OpenOffice.org
Conference 2007 Barcelona

Integrated Tools for Spelling, Style, and Grammar Checking

Daniel Naber

Mindquarry
The Open Source Collaborative Software

Agenda

- About the speaker
- Languages
- Spell checking
- Thesaurus
- Grammar checking

Goal of this Talk

- **Let you know how linguistic tools in OOo work**
- **Let you know how you can improve them**

- **Note: You do not need to be a programmer for that**

About the Speaker

- **Founder and maintainer of OpenThesaurus and LanguageTool**
- **Works for Mindquarry, start-up that offers an Open Source collaboration server**

Languages

Giving a Talk about Languages...

...is difficult:

- **Number of languages worldwide:**
  6000+
- **Number of languages I speak:**
  2

Languages by Number of Speakers

1. Standard Mandarin / Standard Chinese (1200m)
2. Hindi (600m)
3. English (500m)
4. Spanish (400m)
5. French (300m)
6. Russian (280m)
7. Arabic (220m)
8. Bengali (215m)
9. Portuguese (200m)
10. German (130m)

Languages by Number of Speakers (Ctd.)

11. Japanese (127m) . . . . . .

- ...
- ...
- ...
- ??. Romansh (about 35,000 speakers)
  - Spoken in parts of Switzerland
- ??. Saterland Frisian (about 2,000 speakers)
  - Spoken in a town in Germany

Language Support by OOo

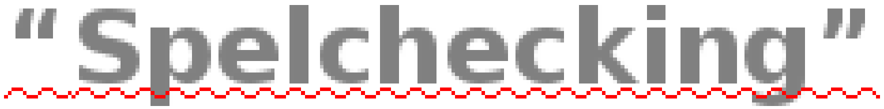
Note: This is not about i18n (the translation of the user interface)

- Spell checking: 88 languages
  - About 8 integrated
- Thesauri: 10 languages
  - About 7 integrated
- Grammar Checking: 7 languages
  - LanguageTool: about 5
  - CoGrOO: 1
  - An Gramadóir: about 3
  - Not integrated (yet?)

Sources:
http://wiki.services.openoffice.org/wiki/Dictionaries

Spell Checking

for example:



“Spelchecking”

Spot the Error

- Some real errors (try a Google search!):

- "Grammer checker wanted..."
- "Grammar checker wanted..."

- "Sorry for my bed english"
- "Sorry for my bad English"

- "That's good to now"
- "That's good to know"
  - Cannot be detected by a spell checker, as now and know are both correct words

Spell Checking

- Dictionaries included in OOo
  - See share/dict/ooo/*.dic and *.aff
- More dictionaries at http://wiki.services.openoffice.org/wiki/Dictionaries and via DicOOo
- Implementation: Hunspell
  - See http://hunspell.sourceforge.net
  - Also used by Firefox and Thunderbird soon
  - Thus work on dictionaries isn't useful for OOo only!
  - Actively developed, current developments include phonetic suggestion, morphology, ...
    - Sponsored by FSF.hu Foundation, Hungary

How does Spell Checking Work?

- **Each word of the text is compared to a large dictionary - if not known, suggest similar**
- **Works on single words, does not know context**
- **The dictionary contains base forms and flags**
- **Example:**
  talk/GZSRD
  expands to
  talk, talking, talked, talker, talkers, talks
  (use unmunch to try it out)
- **Many advanced features, see Hunspell manual**

Create a Spelling Dictionary

1. Research existing dictionaries
2. Collect words and build the .dic and .aff files
3. Announce on the Wiki page (http://wiki.services.openoffice.org/wiki/Dictionaries)
4. Send mail to lingucomponent list so it can be put on the ftp server
   installable via File -> Wizards -> Install new dictionaries

- (similar for thesaurus, only step 2 differs)

Creating a Dictionary from Scratch

- Look for free word lists
  - E.g. Scrabble community
- BootCaT: http://sslmit.unibo.it/~baroni/bootcat.html
  - Set of scripts to build a text corpus
  - Takes a seed list and queries Yahoo or Google

Maintain a Dictionary

- Learn the Unix tools:
  - diff (options: -u, -b, ...)
  - grep
  - wc
  - recode, iconv
- Work with dictionary files:
  - munch
  - unmunch
- Get help with the --help option
- Or with man <command>, e.g. man grep
- Use Unix commands under Windows with Cygwin

