



The Lingucomponent Project - Linguistic Tools in OpenOffice.org

OOoCon 2005, Koper - Capodistria, Slovenia

Daniel Naber

<http://www.danielnaber.de>



Agenda

- About the speaker
- OOo vs. StarOffice
- State of the Lingucomponent project
- **Spell Checker**
- **Thesaurus**
- **Grammar Checker**
- Hyphenation
- Fulltext Search
- OOo 2.0 and beyond
- Q & A

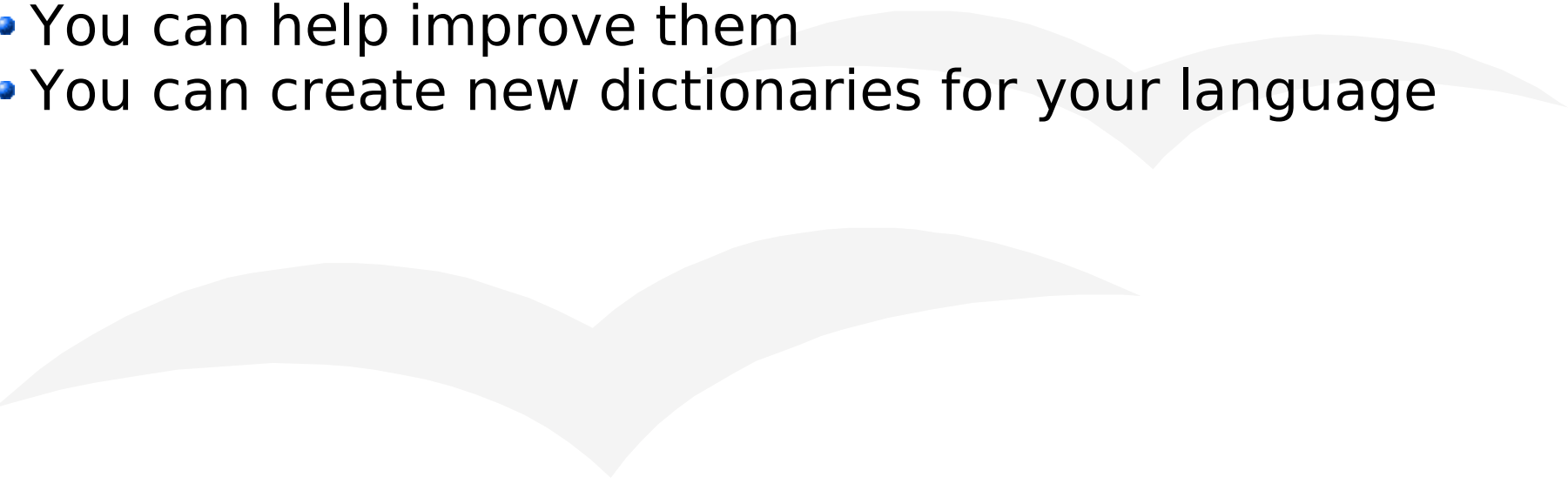


About the Speaker

- OOo:
 - "content developer" for lingucomponent project
 - Maintainer of OpenThesaurus
 - Maintainer of LanguageTool
- Other:
 - Committer at Apache Lucene (Open Source search engine library)
 - Working for IntraFind, Munich, Germany as a software developer



Goal of this Talk

- Understand spell checker, thesaurus & grammar checker so:
 - You can help improve them
 - You can create new dictionaries for your language
- 



OOo vs. StarOffice

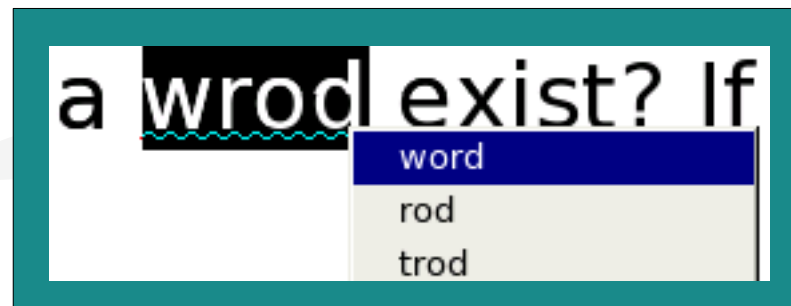
- OOo and StarOffice **do** differ wrt. linguistic features
 - Different dictionaries
 - Different file formats of dictionaries
 - But UNO interface should be the same
- No Sun engineer works on lingucomponent
- OOo only includes a few dictionaries, others need to be downloaded via DicOOo
- StarOffice comes with commercial add-ons for spell checking and thesauri
- Quality comparison: ?

