robert~Erwin Howard}
```

From our point of view, the two drawbacks of its convention are:

◇ if users want to insert a left square bracket only, or a right square bracket only, the solution is to use the LaTeX command `\symbol`:

```
{\symbol{91}Count Zero}        for '['
{Virtual Light\symbol{93}}     for ']'
```

◇ the same 'trick' holds when users wish square brackets to be nested, which is not allowed by MlBibTeX:

```
{Read [\symbol{91}Eons of the Night\symbol{93}] {} by Robert~Erwin Howard}
```

How the language is determined

MlBibTeX considers a **bibliography's language**. Version 1.1 can process documents written:

◇ with the default configuration of LaTeX, that is, without any multilingual package, in which case, the bibliography's language is supposed to be `english`;

◇ by using the babel package, in which case the bibliography's language is the default language, that is, the last option when this package is loaded[14].

Concerning other mulilingual packages, the bibliography's language will be `french` or `german` when MlBibTeX works with the french [19] and german [41] packages, which is planned for Version 1.2.

If the bibliography style used is reference-dependent, the reference's language is given by the `LANGUAGE` field for each entry. If it is document-dependent, the reference's language is always the bibliography's language, but MlBibTeX considers the value given by the `LANGUAGE` field for the following information fields:

```
AUTHOR       EDITOR     PUBLISHER
BOOKTITLE    JOURNAL    TITLE
```

which are processed according to suitable typographical conventions when there is no language switch. To explain this behaviour, let us consider Entry `king1998d` in Figure 8, and let us assume that we would like to refer it in a document written in French. If the bibliography style is reference-dependent, the corresponding reference will be wholly written in English and the result, in the .bbl file, will look like—cf. [7, § 12.2] about the `otherlanguage` environment—:

```
\bibitem{king1998d}
\begin{otherlanguage}{english}Stephen~Edwin King. \emph{Bag of Bones}.
Scribner Book Company, September 1998. Translated into French and German.
\end{otherlanguage}
```

If the bibliography style is document-dependent, only the fields supposed to be in English will be processed according to English typographical conventions, that is, the `TITLE` field, but not the `PUBLISHER` field because there is a language switch with no

[14] This default language is given by the `\bbl@main@language` command, provided by the babel package: cf. [7, § 12.2].

```
@BOOK{king1998d,
        AUTHOR = {Stephen~Edwin King},
        TITLE = {Bag of Bones},
        NOTE = {[Translated into French and German] * english
                [Titre de la traduction fran\c{c}aise : Sac d'os] * french
                [Titel der deutschen \"{U}bersetzung: Sara] * german},
        PUBLISHER = {[Scribner Book Company] * english
                     [Premi\'{e}re \'{e}dition am\'{e}ricaine] * french},
        YEAR = 1998,
        MONTH = sep,
        LANGUAGE = english}
```

Figure 8: Yet another example of language switch.

common part. (In addition, let us recall that MLBIBTEX's Version 1.1 processes the AUTHOR field exactly like BIBTEX, that is, it does not switch over to another language for this field.) So, the result in the .bbl file will look like:—cf. [7, § 12.2] about the \foreignlanguage command—:

```
\bibitem{king1998d}
Stephen~Edwin King. \foreignlanguage{english}{\emph{Bag of Bones}}.
Premi\'{e}re \'{e}dition am\'{e}ricaine, septembre 1998. Titre de la
traduction fran\c{c}aise : Sac d'os.
```

If some fields $F_0, \ldots, F_n$ ($n \geq 0$) are provided by an entry E accessed by means of a CROSSREF field, these fields $F_0, \ldots, F_n$ are processed by considering the value of the LANGUAGE field for E.

Whenever it has to switch over to another language, MLBIBTEX checks that this language will be known when LATEX processes the .bbl file. If not, this other language is replaced by the bibliography's language[15]. This behaviour is different from what happens if direct commands originating from the babel package are used within the value of a BIBTEX field. For example:

```
{\iflanguage{frenchb}{Jessie}{\iflanguage{german}{Das Spiel}{Gerald's Game}}}
```

—cf. [7, § 12.2] about the \iflanguage command—will work only if the corresponding reference belongs to the bibliography of a document using the babel package with at least the frenchb and german options, which can be an actual drawback if .bib files are shared out among several people.

As an example, the bibliography of this article has been put with a reference-dependent bibliography style.

Issue of Compatibility

Due to the huge number of .bib files already written, MLBIBTEX *should* be able to work with them. We ought to have written 'has to' instead of 'should', but syntactic

[15]So, for the example above, we assumed that the babel package was loaded in the LATEX document, with at least the english and frenchb (or french) options.

◇ Commands:

month names	ordinal numbers	other keywords
\bbljan	\bblfirsto	\bbland
\bblfeb	\bblsecondo	\bblchap
\bblmar	\bblthirdo	\bbled
\bblapr	\bblfourtho	\bbledby
\bblmay	\bblfiftho	\bbledn
\bbljun	\bblst	\bbleds
\bbljul	\bblnd	\bblin
\bblaug	\bblrd	\bblmasterthesis
\bblsep	\bblth	\bblno
\bbloct		\bblof
\bblnov		\bblp
\bbldec		\bblphdthesis
		\bblpp
		\bbltechrep
		\bblvol

The commands for month names implement the abbreviations known by BIBTEX and MLBIBTEX:

```
jan   feb   mar   apr   may   jun   jul   aug   sep   oct   nov   dec
```

◇ The `bblquotedtitle` allows a title to be quoted:

```
\begin{bblquotedtitle}%
Nesnesiteln\'{a} lehkost byt\'{\i}
\end{bblquotedtitle}
```

– with respect to English style: “The Unbearable Lightness of Being”
– w.r.t. French style: « L'insoutenable légèreté de l'être »
– w.r.t. German style: „Die Unerträgliche Leichtigkeit des Seins“
– and other styles in interface with MLBIBTEX.

By the way, notice the use of the “`%`” character—which opens a comment—in the example above, just after “`\begin{bblquotedtitle}`”. This aims to bypass the end-of-line character, because LATEX considers it as a space character (*cf.* [32, § 2.2.1]). Another way to avoid any undesirable spacing is to glue the opening of this environment “`\begin{bblquotedtitle}`” and the beginning of the quoted section together, without any space character:

```
\begin{bblquotedtitle}Neuromancer
\end{bblquotedtitle}
```

Table 1: Additional LATEX commands used in bibliographies generated by MLBIBTEX.

conventions have changed... In the same way, some end-users of BIBTEX developed their own bibliography styles, and these styles *should* be able to be used with our version. ‘Should be able to be used...’, too. Here is what we describe precisely below.

Roughly speaking, the whole distribution of MLBIBTEX consists of:

◇ an executable program, called ‘MLBIBTEX’;

◇ a file defining some additional LATEX commands[16] given in Table 1;

[16] Most of them are used in bibliography styles generated by the makebst program, and in interface with the babel package (cf. [13] or [21, § 13.9]). In this case, they have to be defined in a file called babelbst.tex.

◇ some bibliography style files that can be used in addition to the standard bibliography style files. Some begin with the command:

```
REFERENCEDEPENDENT
```

so they put the reference-dependent approach into action. Otherwise, they are document-dependent.

Any existing bibliography style file should work with MLBIBTEX, provided it works with BIBTEX's current implementation[17]. Due to our conventions, these existing bibliography style files are supposed to be document-dependent.

As far as we know, there are only two bug cases when BIBTEX (.bib) files are used with MLBIBTEX:

◇ unbalanced square brackets,

◇ nested square brackets,

—see Subsection *Using square brackets as syntactic tokens*—but we think that they must be very rare in real situations. Anyway, MLBIBTEX knows any field name BIBTEX knows, including names put in compatibility with SCRIBE [42]. If a 'new-fashioned' style is used for files processed by current BIBTEX, there is a bug case:

◇ double-quote character surrounded by square brackets, but not by braces, for example:

```
"Virtual ["Light"]"
```

in the same way, that should hardly ever happen. Otherwise, these files should be processed without any bug but obviously the result may look somewhat strange.

The additional bibliography style files can be used with BIBTEX's current implementation, provided that the `REFERENCEDEPENDENT` command is removed and the LATEX commands given in Table 1 are defined when the .bbl file is processed. So they behave in a 'standard' way. Besides, MLBIBTEX uses two environment variables:

`MLBIBINPUTS` when it searches for a .bib file,
`MLBSTINPUTS` ..... .bst file,

the corresponding environment variables used by BIBTEX's current implementation being `BIBINPUTS` and `BSTINPUTS`. These two sets of environment variables can be given suitable values, so end-users can ensure that 'new-fashioned' files are processed only by MLBIBTEX.

Some Words about the Implementation

In order to be able to master a program in constant progress, we chose to develop MLBIBTEX from scratch, even if we confess that we often consulted the source files of current BIBTEX to get as much experience as possible.

[17]Bug reports or improvement suggestions will be welcome at hufflen@lifc.univ-fcomte.fr. In addition, we will progressively put all the bug reports and any additional information onto the Web page `http://lifc.univ-fcomte.fr/PEOPLE/hufflen/texts/mlbibtex/mlbibtex/mlbibtex.html`.

MLBIBTEX is written in the C programming language [26], since it has become standard and is efficient and widely available on many systems. Besides, its use allowed us to get access to many development tools, especially GNU[18] tools. For example, the scanner and parser have been developed using the GNU scanner and parser generators: flex and bison, corresponding to lex and yacc within 'standard' UNIX [33].

In fact, we adopted an approach of *reverse engineering*, since we recovered design form analysing source files and documentation. But this study allowed us to put a precise modular decomposition into action. This decomposition yielded a precise terminology to name functions and variables, in order to ease the possible improvement of some modules.

Conclusion

Developing the first version of MLBIBTEX was a real challenge for us, since we personally missed this kind of multilingual tool quite often. We also have been very interested in the problems raised by extending BIBTEX's grammar[19]. Concerning ergonomics, we think that our proposals are user-friendly and will need as few adaptations as possible when end-users put BIBTEX files according to MLBIBTEX conventions. But we do not have actual feedback, that is why we presently consider our program as a prototype, being about to belong to the LATEX *legacy* [44], and independent of BIBTEX's future version described in [39]. Our goal is to be fully able to perform some experiments... and other experiments. For this reason, we have preferred a step-by-step approach. So we could change our syntactic conventions if it appears to be preferable.

Besides, we did not forget that many tools are based on BIBTEX[20], and we think that adaptations should be slight if developers of such tools would like to make them conformant to MLBIBTEX, if need be.

On another subject, many extensions of MLBIBTEX are planned:

- ◇ a version based on Unicode,
- ◇ an extension allowing users to define sorting programmes with respect to the lexicographical order associated with a particular language[21],
- ◇ the extension of the language used in bibliography style files, in order to ease file inclusions and avoid the duplication of identical parts from one bibliography style to another.

[18] Recursive acronym for '**G**NU's **N**ot **U**nix'. This project aims to develop *free software*. For more details, see the Web page `http://www.gnu.org`.

[19] In fact, the use of square brackets as syntactic elements originates from our study of the CAMEL citator [6, 5].

[20] For example, bib2bib and bibtex2html, described in [17]. A list of tools based on BIBTEX can be found at the Web page `http://www.ecst.csuchico.edu/~jacobsd/bib/formats/bibtex.html`, maintained by Danna Jacobsen.

[21] Let us consider the Spanish language as an example: words beginning with '`ll...`' are alphabeticised after other words beginning with '`l...`', and before words beginning with '`m...`', that is:

`la...` `ly...` `lla...` `m...`