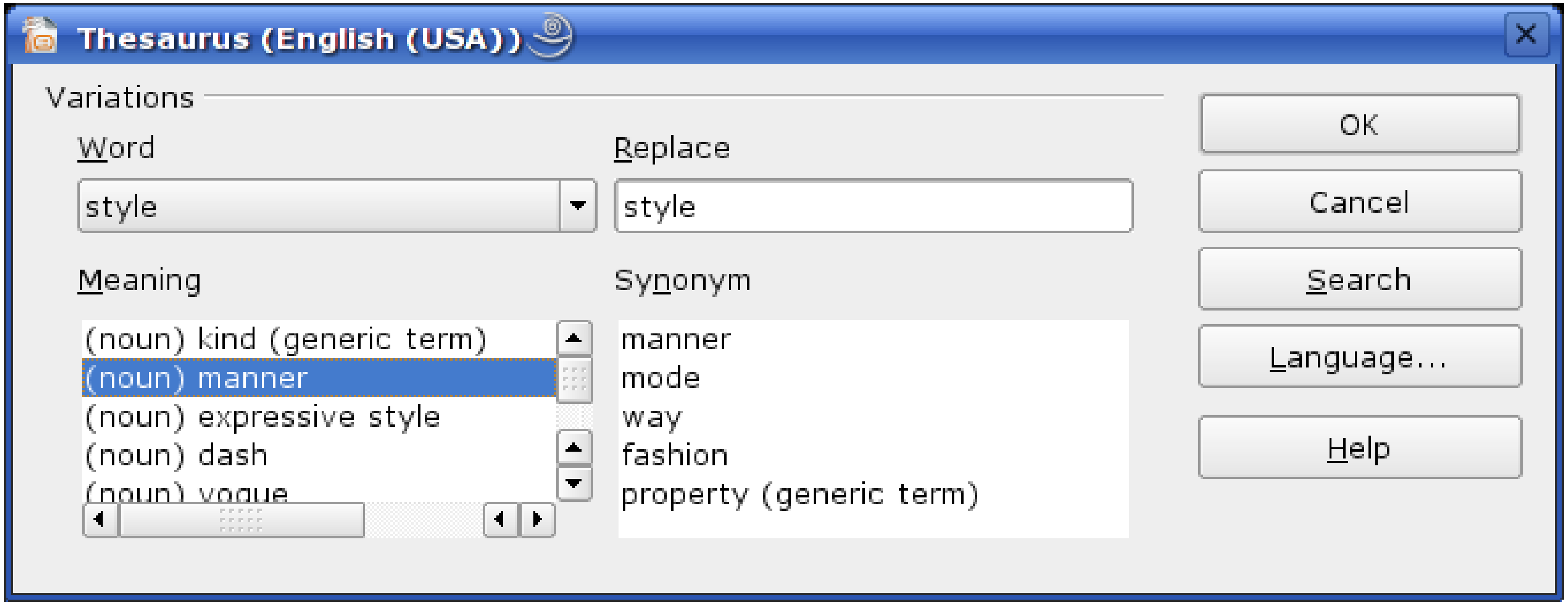
Extend a Dictionary

- Use BootCaT on site with e.g. technical terms
- cat corpus.txt | tr " " "\n" | sort -u | hunspell -l -d xx_YY
  - -l = print misspelled (or unknown) words
- Check all words
  - Remember that some errors are very common, make sure you don't add them

Summary Spell Checking

- Lots of dictionaries available
- It's difficult to judge their quality
- Please subscribe the the mailing list at http://lingucomponent.openoffice.org/ and help dictionary QA

Thesaurus



Thesaurus (English (USA))

Variations

Word: style

Replace: style

Meaning:
(noun) kind (generic term)
(noun) manner
(noun) expressive style
(noun) dash
(noun) vogue

Synonym:
manner
mode
way
fashion
property (generic term)

Thesauri in OpenOffice.org

- Thesaurus = a dictionary that lets you look up words with a similar meaning
- Supported languages: Czech, English, French, German, Greek, Hungarian, Italian, Polish, Russian, Slovak
- Quality differs, probably all could use improvements
- English thesaurus: based on WordNet

Building a Thesaurus

- OpenThesaurus = web application for building and maintaining a thesaurus
- Think of it as a structured Wiki for thesaurus
- OpenThesaurus languages so far
  - German 42,000 words
  - Polish 33,000 words
  - Spanish 21,000 words
  - Slovakian 11,000 words
  - Norwegian 13,000 words
  - Portuguese 13,000 words

OpenThesaurus Requirements (Technical and Personal)

- A little PHP and MySQL knowledge
- A web server with PHP and MySQL
- Time to do research looking for data source (e.g. dictionaries)
- Setup time: 1-4 hours for the software
- 0-20 minutes a day to check user entries + time of the users to actually add words

Summary Thesaurus

- **Most thesauri in OOo built by OpenThesaurus**
- **English thesaurus by WordNet**
- **More work needed**
  - **Adding more languages**
  - **Improvements to existing thesauri**

Why spell checking, thesauri, grammar checking?