State of the Lingucomponent Project

- Tasks:
 - Spell checking
 - Thesaurus
 - Hyphenation
 - (Grammar checking)
- Lead: Kevin Hendricks / Richard Holt
- New lead: László "Laci" Németh

Spell Checking

- Task: does a word exist? If not, suggest similar words
- A short history of nearly every free spell checker
 - Ispell, Aspell
 - MySpell – library in OOo and Mozilla Thunderbird
 - Hunspell – enhanced MySpell, will replace it
- Advantages of Hunspell over MySpell
 - Unicode, support for compound words, improved suggestions
- State:
 - Dictionaries for 45 languages, counting variants (en_GB/en_US) only once
 - Not all included, download via DicOOo





Spell Checker File Format

- Idea: maintain “small” list of words + flags to cover large number of words
 - en: 46.000 words cover 123.000 word forms – unmunch
- Location: <OOoDir>/share/dict/ooo or ~/.openoffice.org2/user/wordbook/
- Hunspell accepts the MySpell .dic and .dat files
- .aff commands:
 - SFX – suffix, e.g. walked, jumped
 - PFX – prefix, e.g. uninteresting
 - REP – replacement table for suggestions

Spell Checker File Format 2

- .dic Example:

- 2 //number of entries
academy/MS
append/S

- .aff Example:

- SFX M Y 1 // number of following entries for "M"
SFX M 0 's . // 0=cut off, 's=append, .=condition

SFX S Y 3

SFX S y ies [^aeiou]y // academy -> academies

SFX S 0 s [aeiou]y

SFX S 0 s [^hsuxyz] // append -> appends

- Will accept academy, academy's, academies, append, appends



Spell Checker File Format 3

- REP example:
 - REP 2
REP ph f
REP f ph
- Will help to suggest face if someone writes phace, or phase for fase



Adding a New Spell Dictionary

- Start with ispell dictionary + modify .aff file
 - See <http://lingucomponent.openoffice.org/dictionary.html>
- Or start from scratch by crawling the web
 - See <http://borel.slu.edu/crubadan/index.html>
- Dictionary is only useful if quite complete and very accurate

