In Praise of Simplicity: Lexicographic Lightweight Markup Language

Vladimír Benko
Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics
E-mail: vladimir.benko@juls.savba.sk

Abstract

Our paper presents a simple markup language – *Lexicographic Lightweight Markup Language* (*LLML*) that has been used for almost the last three decades in the framework of two dozen lexicographic projects carried out by our Institute, as well as in several projects carried out in co-operation with commercial dictionary publishers. While initially trying to solve the problem of insufficient computing power of early *MS-DOS*-based personal computers in early 1990's only, *LLML* is even today the central component of lexicographic workstations our lexicographers work with. Central components of the *LLML* syntax are introduced and exemplified by a sample entry from the *Dictionary of the Contemporary Slovak Language* (*SSSJ*). The final part of the paper describes in short some components of the *LLML*-aware toolbox, i.e., programs that are used in our Institute during compilation, validation, proofreading and typesetting of the respective entries. Some of these tools, however, are just a "bonus", and "low-cost" projects could do even without them.

Keywords: lexicographic data representation, lightweight markup language, XML

1 Introduction

One of the typical features of lexicographic projects is that they usually take many years – in the case of multivolume works, even decades – to complete. The developments in the area of information technologies, on the other hand, are extremely fast. This usually means that several generations of IT components may change during the life cycle of a project.

Today it is mostly taken for granted that dictionary data (at least in the framework of large-scale lexicographic projects) should be represented as "structured text", i.e., encoded in XML and complying to some standard, such as ISO 1951 (2007) or TEI-P5 (2018). The advantages of this approach have described in several works (cf. Derouin & Le Meur 2008).

Nevertheless, we are aware of many projects that – for various reasons – do not use XML and represent dictionary date as "formatted text", i.e., using a standard word processor, such as Microsoft Word (e.g., Apresyan, 2014). Some of them do so just because they are continuing to use the same method as when the project was started years ago, and do not have the resources to change it. The main argument in such a case is usually that "XML is too expensive", having in mind not only the price of the appropriate software – an XML-aware text editor, or even a full-fledged Dictionary-Writing System (*DWS*) –, but also additional "human costs", i.e. salaries for IT specialists necessary to support the software, as well as training costs for the lexicographic team. The Microsoft Word format, on the other hand, seems to be "cheap" – the necessary software is usually available anyway, and almost no additional education for the lexicographers is necessary.

There are, of course, many disadvantages to such an approach, with probably the most important being that it is difficult to enforce uniformity in dictionary entry structure, and such data is almost impossible to validate.

On the other hand, it is also worth noting that traditional lexicographers' "mental model" of a dictionary entry maps directly to typefaces and font styles, and working with a *DWS* requires "mental switching" between two models: a "tree-structured" and a "formatted" one. This involves additional mental burden that – especially the older members of lexicographic terms – by not be easily accepted easily.

In our paper we thus introduce a type of dictionary data representation that may be considered a compromise between fully structured XML format and typographical-only format – using a markup language that is nowadays referred to as *Lightweight Markup Language* (*LML*). The most important feature of such languages is that their syntax is very simple, the data is readily comprehensible in source form, and no special software (besides a generic text editor) is needed.

2 Historical Background and Related Work

> "*Lightweight markup languages were originally used on text-only displays which could not display characters in italics or bold, so informal methods to convey this information had to be developed. This formatting choice was naturally carried forth to plain-text email communications. ... In 1986 international standard SGML provided facilities to define and parse lightweight markup languages using grammars and tag implication. The 1998 W3C XML is a profile of SGML that omits these facilities. However, no SGML DTD for most of the LMLs is known.*" (Wikipedia, 2018).

From this perspective, we can say that it was the conventions developed in e-mail (and USENET) that evolved into languages like *Markdown*[1] & *reStructuredText*[2].

Our markup language, now called *Lexicographic Lightweight Markup Language* (*LLML*), has also a fairly long history, and its first version was developed in 1990 during the project of retro-digitalization of a one-volume monolingual Slovak dictionary that was later republished (KSSJ, 1997). Despite its history, no (English) paper on *LLML* has yet been published. In 1992 this system was introduced internationally , at the Budapest *COMPLEX '92* Conference (Benko, 1992). However, as it did not appear in the Proceedings, and so only the Conference participants were informed about our efforts. Our paper at the *Slovko 2001* Conference (Benko, 2001), on the other hand, was in Slovak only, so became "hidden" to the international lexicographic community.