Why?

- Quality

A text without errors...

- is easier to read
- looks more professional
- is easier to translate with machine translation
- is easier to find (how to find typos?)

Remember: not everybody is a native speaker of the language she has to use.

Other Ways to Improve Text Quality

- **Collaboration**
  - Let other people read and improve your text
  - Edit -> Changes -> Record
  - Use a collaboration system that supports OOo

- **Text readability measures**
  - Optimize for readability (but don't take the numbers too seriously)
  - http://en.wikipedia.org/wiki/Readability_test

Grammar Checking

Spot the Error

- It's getting rather late know.
- It's getting rather late now.

- I thing that's a good idea.
- I think that's a good idea.

- The dogs barks.
  - The dog barks.
  - The dogs bark.

- The king of France is bald.
- ?!?!

What is Grammar?

- A set of rules that describes how sentences are built
  - Word order
    - OpenOffice.org great is.
  - Agreement
    - OpenOffice.org are great.
  - ...

- Grammar checking starts where spell checking ends
  - Grammar checker uses context
  - Works on complete sentences, not on single words

Overview of Free Grammar Checkers

- LanguageTool: English, Polish, German, Dutch, ...
- An Gramadóir: Irish, Welsh, Scottish Gaelic, ...
- CoGrOO: Portuguese

LanguageTool

- Not (just) a grammar checker but a rule checker
- Rule: describes an error
- Rules in Java or XML

- No online checking yet, needs changes in OOo (OOo 2.4?)

spel check

LanguageTool Examples

- It's getting rather late know.
  - (no output)
- I thing that's a good idea.
  - Did you mean 'think' or 'thinks'?
- The dogs barks loudly.
  - Possible agreement error. You should probably use 'bark'.
- The king of France is bald.
  - (no output)

More Built-in Rules

- This repeats a a word.
  - Possible typo: you repeated a word
- It's based on StarOffice , an office suite...
  - Put a space after the comma, but not before the comma

How to use LanguageTool (User's POV)

- Download from www.languagetool.org
- Install the ZIP as an extension
  Tools -> Extension Manager...
- Restart OOo or open a new window
- Call the new menu item
  LanguageTool -> Check Text

How to Write a Rule

- "I thing that's a good idea."
- thing -> think
  - Rule consists of:
    - Internal id (must be unique): I_THINK
    - Name (displayed in configuration dialog):
      Possible typo 'I think...'
    - Error pattern:
      I followed by thing
    - Message (shown if rule matches)
      Did you mean *think*?
    - Correct example (used for automatic testing)
      I think he's right.
    - Incorrect example (used for automatic testing)
      I thing he's right.

Rule Example in XML

- <rule id="YOU_THING" name="Possible typo 'I/you/... thing(think)'">

```
  <pattern mark_from="1">
    <token regexp="yes">I|you</token>
    <token regexp="yes">thing|things</token>
  </pattern>
```

```
  <message>Did you mean
  <suggestion>think</suggestion>?</message>
 <example type="correct">I
 <marker>think</marker> that's a good
 idea.</example>
</rule>
```

Agreement Rule Example (Simplified)

-
  ```
  <pattern>
    <token postag="DT"/>
    <token postag="NNS" />
    <token postag="VBZ" />
  </pattern>
  ```