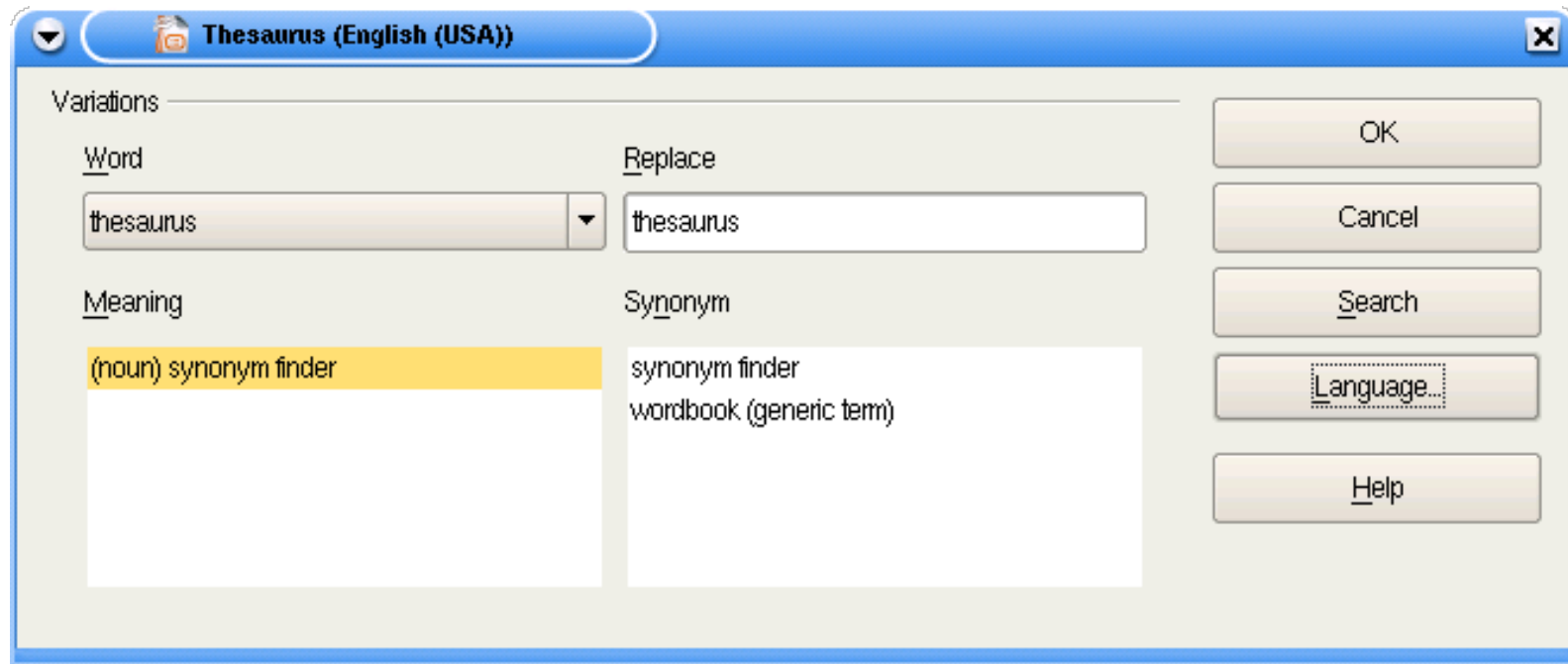


Spell Checking Conclusion

- Many dictionaries exist
- Adding a new one isn't difficult but will be lots of work if no word list exists yet
- TODO:
 - Integrate Hunspell
 - Make use of Hunspell's new compound features for German and Hungarian

Thesaurus

- Task: suggest words with a similar meaning
- Screenshot OOo 2.0:



Hint if it doesn't work: use en_US, not en_GB
(known bug)



Thesaurus State

- 8 languages:
 - English, Czech, French, German, Hungarian, Italian, Polish, Slovakian
(list taken from http://lingucomponent.openoffice.org/thes_dic.html)
- English thesaurus included, others need to be downloaded

Thesaurus File Format

- Changed from OOo 1.x
- .idx example:
 - ISO8859-1 <-- encoding
145866 <-- number of lines
acetyl|139173 <-- word (lowercase) | position in .dat
file in bytes
[...]
five dollar bill|6226683
- .dat Example:
 - ISO8859-1 <-- encoding
acetyl|1 <-- word (lowercase) | number of meanings
(noun)|acetyl group|acetyl radical|ethanoyl group[...]
<-- synonyms
five dollar bill|1
(noun)|fiver|five-spot|bill (generic term)

OpenThesaurus

- Wiki-like PHP web application to create/maintain a thesaurus
- Organized in synsets = meanings. Example:
 - plant, industrial plant <- 1. meaning of plant
 - plant, flora <- 2. meaning of plant
- Features:
 - Search; show random synset
 - Add, delete word in synset
 - Add new synset, delete synset
- Users must be be logged in to make changes
- Relations between synsets: generic term, ...
- Word properties: colloquial, vulgar, ...
- DEMO



OpenThesaurus 2

- Languages: Polish, Spanish, Slovakian, German, in preparation: French, Italian
- QA: admin interface that lists all changes

2005-09-16 22:03:46 dennis: **Spur**, Anflug, Hauch (von), Quentchen
2005-09-16 22:03:28 dennis: **Quentchen**, Anflug, Hauch (von)
2005-09-16 22:02:44 dennis: **Quentchen**, Spur
2005-09-16 22:02:29 dennis: **Spur** [new meaning]