Meanwhile, the language (with only minor modifications) has been used in the preparation of more than 20 monolingual and bilingual dictionaries, and is currently used in the framework of the multivolume *Dictionary of the Contemporary Slovak Language* (three volumes already published, five more to come; SSSJ 2006, 2011, 2016).

3 LLML

We believe that the main point of *LLML* can be described by the keyword "simple". The language elements can be learned within the first day of use, even by novice lexicographers, and a DIN A5 "cheat sheet" typically contains almost everything they need to know. Moreover the *LLML* type of markup can also be considered "natural", as punctuation marks are traditionally used to enhance the structure of highly complex texts.

1 https://daringfireball.net/projects/markdown/

2 http://docutils.sourceforge.net/rst.html

The main elements of the *LLML* syntax can be summarized as follows:

- A dictionary entry is represented a single block of text, entries are separated by a blank line. Though the length of individual lines is not specified by the language itself, it is recommended to keep lines relatively short.
- A line starting with an exclamation mark is used as an entry identifier; its syntax is project dependent. For our retro-digitization projects this has carried information on page and column numbers; in some early projects where dictionary entries had first been compiled on traditional paper slips, these slip numbers were indicated.
- A line starting with a question mark (optionally preceded by whitespace) is considered as a "comment", i.e., will not appear in the final output. Comments are useful for communication between the entry author and editor(s), and provide a device to record editorial decisions.
- "Structural breaks", such as new sense, phraseology zone or run-on, begin on a new line indented by two spaces.
- The respective "information fields" of the entry are indicated by a small set of punctuation and special characters. The actual syntax may slightly differ from one project to another. Table 1 shows the actual syntax used within the *SSSJ* project.

Table 1: Main *LLML* Syntax Elements (*SSSJ* Dialect)

LLML Element	*Default rendering*
"headword"	**headword**
"headwordˆ1"	**headword[1] (headword with index)**
"%substandard headword"	**substandard headword**
"*incorrect headword"	*incorrect headword
"~crossref headword"	**crossref headword**
[pronunciation]	[pronunciation]
|*PoS label|	pos label
|other label|	other label
<etymlology>	⧘etymology⧘
'example text'	*example text*
[*reference]	[reference]
{1}, {2}, …	**1.**, **2.**, … (sense numbers)
{M}, {T}, …	□, □, … (special symbols indicating "structural breaks" in entry structure)
(unmarked)	(unmarked) … definitions, explanations, etc.

As the *LLML* syntax is very similar to that of programming languages, by using text editor featuring user definable syntax highlighting the respective information fields in colors, the lexicographer's work becomes even more user-friendly. We hope that the reader can appreciate its legibility in Figure 1, showing a screenshot of an example entry as displayed by the Notepad++ editor using a custom "language definition".

Identification and comment lines are displayed in gray, so that the entry text itself is highlighted. The cyan vertical line at the right margin indicates the suggested line length, and other colors highlight the respective structure elements.

```
309  !10260
310  "légia" -ie |pl. G| -íí |D| -iám |L| -iách |*ž.| <lat.>
311    {1} |hist., voj.| najväčšia a hlavná bojová jednotka
312  rímskej armády, ktorú tvorilo niekoľko kohort: 'rímske
313  légie'; 'l. cisára Marca Aurelia'; 'veliť légiám'; 'poraziť
314  légie'; 'privolať na pomoc légie z Východu'; 'Rím,
315  zákonodarca a vládca sveta, sídlo nepremožiteľných
316  légií.' [*Anton Hlinka]
317    {2} |voj.| dobrovoľná vojenská jednotka: 'veliteľ,
318  príslušník légie'; 'v roku 1916 vstúpil do
319  československých légií v Rusku'
320    {T} |hist.| 'Biela légia' protikomunistické ilegálne
321  hnutie pôsobiace na Slovensku v r. 1948 - 1955; 'légia
322  Kondor' nacistická vojenská jednotka vyslaná Hitlerom
323  do Španielska počas španielskej občianskej vojny
324    {M} 'Cudzinecká légia' francúzska špeciálna jednotka
325  tvorená zahraničnými dobrovoľníkmi, súčasť francúzskej
326  armády, v súčasnosti nasadzovaná v rámci mierových
327  humanitných operácií vojenských síl OSN
328    {3} |expr.| veľký počet niečoho, množstvo ľudí, dav:
329  'l. básnikov'; 'Mačiek sa vrátila celá légia, lebo sa
330  niekde nakotili a späťnásobili.' [*Š. Žáry]; 'Federálne
331  a miestne vlády zamestnávajú celé légie inšpektorov,
332  ktorí udeľujú vysoké pokuty.' [*HN 2003]
333          ?a IKPMV
334          ?b NK
335          ?? Anton Hlinka - Ozvena slova 1 - Blahozvesť v horizonte
336          ?? ľudskej skúsenosti 2, 1996
337          ?? Štefan Žáry - Apeninský vzduch, 1984
338          ?? Teofil Klas - Putovanie do Loreta, 1999
```