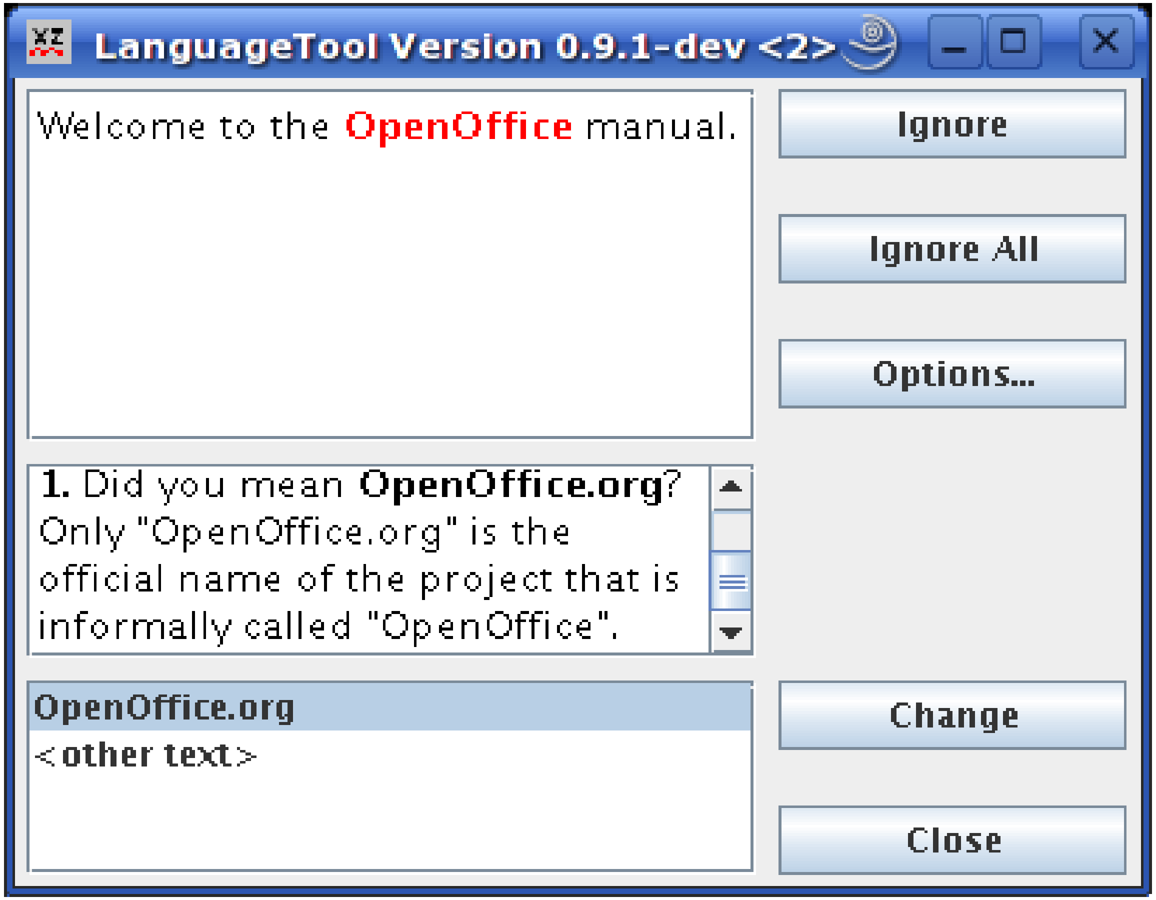
- DT, NNS, VBZ, ... = part-of-speech tags
  - Language dependent

- The dogs barks loudly.
  DT   NNS VBZ
  - DT = determiner (the, a)
  - NNS = plural noun
  - VBZ = 3rd person verb