2005-09-16 11:26:04 kopp@: **schlurfen**, gehen, herumschlendern, herumwandern, latschen, ...
2005-09-16 11:25:43 kopp@: **schlurfen**, gehen, herumschlendern, herumwandern, latschen, ...
2005-09-16 11:23:43 kopp@: Bye: **[-]** >> **[(keine Besonderheit)]**
2005-09-16 11:22:36 kopp@: **Bye [umgangssprachlich]**, Ade, Adieu, Auf bald, Auf Wiedersehen, ...
2005-09-16 11:22:18 kopp@: **Wir sehen uns!**, Ade, Adieu, Auf bald, Auf Wiedersehen, ...
2005-09-16 11:21:37 kopp@: Tschüss, **Wir sehen uns!**

- Not just useful for OOo, other uses:
 - Other Open Source office suites
 - Export for OOo (1.x and 2.x), plain text, KWord
 - Search engines

Thesaurus Uses

The screenshot shows a web browser window titled "iF IntraFind Demo Center - Konqueror". The address bar contains the URL: `iF?sp?query=Trittin+atomkraft+USA&sort=relevance&x=37&y=12&index=Grundform&thesaurus=on&appID=1`. The page header displays "INTRA FIND SOFTWARE AG" and a navigation menu with links like "intrafind.de", "Home", "iFinder", "TopicFinder", "PDF-Hit-Highlighting", "Text-Summarizer", "Wortanalyse", "Datenblätter", and "Hilfe".

The search input field contains "Trittin atomkraft USA" and a red "Finde!" button. Below the search bar, the "Index" is set to "Grundform", and the "Thesaurus-Suche" checkbox is checked. The search results are displayed on page 1, sorted by relevance, showing 6 results in 0.47 seconds.

The first three search results are:

- USA lehnen verbindliche Energie-Ziele auch nach Trittin-Gespräch ab**
... Grüne) ab. «Es ist, wenn überhaupt, nur minimale Bewegung in diesem Punkt zu sehen», sagte Trittin in einem dpa-Gespräch nach einem Treffen mit US-Chefunterhändlerin Paula Dobriansky. Trittin ist seit 16/dpa-20020830-1626-475.xml - Ähnliche Seiten
- Nach Energie-Beschluss scharfe deutsche Kritik an USA und OPEC (Mit Bildern EAR01/03/33 und dpa-Grafik Nr. 6808**
... Heidemarie Wieczorek-Zeul (SPD) kritisierte am Dienstag in Johannesburg die «verheerende Kurzsichtigkeit der OPEC und der USA, die in einem Dinosaurier-Denken verhaftet sind, das nicht zukunftsfähig ist». Umweltminister Jürgen Trittin (Grüne ... 92/dpa-20020903-1406-278.xml - Ähnliche Seiten
- Bundesrat besiegelt Deutschlands Ausstieg aus der Atomenergie**
... Europäischen Union nicht im Alleingang aus der Kernkraft aus, dafür aber am schnellsten. Mit Blick auf die Terroranschläge in den USA am 11. September verwies der Minister auf die Risiken der Atomkraft. Der bayerische Umweltminister Werner Schnappauf ...

On the right side, a sidebar titled "Suche um verwandte Begriffe aus einem manuell gepflegten Thesaurus erweitern:" provides related terms:

- Atomkraft**
 - Atomenergie
 - Kernenergie
 - Kernkraft
- USA**
 - Amerika
 - Vereinigte Staaten
 - Vereinigte Staaten von Amerika

A "Suche erweitern" button is located at the bottom of the sidebar.



German OpenThesaurus

- <http://www.openthesaurus.de>
- about 34.000 words, 14.000 meanings
- 2000 queries/day on the website
- matches for ~75% of queries (not every word has a synonym!)
- full form -> base form mapping: ging -> gehen (not in OOo)
- license: GPL
- imported from EN/DE dictionary



Building a New Thesaurus

- Search for existing synonym lists
- Search for existing synonym lists again
- If that fails, consider using a two-language dictionary
- Use LGPL if possible
- Set up OpenThesaurus on your server (requires PHP + MySQL)
- Don't add proper nouns or abbreviations, the tasks gets too big that way!
- Only add words that actually **have** at least one synonym (or a generic term)