Figure 1: Lexicographic Lightweight Markup Language (*LLML*) text editor screenshot (*Notepad++*)

4 Data Validation

The *LLML* approach to lexicographic data representation does not allow for full-scale data validation, but essential data checks can be performed. A special validation parser had to be written from the very beginning of using *LLML* in order to detect errors with regard to the "well-formedness" of the dictionary data, such as unmatched syntactic structures and non-sequential appearance of sense numbers. With the advent of text editors with color syntax highlighting the former problem became less acute, as unmatched "tags" are usually immediately apparent by "spoiled" colors. The tool, nonetheless, proved to be quite useful and is being gradually enhanced to include checking for the presence and/or absence of whitespace around punctuation, presence of suspicious special characters, etc. The error report produced by the validation parser always contains a detailed error message, including the respective excerpt from the input file indicating the affected line numbers, such as those at Figure 2.

```
157 ******** po čiarke má byť medzera
   157:   {1} |lit.| literárny žáner o živote svätých,v ktorom

228-231 ******** číslo významu mimo poradia
   228:   {1} vlastnosť toho, čo sa zakladá na legendách,
   229: vymyslených, neskutočných veciach: 'l. príbehu';
   230: 'literatúra s črtami legendárnosti'
   231:   {3} vlastnosť toho, čo je nezabudnuteľné, výnimočné,
```

Figure 2: Error report generated by the Validation Parser

The first message (line 157 of the source file) indicates a missing space after a comma, and the second message (lines 228 to 231) states that a sense number out of sequence has been encountered. The lexicographer can usually detect the exact cause of the issue from the error report itself, without the need to study the larger context of the source file.

5 *LLML* Toolbox

In this section we want to mention some other parts of our *LLML*-aware toolbox that are needed to cover the whole process of dictionary creation:

(1) "Paragraph grep" – an open-source *perl* script used to extract dictionary entries. This provides for the creation of ad-hoc lists of entries based on regular expressions,
(2) "Paragraph sort" – custom sort using marked headword as sort keys. A simple modification of a standard (quicksort-based) sort.
(3) Converter to proofreading format (an MS-Word document in RTF format retaining original line breaks and comments, and indicating entry identifiers and line numbers (Figure 3). The line numbers are convenient in subsequent editing of the dictionary data.