```
<the_axiom>             ::=   <information_list> ;
<information_list>      ::=   /* empty */ | <information> <information_list> ;
<information>           ::=   <comment_information> |
                              <preamble_information> |
                              <string_information> |
                              <entry_information> ;
<comment_information>   ::=   "@COMMENT" bibtex_comment ;
<preamble_information>  ::=   "@PREAMBLE" <bibtex_value> ;
<string_information>    ::=
    "@STRING" ("{" bibtex_identifier "=" <bibtex_expression> "}" |
               "(" bibtex_identifier "=" <bibtex_expression> ")") ;
<entry_information>   ::=
    entry_keyword ("{" label <after_label> "}" | "(" label <after_label> ")") ;
<after_label>         ::=   /* empty */ | "," <field_list> ;
<field_list:          ::=   /* empty */ | <field_nelist> ;
<field_nelist>        ::=   <field> | <field> "," <field_nelist> ;
<field>               ::=   <field_name> "=" <bibtex_expression> ;
<bibtex_expression>   ::=   <concatenation_list> ;
<concatenation_list>  ::=   <bibtex_value> | <bibtex_value> "#" <concatenation_list> ;
<bibtex_value>        ::=   bibtex_identifier | natural_number | mlbibtex_atomic_value ;
```

Table 2: Grammar of BIBTEX (and MLBIBTEX).

And finally, we also plan a comparative study of the multilingual features of MLBIBTEX and those provided by XML and XSL[22], which could help us develop further versions of MLBIBTEX.

Annex: mlbibtex's Syntax

Now we outline our syntax precisely. We plan to update this description at each syntax change, in order to ease the writing of tools associated with MLBIBTEX.

MLBIBTEX's parser—which could be used as BIBTEX's parser—is based on the grammar given in Table 2. This grammar is expressed with a formalism close to BNF[23], that is:

- ◇ for each nonterminal symbol, enclosed by '`<`' and '`>`', the expression following the '`::=`' sign and terminated by '`;`' states how it can be expanded: for example, the non-terminal symbol `<the_axiom>` can be expanded only to the non-terminal symbol `<information_list>`;
- ◇ if a non-terminal symbol expands to an empty expression, we emphasise that by putting a C-like comment `/* empty */`;
- ◇ the nonterminal symbol from which the whole grammar is derived, is `<the_axiom>`;

[22] 'e**X**tensible **M**arkup **L**anguage' and 'e**X**tensible **S**tyle **L**anguage'. A good introduction to these languages is [20], in French, and an updated version is available in English: see the Web page `http://webcast.cern.ch/Projects/WebUniversity/AcademicTraining/Goossens/`.

[23] **B**ackus-**N**aur **F**orm. Readers unfamiliar with this formalism can refer to [33] for more details. A more general reference about these techniques is [2].

`bibtex_comment` All the characters are skipped, until next '`@`' character.

`bibtex_identifier` A *constituent* (see below), followed by zero or more constituents or digits.

`constituent` A letter (alphabetical character) or one of the following characters:

```
' ~ @ # $ % ^ & + * - / _ . : ; ? ! < > [ ] \
```

`entry_keyword` Syntactically, the '`@`' character, followed by zero or more constituents or digits, but `@COMMENT`, `@PREAMBLE` and `@STRING` are not entry keywords. Here are the entry keywords used most often:

`@ARTICLE`	`@CONFERENCE`	`@MANUAL`	`@PHDTHESIS`
`@BOOK`	`@INBOOK`	`@MASTERSTHESIS`	`@PROCEEDINGS`
`@BOOKLET`	`@INCOLLECTION`	`@MISC`	`@TECHREPORT`
`@COLLECTION`*	`@INPROCEEDINGS`	`@PATENT`*	`@UNPUBLISHED`

The entry keywords asterisked are not standard, they might occur within some `.bib` files.

`field_name` Syntactically, a letter, followed by zero or more constituents or digits. Here are the fields used more often:

`ABSTRACT`*	`EDITION`	`LCCN`*	`SCHOOL`
`ADDRESS`	`EDITOR`	`LOCATION`*	`SERIES`
`AFFILIATION`*	`HOWPUBLISHED`	`MONTH`	`SIZE`*
`ANNOTE`	`INSTITUTION`	`MRNUMBER`*	`TITLE`
`AUTHOR`	`ISBN`*	`NOTE`	`TYPE`
`BOOKTITLE`	`ISSN`*	`NUMBER`	`URL`*
`CHAPTER`	`JOURNAL`	`ORGANIZATION`	`VOLUME`
`CONTENTS`*	`KEY`	`PAGES`	`YEAR`
`COPYRIGHT`*	`KEYWORDS`*	`PRICE`*	
`CROSSREF`	`LANGUAGE`*	`PUBLISHER`	

The field names asterisked are not used in standard bibliography styles, but can be used by some additional styles or in some LaTeX2ϵ packages.

`label` A non empty-sequence of constituents or digits, the '`{`' and '`}`' characters being also allowed.

`natural_number` A non-empty sequence of digits. Negative numbers or numbers with '`.`' are not allowed.

Table 3: Lexical tokens of MlBibTeX, except for field's values.

◇ expressions enclosed by two double quote characters are reserved words and symbols,

◇ symbols not surrounded by braces are terminal, that is, they are tokens read by the scanner,

◇ the '`|`' sign means an alternative,

◇ parentheses are used to override precedence.

We describe the tokens common to BibTeX and MlBibTeX in Table 3, whereas the description of the grammar ruling the possible values for MlBibTeX's fields are given in Table 4.

Let us recall that:

◇ the reserved words of BibTeX and MlBibTeX are read by a case insensitive scan-