Thesaurus Conclusion

- 8 languages covered -> more thesauri needed
- Easy to build using OpenThesaurus
- Still quite some work, long-term commitment required
- TODO:
 - integrate László's Unicode patch (issue #54268)
 - integrate hunstem: find synonyms for walks, not just for walk
 - next step: inflect synonyms:
"He walks home" -> "He runs home"
walks --baseform--> walk --synonym--> run
--inflect--> runs



Grammar Checking

- Tasks:
 - Check whether the text is grammatically correct, suggest fixes if needed
 - Does the text contain "forbidden" words
 - Avoid different spellings in one text (aufwendig <-> aufwändig)
 - Does the text contain typical errors:
 - it's <-> its
 - there <-> their <-> they're
 - **Any** check you can think of and that a computer can handle
- Grammar checker = spell checking + context

Grammar Checking Examples 1

- "Sorry for my bed English"
 - 1000+ matches with Google for the above phrase!



- "Sorry for my bat English"
 - 180+ matches with Google for the above phrase!
- Are these grammar problems?
A spell checker cannot detect them.

Grammar Checking Examples 2

- False Friends
 - German example: “bekommen” sounds like “become” but means something totally different



"I become a cheeseburger"



Grammar Checker State

- No grammar checker in OOo
- No interface available
- Some integration via macro
- Some Open Source grammar checkers:
 - An Gramadoir (by Kevin Scannell)
 - Good support for Irish
 - Other languages in development: Afrikaans, Cornish, Esperanto, French, ...
 - Published under GPL
 - Written in Perl
 - LanguageTool:
 - Support for English and German
 - Published under LGPL
 - Written in Java



LanguageTool Approach

- Split text into sentences
- Part-of-speech tagging per sentence:
Thanks/NNS for/IN the/DT responds/NNS./.
- Different for German! several tags & properties per word
- Apply XML error rules:
<pattern>DT "responds" </pattern>
- Apply Java error rules
 - e.g. word repetition
 - e.g. a hour -> an hour
- Print all rules that match
- Show Demo



LanguageTool Example

- Input text: Thanks for the responds.
- Rule:
 - ```
<rule id="DT_RESPONDS" name="Confusion of responds/response">
 <pattern lang="en">DT "responds" </pattern>
 <message>Did you mean
 response? </message>
 <example type="correct">Thanks for the
 response.</example>
 <example type="incorrect">Thanks for the
 responds.</example>
</rule>
```
- Output: 1.) Line 1, column 15  
Message: Did you mean 'response'?

# Adding a Language to LanguageTool

- LanguageTool is extensible, i.e. you just need to implement given Interfaces (e.g. *Tagger*)
- Find or implement a part-of-speech tagger for your language
- Collect grammar rules and/or real-world errors
- Write error rules in XML
- Write error rules in Java
- See <http://www.danielnaber.de/languageTool/>