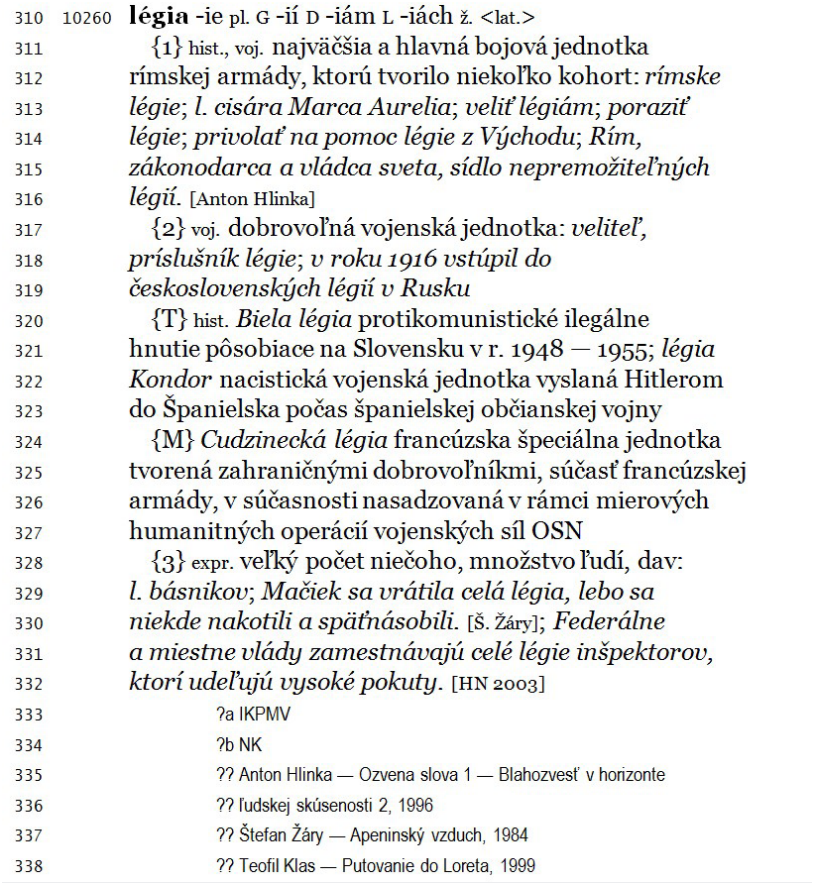
310 10260 **légia** -ie pl. G -íí D -iám L -iách ž. <lat.>
311 {1} hist., voj. najväčšia a hlavná bojová jednotka
312 rímskej armády, ktorú tvorilo niekoľko kohort: *rímske*
313 *légie*; *l. cisára Marca Aurelia*; *veliť légiám*; *poraziť*
314 *légie*; *privolať na pomoc légie z Východu*; *Rím,*
315 *zákonodarca a vládca sveta, sídlo nepremožiteľných*
316 *légií.* [Anton Hlinka]
317 {2} voj. dobrovoľná vojenská jednotka: *veliteľ,*
318 *príslušník légie*; *v roku 1916 vstúpil do*
319 *československých légií v Rusku*
320 {T} hist. *Biela légia* protikomunistické ilegálne
321 hnutie pôsobiace na Slovensku v r. 1948 — 1955; *légia*
322 *Kondor* nacistická vojenská jednotka vyslaná Hitlerom
323 do Španielska počas španielskej občianskej vojny
324 {M} *Cudzinecká légia* francúzska špeciálna jednotka
325 tvorená zahraničnými dobrovoľníkmi, súčasť francúzskej
326 armády, v súčasnosti nasadzovaná v rámci mierových
327 humanitných operácií vojenských síl OSN
328 {3} expr. veľký počet niečoho, množstvo ľudí, dav:
329 *l. básnikov*; *Mačiek sa vrátila celá légia, lebo sa*
330 *niekde nakotili a späťnásobili.* [Š. Žáry]; *Federálne*
331 *a miestne vlády zamestnávajú celé légie inšpektorov,*
332 *ktorí udeľujú vysoké pokuty.* [HN 2003]
333 ?a IKPMV
334 ?b NK
335 ?? Anton Hlinka — Ozvena slova 1 — Blahozvesť v horizonte
336 ?? ľudskej skúsenosti 2, 1996
337 ?? Štefan Žáry — Apeninský vzduch, 1984
338 ?? Teofil Klas — Putovanie do Loreta, 1999

Figure 3: Proofreading format (line breaks and comments retained, line numbers indicated)

(4) Converter to typesetting format (entry identifiers and comments deleted, Figure 4). This is in fact the only "compulsory" part of the toolbox. It works in two phases: firstly the *LLML* data is converted to "presentation type" of *XML* format, i.e., indicating the respective typefaces the information fields are mapped into. The second step can use any standard tool for *XML* conversion, in our case generating an *RTF* format that can be imported into the respective publishing system.

légia *-ie* pl. G *-ií* D *-iám* L *-iách* ž. ⟨lat.⟩ **1.** hist., voj. ▸ najväčšia a hlavná bojová jednotka rímskej armády, ktorú tvorilo niekoľko kohort: *rímske légie; l. cisára Marca Aurelia; veliť légiám; poraziť légie; privolať na pomoc légie z Východu; Rím, zákonodarca a vládca sveta, sídlo nepremožiteľných légií.* [Anton Hlinka] **2.** voj. ▸ dobrovoľná vojenská jednotka: *veliteľ, príslušník légie; v roku 1916 vstúpil do československých légií v Rusku* □ hist. *Biela légia* protikomunistické ilegálne hnutie pôsobiace na Slovensku v r. 1948 – 1955; *légia Kondor* nacistická vojenská jednotka vyslaná Hitlerom do Španielska počas španielskej občianskej vojny □ *Cudzinecká légia* francúzska špeciálna jednotka tvorená zahraničnými dobrovoľníkmi, súčasť francúzskej armády, v súčasnosti nasadzovaná v rámci mierových humanitných operácií vojenských síl OSN **3.** expr. ▸ veľký počet niečoho, množstvo ľudí, dav: *l. básnikov; Mačiek sa vrátila celá légia, lebo sa niekde nakotili a späťnásobili.* [Š. Žáry]; *Federálne a miestne vlády zamestnávajú celé légie inšpektorov, ktorí udeľujú vysoké pokuty.* [HN 2003]

