A Spell Checker for Esperanto

Final Project Report (December 2008 to July/November 2009)

by Bc. Marek Blahuš <muni@blahus.cz> učo 172464

This report resumes the overall results of the project “A Spell Checker for Esperanto” that have been accomplished during its development which lasted in three stages, i.e. Stage 1 (December 2008 to February 2009), Stage 2 (March 2009 to April 2009), Stage 3 (May 2009 to July 2009), plus the extra months (August 2009 to November 2009). The report resumes what has been written in checkpoint reports published after the completion of every stage, and its content is to certain extent parallel to that of the project’s website.

“A Spell Checker for Esperanto” is a project in the field of Natural Language Processing, aiming at designing and implementing a spell checking dictionary for the Esperanto language using the Hunspell spell checker, which has been financed in its recent phase by the Students' Research and Development Projects scholarship of the Faculty of Informatics at the Masaryk University in Brno, Czech Republic.

Overview of the Development

The goals of the first stage were enhancing the spell checker’s functionality and performing preliminary research on the possibilities of its integration in the OpenOffice.org office application suite.

As an outcome, the existing prototype spell checker was enhanced in several directions, in particular by extending its coverage to all official, as well as unofficial and “international” prefixes and suffixes, as described in the PMEG reference grammar. An insight into these enhancements were given in form of general descriptions of each of them along with examples of words, word groups or sentences for which the previous version of the software did not provide correct result, while their treatment were fixed in the new version. Furthermore, a system of rules were devised and implemented to cope with homely forms of personal names, and the somewhat related issue of capitalization in proper names was also dealt with, by means of closely following the reference PIV dictionary. The preliminary research on possibilities of the spell checker’s integration in OpenOffice.org has identified two dimensions of the task – technical details of the integration itself being one of them, proper licensing of the dictionary as a precondition for this being the second one.

The goals of the second stage were solving out problems related to the spell checker’s integration in the OpenOffice.org environment, and arranging its inclusion as a part of the official Esperanto language distribution of the office suite.

As an outcome, an analysis of the current state of development of three software packages in question (the OpenOffice.org office suite, its Esperanto language localization and the discussed spell checker itself), maintained by three different bodies (the OpenOffice.org community, its Esperanto translation team and the scholarship holder itself) was given. These three pieces of software depend on each another and their proper harmonization is necessary in order to produce a spell checker that may be provided as an optional and/or native tool in the office suite. As a concrete result of this discussion, a suitable approach was found (consisting particularly of making the dictionary an XML-based dictionary extension and licensing it under the triple GPL/LGPL/MPL license) and, as its implementation, the first version of the spell checker in a form readily usable in OpenOffice.org was produced.
The goals of the third stage were to set up a bug-report system and to use it for collecting feedback from users who are testing the spell checker in the environment of OpenOffice.org. Comments from the users were to be evaluated and necessary measures taken in order to remove any bugs reported.

As explained in detail in the third stage’s report, the extend of the time available for working out the goals of this stage was prolonged, based on unexpected medical emergency suffered by the scholarship’s holder, with permission from the scholarship’s administrator and to no additional costs.

As an outcome of the third stage, a working bug-report system was set up, located right at the official OpenOffice.org website. By means of it, users were made able to post their comments while the related discussions can be easily tracked. In order to speed up the process of creating a proper category for the bug reports, contact with Pavel Janík, a Czech member of the OpenOffice.org development team, was established. Since then, a number of the spell-checker’s testers have provided useful comments and requests for corrections and improvements, which were consecutively implemented in the spell checker. Yet, because of the long-time nature of the bug-report system and the fact of the spell checker being actually used by users at the moment, the debugging has become a long-time process which is certain to go on also further beyond the time span of the project’s scholarship.

At the moment, the resulting spell-checker software has been published as an easily installable extension in the official OpenOffice.org repository for Extensions, where any interested user can easily find it. As the Esperanto localization team was just working on translating the new OpenOffice.org 3 into Esperanto at the time of the project’s third phase, their general consensus has been found on integrating the new spell checker in the package once it’s Esperanto localization is ready. This was expected to happen in early 2010. Further details on the actual appearance of the new Esperanto package of OpenOffice.org and the spell checker as its integral part will be provided during the project’s official defense to be held in Brno in 2010.

As a summary: The existing prototype spell-checker resulting from the Bachelor’s thesis has been turned into a readily usable spell-checking tool for Esperanto, recognizing a plenty of the languages morphological phenomena, tested by first actual users, usable among others within the framework of the OpenOffice.org office suite, and acknowledged as an official part of the package’s Esperanto localization.

The project’s goals, as described in the scholarship application, have been fulfilled.

Public Presentations

In the course of the project’s development, it has been presented on several occasions, both to the community of experts and the general public. These presentations, apart from popularizing the project and preparing a user basis for its result, were been followed by discussions, which provided additional input and suggestions of features to be implemented.

Following is the list of these public presentations (including three made between the project’s defense as a Bachelor’s thesis and the start of the scholarship project):

- 2008-07-25 Rotterdam (NL)
  93rd World Congress of Esperanto, Meeting of Computer Linguists

- 2008-09-27 Stockholm (SE)
  E@I Grammar Checker (Lingvohelpilo) Developers Meeting
• 2008-11-22 Berlin (DE)
  Gesellschaft für Interlinguistik (Society for Interlinguistics), 18. Tagung

• 2009-02-06 Antwerpen (BE)
  La Verda Stelo, local Esperanto speakers group

• 2009-02-09 Ottignies-Louvain-la-Neuve (BE)
  Kvinfolio, local Esperanto speakers group

• 2009-11-25 Brno (CZ)
  Faculty of Informatics, Seminar of the Natural Language Processing Laboratory

• 2009-12-04 Karlova Studánka (CZ)
  Third Workshop on Recent Advances in Slavonic Natural Language Processing
  RASLAN 2009

There are also publications in which the project has been described:
