Subject: Re: Spell checking on linux
Posted by dolik.rce on Fri, 30 Apr 2010 11:53:47 GMT
View Forum Message <> Reply to Message

Hi all!

Koldo: They are all aspell dictionaries, acquired via ubuntu repositories. All of them are GPL. I didn't look at openoffice files at all.

Mindtraveller: Yes, as I mentioned before, the quality varies. Most of the small ones (with exception of russian and ukraine) are not widely used, so there was probably not much effort to build the dictionaries. If anyone can supply better wordlist, we can substitute them.

Also, the hungarian file looked suspicious from the other side it is too big - there were too much words. To compare: 12M in hungarian, compared to 4M for czech. Most of them looked like concatenated from two words, with common prefixes, but not even google could find any of these suspicious words... Anyone capable of checking them?

Mr. Ped: I was thinking about similar approach. Putting them to sf.net with GPL license and writing a download wizard into theide. As you said the problem is with end users who want to redistribute them with non-GPL apps. I guess they would have to reuse the wizard or supply their own dictionaries.

Mirek: Yes, it worked without problems. Also, since I got the wordlist directly from aspell program, it was all in UTF-8 and I didn't have to bother with the charset conversion. The entire process went smooth and fully automatically. Only exception was the hungarian - it choked my computer, 512 MB RAM was not enough, so I had to convert this one on different computer, but that is not a fault in algorithm but in my hardware

Also one more thing: I'm not sure about some of the languages codes, because in aspell the were designated with only two letter code, like "cs". For the sake of automatizaion, I expended it in such cases presuming the second part is the same as first, but in some cases it is wrong, e.g. cs-cs should be cs-cz... This should be checked/solved before making them available for download.

Best regards,
Honza