```
<mlbibtex_atomic_value>     ::=   '"' <no_double_quote>* '"' | '{' <between_braces> '}' ;
<no_double_quote>           ::=
    <almost_any> |
    '$' <any_math>* '$' |
    '[' <between_square_brackets>* ']' <after_square_brackets> |
    '{' <between_braces> '}' ;
<between_square_brackets>   ::=
    <almost_any> | '"' | '{' <between_square_brackets> '}' | '$' <any>* '$' ;
<after_square_brackets>   ::=
    /* empty */ |
    '*' <bibtex_identifier> |
    '!' <bibtex_identifier> |
    ':' <bibtex_identifier> |
    '{' '}' ;
<between_braces>   ::=   <no_double_quote> | '"' ;
```

where:

`<almost_any>` is for any characters—including the space, tabulation, or end-of-line character—but braces, square brackets, '`"`' and '`$`';

`<any_math>` is for any character—including the space, tabulation, or end-of-line character—but '`$`';

`'...'` states a character that actually appears within the value;

`<...>*` means 'zero or more occurrences of `<...>`'.

Table 4: Lexical grammar of MLBIBTEX values.

ner[24]: for example, '`@BOOK`', '`@book`', '`@Book`', '`BoOk`' are suitable for the entry type `BOOK`;

- ◇ on the contrary, the values of BIBTEX and MLBIBTEX fields—that is, the possible values for the `bibtex_atomic_token` token—keep their own case;
- ◇ BIBTEX and MLBIBTEX ignore any text that is not inside an entry, so the `@COMMENT` command, used to put any texts outside commands, is not really necessary;
- ◇ from a syntactic point of view, some labels used to refer BIBTEX or MLBIBTEX entries, may be accepted by them, but be wrong arguments of the LATEX commands `\bibitem` and `\cite`;
- ◇ space, tabulation, and end-of-line characters between syntactic categories are irrelevant in the specification given in Table 2, but are relevant in Tables 3 and 4.

Acknowledgements

Thanks to Oren Patashnik for his valuable comments about this article, and Simon Pepping who proof-read it.

[24] On the contrary, LATEX is case sensitive.

References

[1] AFNOR Z44-045 : *Documentation — Références bibliographiques — Contenu, forme et structure*. Norme disponible auprès de l'AFNOR, voir `http://www.afnor.fr`. Décembre 1987.

[2] Alfred V. Aho, Ravi Sethi and Jeffrey D. Ullman: *Compilers, Principles, Techniques and Tools*. Addison-Wesley Publishing Company. 1986.

[3] ANSI/NISO Z39.71–1999: *Holding Statements for Bibliographic Items*. See `http://www.ansi.org`. 1999.

[4] Peter Antman: *Oxford Style Package*. Version 0.4. 1997. See CTAN: `biblio/bibtex/contrib/oxford/`.

[5] Frank G. Bennett, Jr: *The* Law *Module for the* Camel *Bibliography Engine*. July 1995. See `http://www.loria.fr/services/tex/english/bibdex.html`.

[6] Frank G. Bennett, Jr: *User's Guide to the* Camel *Citator*. July 1995. See `http://www.loria.fr/services/tex/english/bibdex.html`.

[7] Joannes Braams: *Babel, a Multilingual Package for Use with* LaTeX*'s Standard Document Classes*. Version 3.7. February 2001. See CTAN: `macros/latex/required/babel/babel.dvi`.

[8] British Standards 1629:1989: *Recommendation for References to Published Materials*. See `http://bsonline.techindex.co.uk`. 1989.

[9] British Standards 5605:1990: *Recommendations for Citing and Referencing Published Materials*. See `http://bsonline.techindex.co.uk`. 1990.

[10] Judith Butcher: *Copy-Editing. The Cambridge Handbook for Editors, Authors, Publishers*. 3rd edition. Cambridge University Press. 1992.

[11] *The Chicago Manual of Style*. The University of Chicago Press. The 14th edition of a manual of style revised and expanded. 1993.

[12] *Code typographique. Choix de règles à l'usage des professionnels du livre*. Fédération nationale du personnel d'encadrement des industries polygraphiques et de la communication, Paris. 17e édition. 1993.

[13] Patrick W. Daly: *Customizing Bibliographic Style Files*. Version 3.2. February 1999. Part of LaTeX' distribution.

[14] Bernard Desgraupes : LaTeX. *Apprentissage, guide et référence*. Vuibert Informatique, Paris. 2000.