More Examples

- E.g. for technical documentation
- Not part of LanguageTool, but easy to add:

  -
    ```
    <pattern>
      <token>OpenOffice<token>
      <token negate="yes">.</token>
      <token negate="yes">org</token>
    </pattern>
    ```



Welcome to the OpenOffice manual.

1. Did you mean **OpenOffice.org**? Only "OpenOffice.org" is the official name of the project that is informally called "OpenOffice".

OpenOffice.org
<other text>

Hints for Writing Rules

- Easier rule editing using the XMLmind XML Editor
- Something doesn't work? Use LanguageTool on the command line with the -v option

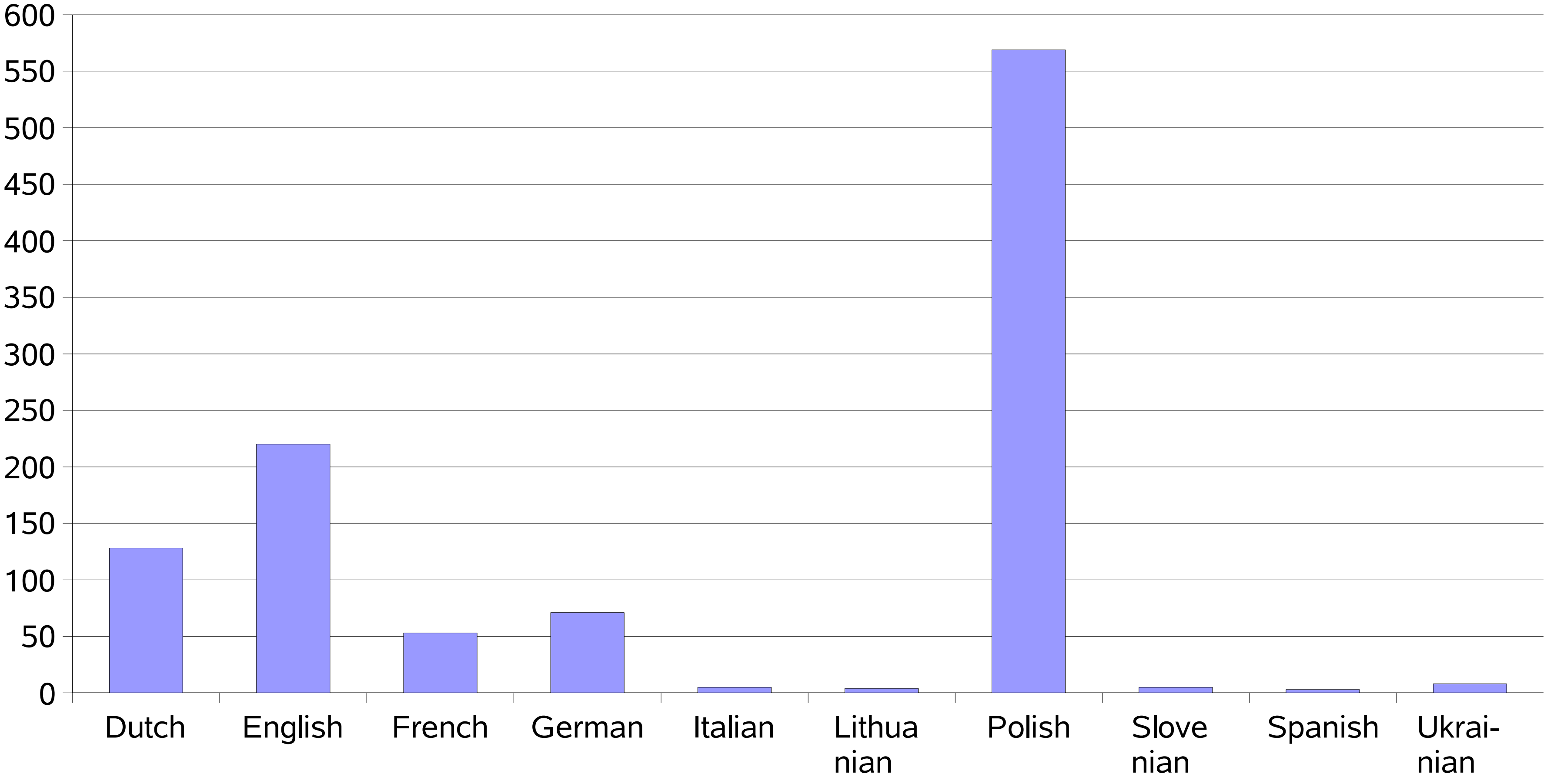
- Remember: don't produce false alarms
  - Check your rules against large amounts of correct text

How to Write a Rule (Java Programmer's POV)

- public class MyRule extends Rule { ... }
- Implement these methods:
  - String getId()
  - String getDescription()
  - RuleMatch[] match(final AnalyzedSentence text):
    - Called once per sentence
    - Lets you iterate over all words an their annotations
    - If you find errors, create RuleMatch objects and return them

LanguageTool Supported Languages

- **By number of rules**



Summary LanguageTool

- Rule-based
- Not limited to grammar
  - Useful for technical documentation
- Can easily be extended to more languages
- No programming knowledge required

Conclusion

- We have a lot of spell checker dictionaries, but could use even more
  - Quality can probably be improved
  - We need to know more about current quality
- We need more thesauri
- We need more grammar checking rules

- You can help – it's not difficult, no programming needed
- Your work will be useful beyond OpenOffice.org
- Consider joining the lingucomponent project

Thank you!

Daniel Naber **www.danielnaber.de**

Mindquarry **www.mindquarry.com**

Lingucomponent Project
http://lingucomponent.openoffice.org/

Thanks to by Björn Jacke, Marcin Miłkowski, and László Németh

License

- These slides: http://creativecommons.org/licenses/by/3.0/
- Stop sign image: http://flickr.com/photos/cocoen/451038962/ (http://creativecommons.org/licenses/by-nc-sa/2.0/deed.en)