Subject: Spell checking on linux
Posted by dolik.rce on Sat, 24 Apr 2010 19:00:16 GMT
View Forum Message <> Reply to Message

Hi!

Today, I decided that I should use spellchecker when writing a new page for uppweb. I knew there is one, but to my great surprise, it didn't work.

After consultation with Koldo and digging through sources, I found where is the problem: There is no dictionary in Linux releases. Well, no problem - I said to myself - let's download the dictionary from svn or sf.net. Another surprise: there is no English dictionary on sf.net and there is even no dictionaries at all in svn.

So I had to download the win installer and extract the en-us.scd file from there. I put it to ~/.upp/theide where theide is supposed to look for the dictionaries. And here comes the last surprise - it didn't help either.

Finally, after reading the sources more carefully, I noticed it is actually looking for "EN-US.scd" and that makes a great difference on case-sensitive filesystems.

So I have few proposals: Add basic (at least en-us) dictionaries to the src packages. I'll take care about debian packages.
 Put the files for all available languages on one common place for download (sf.net as it is now is fine, just add English ones please)
 Make the files work corectly on Linux - i.e. either convert the basename to uppercase OR change the code in sGetSpeller() to ...

```
pp << spell_path << ';' << getenv("LIB") << ';' << getenv("PATH") << ';';
String path = GetFileOnPath(ToLower(LNGAsText(lang)) + ".udc", pp);
if(IsNull(path))
 path = GetFileOnPath(ToLower(LNGAsText(lang)) + ".scd", pp);
if(IsNull(path))
 return false;
FileIn in(path);
...
```

 The usual complaint: This should be documented somewhere. If nobody disagrees, I'll put an article to theide docs explaining how to obtain and install additional dictionaries inside theide.

Best regards,
Honza

Subject: Re: Spell checking on linux
Posted by Mindtraveller on Sat, 24 Apr 2010 20:20:19 GMT
View Forum Message <> Reply to Message

My vote too for documenting spellchecking in U++. Honza, thank you.

Subject: Re: Spell checking on linux
Posted by koldo on Sat, 24 Apr 2010 20:29:53 GMT
View Forum Message <> Reply to Message

I agree too.

Perhaps in addition I would vote for having a secondary tar.gz compressed file including dictionaries and another for windows including SDL and MinGW.

Subject: Re: Spell checking on linux
Posted by koldo on Sat, 24 Apr 2010 22:12:17 GMT
View Forum Message <> Reply to Message

Hello Honza

Perhaps other option would be to include in Download page links to "Speller Dictionaries" folder and to SDL.

Subject: Re: Spell checking on linux
Posted by dolik.rce on Sat, 24 Apr 2010 22:40:41 GMT
View Forum Message <> Reply to Message

koldo wrote on Sun, 25 April 2010 00:12Hello Honza

Perhaps other option would be to include in Download page links to "Speller Dictionaries" folder and to SDL.
I actually like this one even better... The files are there already and anyone can just select those which he needs. "Additional resources" section with links on Download page should cover it just fine. I'll do it together with the documentation...

Honza

Subject: Re: Spell checking on linux
Posted by mirek on Mon, 26 Apr 2010 08:41:17 GMT
View Forum Message <> Reply to Message

dolik.rce wrote on Sat, 24 April 2010 15:00
[*] Make the files work corectly on Linux - i.e. either convert the basename to uppercase OR change the code in sGetSpeller() to ...

```
pp << spell_path << ';' << getenv("LIB") << ';' << getenv("PATH") << ';';
String path = GetFileOnPath(ToLower(LNGAsText(lang)) + ".udc", pp);
if(IsNull(path))
 path = GetFileOnPath(ToLower(LNGAsText(lang)) + ".scd", pp);
if(IsNull(path))
```

```
 return false;
FileIn in(path);
...
```

Fix applied.

I believe en-us dictionary should be part of .deb.

I believe it all should be documented, including dictionary creation process.

Also, note that that there is SpellerDictionaries folder on sf.net. Maybe somebody should work on getting more .scd files there

All you need is a plain text list of all correct words for language, then use uppbox/MakeSpellScd to create .scd file.