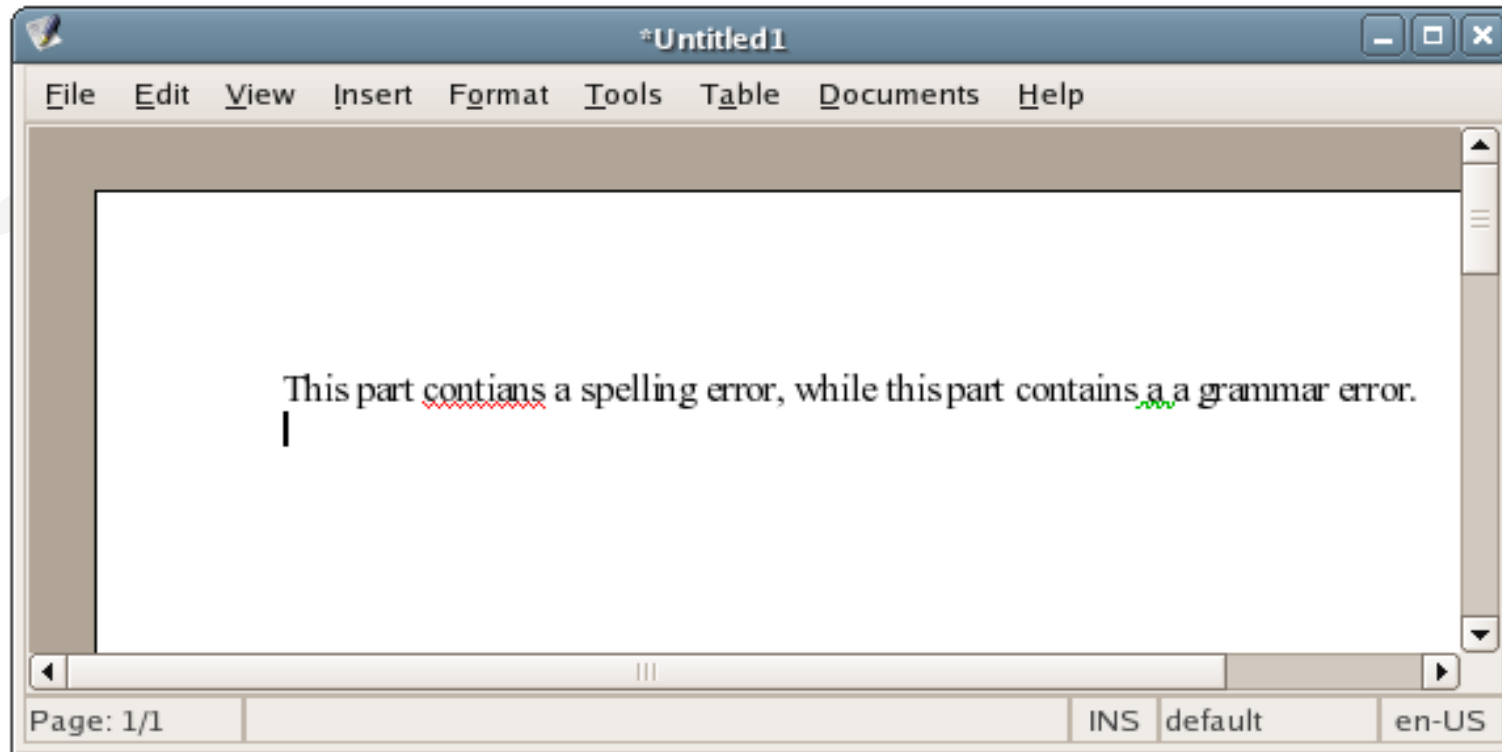


# Link Grammar

- See <http://www.link.cs.cmu.edu/link/>
- Alternative approach: don't search for errors, but parse each sentence like a compiler for a programming language
- Cannot parse: must be an error
- Advantage:
  - Will find any kind of error
- Problems:
  - “Dear user, there's an error somewhere” is not enough: what kind of error? How to fix it?
  - English only, no easy way to adapt it to a different language
  - Grammar must be complete, otherwise it will complain about many correct sentences

# Link Grammar 2

- Implemented in AbiWord 2.3.5 (a development version)
- No feedback other than “there's an error” yet





# Grammar Checker Conclusion

- No mature Open Source grammar checker for a mainstream language available yet
- Very, very difficult to get a “complete” checker
- Help needed on OOo side
  - Adding an interface, similar to that of the spell checker
  - Should support on-the-fly checking
  - Keli Hu (hukeli at gmail com) is working on this
- Help needed on the grammar checkers
  - Adding more error rules
  - Adding more languages
  - No knowledge of OOo required
  - For many cases no programming knowledge required at all



# Hyphenation

- Task: insert hyphen at the end of a line, like  
follow the wa-  
ter
- Activate:
  - Hyphenation dictionary must be installed (English and others included by default)
  - Paragraph Style -> Text Flow -> Hyphenation: check "automatically"
- See <http://lingucomponent.openoffice.org/hyphenator.html>
- TODO: some hyphenation points missed, especially for languages with compound words like German

