WEB LEXICOGRAPHY FOR AND BY NON-TECH PEOPLE

Dmitri Dmitriev

Institute of Linguistic Research, Russian Academy of Sciences, St Petersburg
dmitri@globbie.net

ABSTRACT

Globbie Neologia is a set of open source software tools, GUI and methodologies for building a collaborative network of lexicographical sites. The ultimate goal is to make the process of lexicon compilation less complicated and more accessible to nontech people. The crowdsourced contributions are evaluated on-the-fly by the NLP service to conform to adopted formalism. The resulting linguistic descriptions are ready for use in digital applications, including crosslanguage search engines, machine translation, language learning etc.

Index Terms— web lexicography, crowdsourcing, NLP, SVG based GUI

1. INTRODUCTION

Successful crosscultural mediation and communication remains a challenge at present. It is especially true for the digital web presence of minority languages and cultures, due to the known complexity of natural language in general, and obvious shortage of resources for language study and development — scholars, software developers, language enthusiasts, equipment, operating funds etc. The resolution of the pioneer International Conference named "Languages of minorities in computer technologies: experience, goals, and perspectives" that took place in YoshkarOla (Russia, 25-27 April 2011) attracts attention to the following problem:

"The speakers of minority languages, i.e., languages with a comparatively small number of speakers and limited technological, infrastructural and other possibilities for use) of Russia and many other countries are experiencing serious problems in connection with the presence of the so-called "information" or "digital divide," described as a "limit on a social group's abilities as a result of its lack of access to modern means of communication in its native language."

The continued preservation of the situation whereby the majority of the languages of Russia are not represented in the electronic information space, despite the latter's rapid development, will only exacerbate this situation" [1].

The Resolution explores the importance of availability of the specific language tools as follows:

- electronic dictionaries under open licenses;
- creation and introduction of spell-checking systems;
- packages of educational software localized in minority languages;
- search engines with complete support for the minority languages, as well as localization of search engine interfaces in these languages;
- educational computer programs, electronic textbooks and open-access repositories of multimedia materials in minority languages, for use in education to teach such languages;
- machine translation systems.

The international community has developed a longstanding, consistent strategy to meet these needs. We can mention here the continuing efforts of the International Conference on Minority Languages (ICML) [2], and more specifically—the International Speech Communication Association (ISCA) Special Interest Group (SIG, http://www.isca-speech.org) on Speech And Language Technology for MInority Languages (SALTMIL) [3].

The Globbie Neologia project was started in 2011 to develop basic linguistic technologies and provide methodological assistance to local language communities in North Eurasia. The project is based on first-hand experience accumulated through many projects in lexicography, software localization, machine translation in Russian, Tatar, Mari, Abkhaz, Armenian, and other languages of the former Soviet Union.The distinctive features of suggested solution are as follows:

1. constant evaluation of the input data through the use of natural language processing (NLP),
2. attractive GUI with vector graphics,
3. intelligent faceted search,
4. dynamic concept ontologies (Concept Maps),
5. distributed network architecture with focus on crowdsourcing.

2. EXISTING PROJECTS AND TOOLS

It is vital to keep in mind that many language development projects have long been carried out by the local groups of language enthusiasts, regional Internet communities that

accumulate language resources, localize the interfaces of popular web services (search engines, social networks) and software (Mari, Komi, Tatar, Udmurt, Chuvash, Yakut, and other), cf. http://komikyv.ru/, http://www.sakhatyla.ru, http://www.buryadxelen.com, http://www.udmurt.info, http://www.chuvash.org, http://dict.marlamuter.ru/.

The Yoshkar-Ola Conference (April 2011) clearly identified the need for better cross-communication between these projects and for the scholarly and technical guidance by professionals.

First and foremost, we need to note the experience of SIL International. For decades, SIL International has been a recognized leader in producing software for linguistic research, lexicon compilation and translation. SIL has put significant time and effort into developing tools that are beneficial for local language communities. Software tools like Flex, Lexique Pro, WeSay, Webonary are used by many field linguists in many parts of the world. Some of these tools are aimed to publish dictionaries on the web, while others can be used to merge the linguistic materials from desktop applications into a common online repository. However, the task of crowdsourcing the lexicon brings a new challenge: the tools must be simple enough to use by non-tech people who often lack linguistic education.