Mirek

Subject: Re: Spell checking on linux
Posted by dolik.rce on Mon, 26 Apr 2010 11:16:33 GMT
View Forum Message <> Reply to Message

luzr wrote on Mon, 26 April 2010 10:41Also, note that that there is SpellerDictionaries folder on sf.net. Maybe somebody should work on getting more .scd files there

All you need is a plain text list of all correct words for language, then use uppbox/MakeSpellScd to create .scd file.
I'm aware of the sf.net dictionaries, but the selection is rather limited I was already looking for opensource dictionaries that could be converted to scd. The best solution I found so far would be to convert aspell dictionaries. They are GPL licensed, which is not optimal, but better than nothing. I'll try to investigate how to convert them.

A part of the documentation will definitely be a list of available dictionaries with direct download links to sf.net. Also, thinking about it now, it should be rather easy to add a menu item to theide "Fetch spellcheck dictionary..." which would ask for language and try to fetch the file from sf.net and install it.

Honza

Subject: Re: Spell checking on linux
Posted by dolik.rce on Wed, 28 Apr 2010 15:27:24 GMT
View Forum Message <> Reply to Message

Hi all!

Little update: After a bit of investigation (i.e. reading the manual) I have found really simple way of converting aspell dictionaries to wordlists in suitable format: #!/bin/bash

```
DICTS=$( aspell dump dicts )

for d in $DICTS
do
 echo "Exporting $d ..."
 aspell dump master $d | aspell expand | sed 's/ /\n/g;' > wordlist.$d
done

echo "Finished!"
```

The above script will ask aspell for a list of installed dictionaries (only master ones, but can be extended to include personal dictionaries as well). Then it cycles through all of them and creates the worldists in current directory.

For list of available languages, see official download page. If I count correctly, there is something around 90 languages which can be easily converted to .scd files.

Now the question is what do we have to do to satisfy the GPL license? Is it enough to mention the dictionaries are GPL-licensed in About-box of theide and on the sf.net download page? I never really understood those licenses :-/

Regards,
Honza

Subject: Re: Spell checking on linux
Posted by dolik.rce on Thu, 29 Apr 2010 20:45:32 GMT
View Forum Message <> Reply to Message

Another update: I converted 68 dictionaries that are available from ubuntu repositories. Let's call that a standard selection The dictionaries contain data varying from poor to excellent, but I was to lazy to go through the documentation and select only the good ones. Also bad dictionary is better than none

I won't upload them anywhere just yet since I'm still unsure about the licensing. If someone wishes to test his language, just contact me. Here comes the list, the second column is filesize:
af-af.udc 428K
am-am.udc 137
ar-ar.udc 364
bg-bg.udc 231K
bn-bn.udc 467
br-br.udc 62K
ca-ca.udc 745K

cs-cs.udc 3,5M
cy-cy.udc 180K
da-da.udc 388K
de-at.udc 474K
de-de.udc 473K
de-ch.udc 472K
el-el.udc 1,5K
en-ca.udc 226K
en-gb.udc 226K
en-us.udc 226K
eo-eo.udc 158K
es-es.udc 459K
et-et.udc 556K
eu-eu.udc 408K
fa-fa.udc 1,2K
fi-fi.udc 349K
fo-fo.udc 241K
fr-fr.udc 375K
fr-ch.udc 375K
ga-ga.udc 402K
gl-gl.udc 67K
gu-gu.udc 398
he-he.udc 1,8K
hi-hi.udc 410
hr-hr.udc 368K
hu-hu.udc 4,8M
hy-hy.udc 559
is-is.udc 457K
it-it.udc 346K
ku-ku.udc 17K
lt-lt.udc 567K
lv-lv.udc 548K
ml-ml.udc 711
mr-mr.udc 406
nb-nb.udc 1,2M
nl-nl.udc 605K
nn-nn.udc 810K
no-no.udc 1,2M
nr-nr.udc 44K
ns-ns.udc 21K
or-or.udc 115
pa-pa.udc 126
pl-pl.udc 1,8M
pt-br.udc 1,4M
pt-pt.udc 106K
ro-ro.udc 727K
ru-ru.udc 8,6K
sk-sk.udc 793K

sl-sl.udc 582K
ss-ss.udc 64K
st-st.udc 21K
sv-sv.udc 139K
ta-ta.udc 211
te-te.udc 550
tl-ph.udc 43K
tn-tn.udc 31K
ts-ts.udc 67K
uk-uk.udc 6,1K
uz-uz.udc 462
xh-xh.udc 56K
zu-zu.udc 203K