[15] Duden: *Rechtschreibung der deutschen Sprache und der Fremdwörter*. 19. Auflage. Bibliographisches Institut, Mannheim. 1986.

[16] Michael J. Ferguson: "A Multilingual TeX". *TUGboat*, Vol. 6, no. 2, p. 57–58. July 1985.

[17] Jean-Christophe Filliâtre and Claude Marché: BibTeX*2*HTML*: A Translator of* BibTeX *bibliographies into* HTML. February 2001. See `http://www.loria.fr/services/tex/english/outils.html`.

[18] Daniel Flipo: *A babel Language Definition File for French*. Version v1.5e.

March 2001. See `http://www.tex.ac.uk`.

[19] Bernard GAULLE : *Notice d'utilisation du style french multilingue pour* LATEX. Version pro V5.01. Janvier 2001. Voir sur CTAN : `loria/language/french/pro/french/ALIRE.pdf`.

[20] Michel GOOSSENS : « XML et XSL : un nouveau départ pour le Web ». *Cahiers GUTenberg*, Vol. 33–34, p. 3–126. Novembre 1999.

[21] Michel GOOSSENS, Frank MITTELBACH and Alexander SAMARIN: *The* LATEX *Companion*. Addison-Wesley Publishing Company, Reading, Massachusetts. 1994.

[22] *Hart's Rules for Composers and Readers at the University Press*. Oxford University Press. 39th edition. 1999.

[23] Jean-Michel HUFFLEN : « Typographie : les conventions, la tradition, les goûts, ... et LATEX ». *Cahiers GUTenberg*, Vol. 35–36, p. 169–214. Mai 2000.

[24] Jean-Michel HUFFLEN : « Vers une extension multilingue de BIBTEX ». *Cahiers GUTenberg*, Vol. 39–40, p. 23–38. Mai 2001.

[25] Jean-Michel HUFFLEN, Denis RŒGEL et Karl TOMBRE : *Guide local* (LA)TEX *du LORIA. Millésime 1998*. Rapport technique 98–R–214, LORIA. Septembre 1998.

[26] Brian W. KERNIGHAN and Denis M. RITCHIE: *The C Programming Language*. 2nd edition. Prentice Hall. 1988.

[27] Donal Ervin KNUTH and Pierre MACKAY: "Mixing Right-to-Left Texts with Left-to-Right Texts". *TUGboat*, Vol. 8, no. 1, p. 14–25. April 1987.

[28] Donald Ervin KNUTH: *Computers & Typesetting. Vol. A: the TEXbook*. Addison-Wesley Publishing Company, Reading, Massachusetts. 1984.

[29] Helmut KOPKA: LATEX—*Band I: Einführung*. 2. Auflage. Addison-Wesley Longmann. 1997.

[30] Helmut KOPKA: LATEX—*Band II: Ergänzungen mit einer Einführung in* METAFONT. 2. Auflage. Addison-Wesley Longmann. 1997.

[31] Helmut KOPKA: LATEX—*Band III: Erweiterungen*. Addison-Wesley Longmann. 1997.

[32] Leslie LAMPORT: LATEX. *A Document Preparation System. User's Guide and Reference Manual*. Addison-Wesley Publishing Company, Reading, Massachusetts. 1994.

[33] John LEVINE, Tony MASON and Doug BROWN: *lex & yacc*. 2nd edition. O'Reilly. October 1992.

[34] *Lexique des règles typographiques en usage à l'Imprimerie Nationale*. Imprimerie Nationale. 1990.

[35] Wenzel MATIASKE: *Multilinguale Zitierformate*. Oktober 1995. Siehe CTAN: `macros/latex/contrib/supported/mlbib/`.

[36] NTS TEAM and Peter BREITENLOHNER: *The ε-TEX Manual*. February 1998.

Part of LaTeX' distribution.

[37] Oren Patashnik: *Designing* BibTeX *styles*. February 1988. Part of LaTeX' distribution.

[38] Oren Patashnik: BibTeX*ing*. February 1988. Part of LaTeX' distribution.

[39] Oren Patashnik: "BibTeX 1.0". *TUGboat*, Vol. 15, no. 3, p. 269–273. September 1994.

[40] John Plaice and Yannis Haralambous: *Draft Documentation for the* Ω *System*. March 1998. See `http://www.loria.fr/services/tex/english/moteurs.html`.

[41] Bernd Raichle: *Die Makropakete „german" und „ngerman" für* LaTeX*2ε, LaTeX 2.09, Plain-TEX and andere darauf Basierende Formate*. Version 2.5. Juli 1998. Im Software LaTeX.

[42] Brian Reid: Scribe *Document Production System User Manual*. Technical Report, Unilogic, Ltd. 1984.

[43] Christian Rolland : LaTeX *par la pratique*. Éditions O'Reilly. Octobre 1999.

[44] Chris A. Rowley: "The LaTeX Legacy. 2.09 and All That". In: *Annual* ACM *Symposium on Principles of Distributed Computing*. Newport, Rhode Island. August 2001.

[45] *The Unicode Standard Version 3.0*. Addison-Wesley. February 2000.