Wiktionary is by all means the most popular web project today that relies on crowdsourcing to exploit the power of networking. Wiktionary uses the MediaWiki engine that is free server-based software which is licensed under the GNU General Public License (GPL) [4]. Even though MediaWiki provides effective technical infrastructure for joining the efforts of language enthusiasts worldwide, its textual markup does require certain skills from its contributors. Wiktionary pages use MediaWiki's wikitext format, so that users without knowledge of HTML or CSS can edit them. Let us take a quick look at an excerpt from the wiki entry of the English word “format”:

```
===Noun===
{{en-noun}}

# The [[layout]] of a [[document]].
# {{context|hence|lang=en}} The [[form]] of [[presentation]] of something.
#: ''The radio station changed the '''format''' of its evening program.''
# {{context|computing|lang=en}} A [[file type]].
```

One is likely to agree that it is not trivial for a lay person to use this kind of notation to make a meaningful and valid contribution. In the context of under-resourced social environment it becomes a substantial hindrance for attracting free volunteers.

Moreover, professional lexicographers will find this formalism utterly incomplete from their scholarly perspective. Lexicographical entry in standard published dictionaries in many cases represents a deep hierarchical data structure. It contains nested sense definitions, examples of usage, references to corpora, derivation pointers, cross-references and many other pieces of information. The TEI Guidelines of the P5 lexicographical encoding make this introductory remark: “Both typographically and structurally, print dictionaries are extremely complex” [5]. One brief example from TEI schema for dictionaries illustrates the complicated mechanism of cross-anchors within nested sense descriptions:

```
<sense n="4">
 <usg type="dom">imprim</usg>
 <def>Donner a (une ligne) une longeur convenable au moyen de
 <ref target="#blanc-2.1.3">blancs (2, sens 1, 3)</ref>
 </def>
</sense>
<entry xml:id="blanc" n="2">
<!-- ... -->
 <sense n="1">
<!-- ... -->
  <def xml:id="blanc-2.1.3">...</def>
<!-- ... -->
 </sense>
<!-- ... -->
</entry>
```

Furthermore, if we look at the Wiki format as computer linguists we might notice that it can not be directly used in automatic text analysis or synthesis. The textual definitions in Wiktionary are intended to be read by humans, therefore they are written as plain sequences of characters (#PCDATA in XML) and are neither syntax trees nor conceptual graph structures.

The above mentioned considerations have motivated and challenged us to search for a better solution.

3. GOALS AND PRINCIPLES OF USER-FRIENDLY WEB LEXICOGRAPHY

The key difference of the Globbie Neologia project is that it aims to help create formalized digital descriptions of lexical codes that must be ready-to-use in systems of automatic language analysis, but to do so in a user-friendly way with no requirements to study formal logical markup or sophisticated interfaces. The cornerstones of our solution are as follows:

1. Provide a dialogue-based, user-friendly interface for non-tech audience, supporting intelligent recognition and semantic restatement of every user's input.
2. Using robust linguistic parsing technology to interpret the encoded input data, removing the burden of formalized input of data, required by rigid DBMS, providing easy-to-use tools for enriching the lexicons and knowledge bases.
3. Intelligent search engine: consistent use of cross-language faceted search.
4. Communication and collaboration assistance for the language communities with the focus on North Eurasia.
5. Content produced by local communities automatically becomes ready for both human perception (to be used in language learning), and for derivative computer products, such as spell checkers, interactive search engines, translation tools etc.

In order to hide the complexity of lexicon compilation behind the scenes, we had to substantially redesign the software "back-office" architecture.

4. SOFTWARE ARCHITECTURE

4.1. Main components

The following diagram shows the interaction of the main software components of Globbie Neologia. All server-side services are written in ANSI C and run on Linux.

4.2. Message-oriented services

ZeroMQ messaging library is used for all inter-communications of services. Each service runs as a separate process and provides an asynchronous interface: it reads the queue of incoming messages and sends replies through a delivery service. To provide maximum productivity and throughput, each service uses horizontal partitioning where tasks are evenly distributed among multiple agents.

4.3. Web server

Web content is served by a standard Apache server with extra Globbie module that is used to provide connectivity between the HTTP-requests and inner ZeroMQ messaging protocol.

4.4. Data storage

Oracle Berkeley DB is used for disk storage. To meet the performance requirements, much of the frequently used data is cached in RAM.

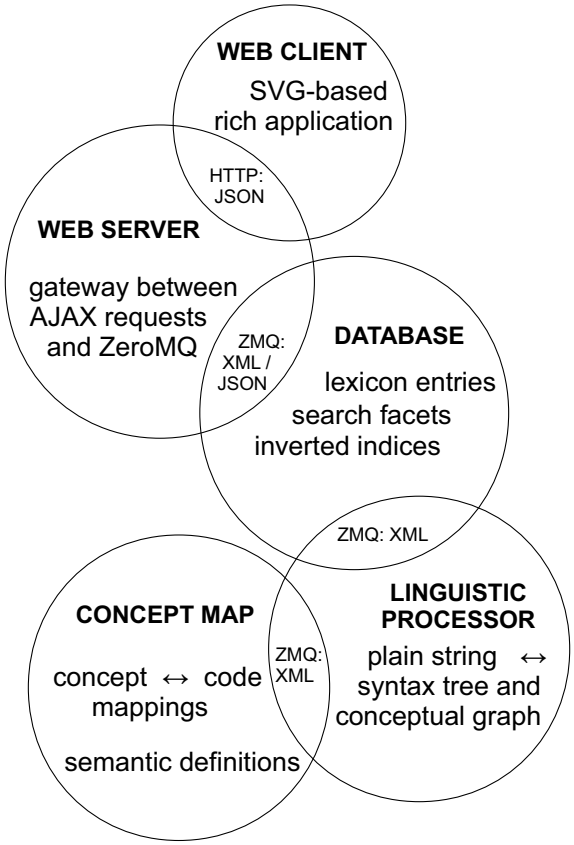

Figure 1: Globbie Neologia main software components

4.5. Linguistic processor

The natural language processing service in Globbie Neologia is based on the in-house OOmnik linguistic processor written in ANSI C. It takes any sequence of bytes as its input and applies various parsing strategies to build a valid conceptual graph with interconnected concepts.

Linguistic processor is activated each time a free string of text is given as input to the database. Lexicographical definitions are verified against a controlled lexicon to be consistent, meaningful and accurate. Possible interpretations are presented to the user so that she could confirm that this was indeed the intended meaning of the definition. This technique of machine-assisted editing helps maintain a consistent conceptual network shared by cooperating contributors. Concept Map service is used to synchronize this formalized set of concepts.

The current parsing speed is around 200K of input data per sec on a single stock Pentium core.

Another important thing about the linguistic processor is that all language coding systems exist as external XML-based modules. This allows flexibility in extending the system without recompiling software components. Every effort was taken to keep the linguistic core of a processor language-independent.

The details of the processor's architecture, linguistic formalisms and public API are to be described elsewhere. A quick illustration of the XML-output produced by OOmnik as a result of linguistic parsing and semantic analysis is given below.

Original input of the Russian free text lexicographical definition:

"Венская конвенция об охране озонового слоя"

English equivalent:

"Vienna convention for the protection of the ozone layer"

Automatic interpretation by OOmnik (layers involved: byte/character decoding, morphological segmentation, syntax analysis, semantic domain attribution, sense disambiguation):

```
<conc>
 <head>
  <conc name="КОНВЕНЦИЯ / CONVENTION" parent="ДОГОВОР /
    AGREEMENT"/>
 </head>

 <specs>
  <spec name="RELATION">
   <conc name="ВЕНА / VIENNA" parent="ГЕО НАЗВАНИЕ /
    GEO NAME"/>
  </spec>
  <spec name="CONTENT">
    <conc>
     <head>
      <conc name="ОХРАНА / PROTECTION " parent="НАПАДЕНИЕ-
        ЗАЩИТА / ATTACK-DEFENCE"/>
     </head>

     <specs>
      <spec name="DIRECT OBJECT">
       <conc>
        <head>
         <conc name="СЛОЙ / LAYER" parent="ЧАСТЬ ЦЕЛОГО /
           PART OF A WHOLE"/>
        </head>
        <specs>
         <spec name="RELATION">
          <conc name="ОЗОН / OZONE" parent="КОНКРЕТНЫЕ ГАЗЫ /
           SPECIFIC GAS"/>
         </spec>
        </specs>
       </conc>
      </spec>
     </specs>
    </conc>
   </spec>
  </specs>
</conc>
```

5. RECORD FORMATS

XML has been widely accepted as a flexible format for representing lexicon descriptions. There are many flavors (open or in-house) of XML-schemas out there to suit lexicographers' multifold needs. XML markup in Globbie Neologia is used only as an internal storage format. There is no manual interaction with XML by contributors.

5.1. TEI P5

TEI P5 standard encoding can be used as input format to the lexical database of Globbie Neologia. Lexicographical materials can also be exported from the database to plain text files using this schema.

5.1. Inner DB format

Inside the database XML input is split into segments to provide quick random access to pieces of information without parsing large chunks of XML in RAM. Below is given an excerpt from the "web-technology" Russian entry.

```
<usg_list>
<usg id="U:0">
 <def_t><t>Компьютерные технологии, используемые во
  <link ref="ВСЕМИРНАЯ ПАУТИНА">Всемирной паутине</link>,
  <link ref="ИНТЕРНЕТ">Интернете</link>;
  <link ref="ИНТЕРНЕТ-ТЕХНОЛОГИЯ">интернет-технологии</link>.
  </t> </def_t>
 <cit_list>
 <cit><text_t><t>В последнее время все больше внимания уделяется
   технологиям реального времени, в том
   числе, в первую очередь, технологии «всемирной паутины»..
   (далее в тексте – веб-технология).. Веб-технология представляет собой
   с точки зрения пользователей-учащихся совокупность открытых
   для он-лайнового доступа информационно-поисковых систем
   (веб-серверов).</t></text_t>
```

```
  <src>ИРР, 1996, 6.</src>
 </cit>
 <cit><text_t><t>Далеко не все хранители
    драгоценных сведений уже перешли на Web-технологию,
   хотя, надо признать, сегодня это наиболее динамично
   развивающийся сегмент Сети.</t></text_t>
    <src>Ин, 1996, 39.</src>
 </cit>
 <cit><text_t><t>Изобретатель веб-технологии Тим Бернерс-Ли не раз
   имел основания быть недовольным политикой компаний
   Microsoft и Netscape.</t></text_t>
  <src> Э-А, 1999, 4.</src>
 </cit>
</cit_list>
```