Subject: Re: Spell checking on linux
Posted by koldo on Fri, 30 Apr 2010 04:06:55 GMT
View Forum Message <> Reply to Message

Hello Honza

Huge job!

About licensing, where have you got these dictionaries?

In general in every dictionary compressed file is the license in a text file. And in the same place, for example, here: http://wiki.services.openoffice.org/wiki/Dictionaries , not all licenses are the same.

Subject: Re: Spell checking on linux
Posted by Mindtraveller on Fri, 30 Apr 2010 06:10:52 GMT
View Forum Message <> Reply to Message

el-el.udc 1,5K
fa-fa.udc 1,2K
he-he.udc 1,8K
ku-ku.udc 17K
nr-nr.udc 44K
ns-ns.udc 21K
ru-ru.udc 8,6K
uk-uk.udc 6,1K

Looks like some langauge packs are almost empty. Did you try to take OpenOffice spellchecking files?

Subject: Re: Spell checking on linux
Posted by mr_ped on Fri, 30 Apr 2010 07:40:01 GMT
View Forum Message <> Reply to Message

I'm not sure how to deal with the GPL in dictionaries.

The mingw was included with GPL, although I'm afraid that's not ok (unless mingw has some exception), and I'm afraid the dictionaries would taint the whole package with GPL.

It would be easy to go around this by including just some wizard script (under BSD), which would download and convert dictionaries for the user who run it (I think this is trivial workaround for any programmer, so it would suit well our target users), but you hit the GPL issue again later when deploying the app. If you would include spell-check+dicts, you would have either to use such workaround even for final user (may work not as well as for u++ users), or be sure how to distribute the GPL dicts properly under their license. (it would be very likely OK to make them available as additional download on your site, so every user will get BSD package as deployment, and he can download any GPL stuff separately later on his own. I'm not sure if it's possible to create a legal bundle of both, I would say "no" unless you are very clever and creative.)

Subject: Re: Spell checking on linux
Posted by mirek on Fri, 30 Apr 2010 08:55:43 GMT
View Forum Message <> Reply to Message

mr_ped wrote on Fri, 30 April 2010 03:40I'm not sure how to deal with the GPL in dictionaries.

The mingw was included with GPL, although I'm afraid that's not ok (unless mingw has some exception), and I'm afraid the dictionaries would taint the whole package with GPL.

It would be easy to go around this by including just some wizard script (under BSD), which would download and convert dictionaries for the user who run it (I think this is trivial workaround for any programmer, so it would suit well our target users), but you hit the GPL issue again later when deploying the app. If you would include spell-check+dicts, you would have either to use such workaround even for final user (may work not as well as for u++ users), or be sure how to distribute the GPL dicts properly under their license. (it would be very likely OK to make them available as additional download on your site, so every user will get BSD package as deployment, and he can download any GPL stuff separately later on his own. I'm not sure if it's possible to create a legal bundle of both, I would say "no" unless you are very clever and creative.)

I guess for starters, putting them to sf.net and not making them the part of release would do.

I guess nobody wants to download all of them anyway.

Mirek

Subject: Re: Spell checking on linux

Posted by mirek on Fri, 30 Apr 2010 08:56:52 GMT
View Forum Message <> Reply to Message

P.S.: Looking at sizes (those that are evidently OK), I guess the new compression algorithm works quite well.... despite being quite primitive.

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

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 loking 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 "suspicous 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

Subject: Re: Spell checking on linux
Posted by dolik.rce on Sat, 15 May 2010 22:17:02 GMT

View Forum Message <> Reply to Message

Hi,

Here are the dictionaries, together with the original sources. Can someone put them to sf.net? Or I can do it myself, if someone grants me the rights (my sf.net account is dolik_rce).