Figure 4: Final format (typeset entry)

6 Conclusion and Further Work

From the early 1990s, when our first *LLML*-based projects started, we considered it as something temporary that should (and will) eventually be replaced by a more sophisticated representation. During the SGML period, however, our computing equipment was firstly not powerful enough to implement it, and secondly the *SGML*-aware software was also far beyond what we could afford to buy. With the advent of XML, the situation has changed dramatically, both in terms of the computing power of our equipment and availability of affordable software tools.

We have observed the efforts of introducing the *XML* technology at our partner institutions that, surprisingly, turned out to be not as straightforward and easy as we would imagine. Though the computing power of modern workstations is no longer the main problem, another scarce resource appeared: the *XML*-based projects require much more (human) IT support. This is probably why we are still reluctant to switch our main project to it. We realize, however, that the day is approaching, and that better interoperability will most likely be the motivating main reason.

One of our anonymous reviewers noted that "… *I accept that you use the framework what you describe, but I doubt that it could be recommended for other (new) dictionary projects as a standard instead of XML*". This is naturally difficult to argue against, yet there is at least one area where the *LLML*-based approach can be of an advantage even today: the dictionary retro-digitization projects involving manual proofing of OCR-ed material. According to our experience, it is convenient here to split the process into two separate phases, with the first aimed to achieving only the "typographical identity" – an explicit, *simple* markup can ease the whole process significantly.

References

Benko, V. (1992). *Late Computational Support for a Dictionary Project.* Presentation at the COMPLEX '92 International Conference. Budapest, Hungary. (unpublished).

Benko, V. (2001). *Počítačová podpora lexikografických projektov – retrospektívny pohľad*. (Computational Support of Lexicograpphis Projects – A Retrospective View). In: Jarošová, A. (ed) *Slovenčina a čeština v počítačovom spracovaní*. Proceedings of the Slovko 2001 Conference. Bratislava: VEDA.

ASRYa (2014). Apresyan, Yu. (Ed.) Aktivnyj slovar' russkogo yazyka. Tom 1. A – B. Yazyki slavyanskoj kul'tury, Moskva.

Derouin, Marie-Jeanne and Le Meur, André (2008). ISO-Standards for Lexicography and Dictionary Publishing. Proceedings of the 13th EURALEX International Congres. Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, 2008. pp. 663 – 668, ISBN 978-84-96742-67-3.

KSSJ (1997). Krátky slovník slovenského jazyka. Red. J. Kačala – M. Pisárčiková. 3. dopl. a preprac. vyd. Bratislava: Veda 1997. 943 s. ISBN 80-224-0464-0

ISO (2007). ISO 1951:2006(en), Presentation/representation of entries in dictionaries — Requirements, recommendations and information.

SSSJ I (2006). *Slovník súčasného slovenského jazyka. A – G*. Hl. red. K. Buzássyová – A. Jarošová. Bratislava: Veda, vydavateľstvo Slovenskej akadémie vied 2006. 1134 p. ISBN 978-80-224-0932-4

SSSJ II (2011). *Slovník súčasného slovenského jazyka. H – L*. Ved. red. A. Jarošová – K. Buzássyová. Bratislava: Veda, vydavateľstvo Slovenskej akadémie vied 2011. 1087 p. ISBN 978-80-224-1172-1

SSSJ III (2016). *Slovník súčasného slovenského jazyka. M – N.* Ved. red. A. Jarošová, Bratislava: Veda, vydavateľstvo SAV 2015. 1100 s. ISBN 978-80-224-1485-2.

TEI (2018). TEI Consortium, eds. TEI P5: *Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium. http://www.tei-c.org/Guidelines/P5/

Wikipedia (2018). Lightweight markup language. https://en.wikipedia.org/wiki/Lightweight_markup_language. [15/06/2018]

Acknowledgement

This work has been, in part, financially supported by the Slovak VEGA Grant Agency, Project No. *2/0017/17*.