6. INTERLINGUA DEFINITIONS

Wordsense definitions are stored in both human-readable format and also as structured conceptual interlingua. In our view, "interlingua is not an universally accepted classification of the world, but rather a kind of intermediary pseudo-Person with its own type of mental organization that can map its inner conceptual ontology with reality that we know. It is just like an interpreter who stands between you and a foreign speaker. You pass all your ideas through the mind of your interpreter" [6]. Conceptual graphs in Globbie Neologia are stored internally in a compact format that is similar to S-expressions of Lisp.

7. SVG BASED GUI

SVG is an XML-based language that is used to describe two-dimensional vector graphics and text. It is an open W3C standard and is currently supported in most modern browsers. Use of vector graphics makes it easier to design innovative interfaces that can be much more expressive than traditional ones, eg. radial presentation of lexicon items instead of blocked rectangulars.

8. CASE FOR RUSSIAN

Even though one can hardly think of Russian language as an example of under-resourced language, it can be labelled under-resourced when it comes to semantic digital representation. Traditional Russian dictionaries are not yet fully converted to a format that is satisfactory for machine-learning or for use in intelligent search engines.

In 2014 we launch a special web portal named **Neologia.ru** that is specifically designed to monitor lexical innovations of Russian language using the power of crowdsourcing. This project received financial support from the Russian Foundation for Humanities (grant № 12-04-12065).

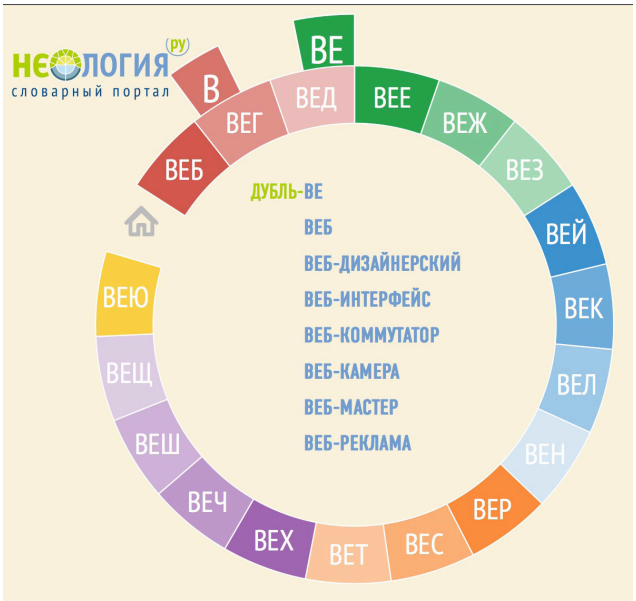

Figure 2: SVG-based GUI in Neologia.ru

9. CONCLUSION

Web crowdsourcing has tremendous potential nowadays. Quite a few enthusiasts who are passionate about their language and culture and would be happy to contribute to the compilation of lexicons and everyday language monitoring. Consistent, reliable and up-to-date digital lexicon is the core of every modern language technology. We believe that our technical and methodological solutions will assist many under-resourced language communities in developing a good online mother-tongue presence.

REFERENCES

[1] http://www.fennougria.ee/public/final_document-eng.pdf
[2] http://icml14.uni-graz.at/docu.html
[3] Climent Nadeu, Donncha Cróinín, Bojan Petek, Kepa Sarasola, Briony Williams. ISCA SALTMIL SIG: speech and language technology for minority languages. In proceeding of: EUROSPEECH 2001 Scandinavia, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event, Aalborg, Denmark, September 3-7, 2001.
[4] http://www.mediawiki.org/wiki/MediaWiki
[5] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html
[6] Dmitri Dmitriev, Semantic Interlingua for the Knowledge Base Creation. *Electronic Government - Workshop and Poster Proceedings of the Fourth International EGOV Conference 2005*, August 22-26, 2005, Copenhagen, Denmark. Schriftenreihe Informatik 13 Universitätsverlag Rudolf Trauner, Linz, Austria 2005, ISBN 3-85487-830-3. P. 11—18.