Also, I propose to make a little change to the function finding the files:

```
Speller *sGetSpeller(int lang)
{
 static ArrayMap<int, Speller> speller;
 int q = speller.Find(lang);
 if(q < 0) {
  String pp;
  String dir = ConfigFile("scd");
  for(;;) {
   pp << dir << ';';
   String d = GetFileFolder(dir);
   if(d == dir) break;
   dir = d;
  }
  pp << spell_path << ';' << getenv("LIB") << ';' << getenv("PATH") << ';';
  String path = GetFileOnPath(ToLower(LNGAsText(lang)) + ".udc", pp);
  if(IsNull(path))
   path = GetFileOnPath(ToLower(LNGAsText(lang)) + ".scd", pp);
  if(IsNull(path))        // This is
   path = GetFileOnPath(ToLower(LNGAsText(lang).Left(2)) + ".udc", pp); // added
  if(IsNull(path))
   return false;
  FileIn in(path);
  if(!in)
   return false;
```

... It will allow the dictionaries to have just the language part of name (i.e. no country code). This way, even if the user selects country which does not have specific dictionary yet (e.g. EN-NZ) he still gets a spell checker for given language (that is en.udc for the New Zealand exmple). IMHO it is better to serve a more general dictionary than nothing.

Best regards,
Honza

Subject: Re: Spell checking on linux
Posted by mirek on Sun, 16 May 2010 06:55:21 GMT

View Forum Message <> Reply to Message

dolik.rce wrote on Sat, 15 May 2010 18:17Hi,

Here are the dictionaries, together with the original sources. Can someone put them to sf.net? Or I can do it myself, if someone grants me the rights (my sf.net account is dolik_rce).

Rights granted.

Quote:

Also, I propose to make a little change to the function finding the files: Speller *sGetSpeller(int lang)

```
{
 static ArrayMap<int, Speller> speller;
 int q = speller.Find(lang);
 if(q < 0) {
  String pp;
  String dir = ConfigFile("scd");
  for(;;) {
   pp << dir << ';';
   String d = GetFileFolder(dir);
   if(d == dir) break;
   dir = d;
  }
  pp << spell_path << ';' << getenv("LIB") << ';' << getenv("PATH") << ';';
  String path = GetFileOnPath(ToLower(LNGAsText(lang)) + ".udc", pp);
  if(IsNull(path))
   path = GetFileOnPath(ToLower(LNGAsText(lang)) + ".scd", pp);
  if(IsNull(path))        // This is
   path = GetFileOnPath(ToLower(LNGAsText(lang).Left(2)) + ".udc", pp); // added
  if(IsNull(path))
   return false;
  FileIn in(path);
  if(!in)
   return false;
```

  ... It will allow the dictionaries to have just the language part of name (i.e. no country code). This way, even if the user selects country which does not have specific dictionary yet (e.g. EN-NZ) he still gets a spell checker for given language (that is en.udc for the New Zealand exmple). IMHO it is better to serve a more general dictionary than nothing.

OK, byt do we really have those "more general" dictionaries?

Maybe we should scan for something like "en*.udc" instead?

Mirek

Subject: Re: Spell checking on linux
Posted by dolik.rce on Sun, 16 May 2010 07:33:47 GMT

View Forum Message <> Reply to Message

luzr wrote on Sun, 16 May 2010 08:55OK, byt do we really have those "more general" dictionaries?

Maybe we should scan for something like "en*.udc" instead?

Hi Mirek,

Thanks for the rights, I'll upload it today.

This idea is based on the Aspell dictionaries. Most of them have only the general form. Some, like en, are to my best knowledge combination of all words from all english dialects. There is only single case where the general dictionary does not exist (pt-pt,pt-br, no pt).

Maybe we could scan for en*.udc as a last resort after checking for the general dictionary, if the general does not exist. But if someone asks for Angolan Portuguese, how would you decide if it is better to use pt-pt or pt-br?

Honza