
Subject: Re: Spell checking on linux
Posted by [dolik.rce](#) on Sat, 01 May 2010 20:05:54 GMT
[View Forum Message](#) <> [Reply to Message](#)

Very interesting... The dictionaries with small sizes are actually not that wrong as you assumed. I checked the wordcounts:

- 407752 wordlist.el
- 339747 wordlist.fa
- 455264 wordlist.he
- 14268 wordlist.ku
- 12497 wordlist.nr
- 6234 wordlist.ns
- 1029 wordlist.or
- 2045 wordlist.pa
- 732571 wordlist.ru
- 181779 wordlist.uk

For comparison:

- 417350 wordlist.bg
- 4669281 wordlist.cs
- 307891 wordlist.da
- 135275 wordlist.en_US
- 629569 wordlist.fr_FR
- 12939123 wordlist.hu
- 48490 wordlist.pt_PT
- 859141 wordlist.sk_SK

After looking at those numbers for a while, I got an idea, that some of the wordlists may contain multiple entries for a single word. So I ran some of them through sort -u and here is what I got

- 135275 wordlist.en_US
- 135275 wordlist.en_US.sorted
- 732571 wordlist.ru
- 1236 wordlist.ru.sorted
- 4669281 wordlist.cs
- 4269350 wordlist.cs.sorted

Conclusion:

The Russian dictionary is very poor and the same might be true for other ones on "suspicious list". So someone will have to do some quality control... I can check the word counts, but it would be nice if someone who actually speaks the given language checked it after me.

Regards,
Honza