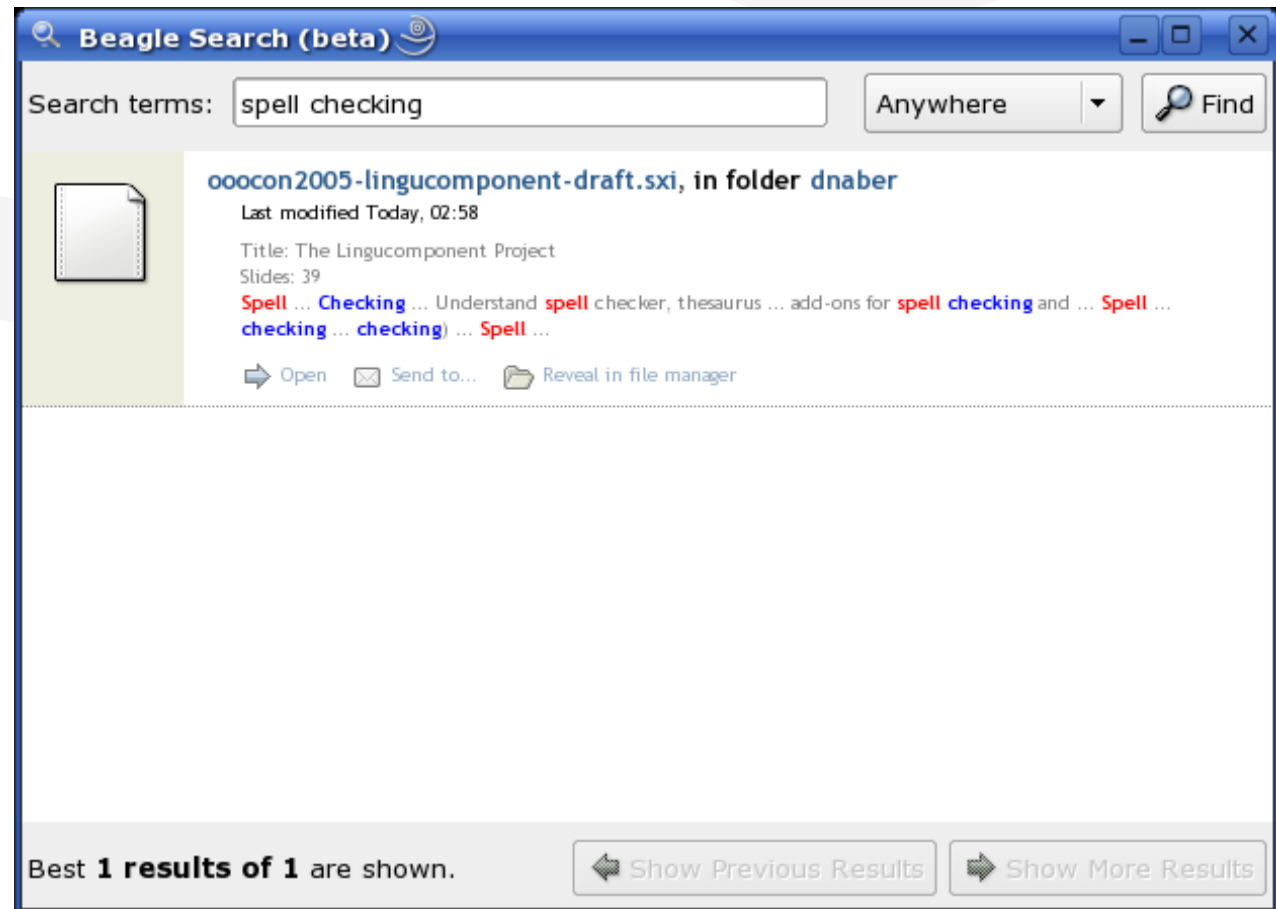


# UNO

- Adding your own implementation of spell checking, thesaurus and hyphenation
- C++, Java, Python, ...
- Example:
  - ```
public class MyThesaurus extends ComponentBase
implements XThesaurus, [...] {
    public XMeaning[] queryMeanings(String term,
    Locale locale, PropertyValue[] properties) {
        XMeaning[] meanings = new XMeaning[] {...};
        return meanings;
    }
}
```
- Documentation at <http://api.openoffice.org/DevelopersGuide/DevelopersGuide.html>

Fulltext Search

- OOo files are XML files compressed with ZIP
- OOo and StarOffice use the same file format
- Beagle -- <http://beaglewiki.org/>





Fulltext Search 2

- Look -- <http://www.danielnaber.de/look/>
- FoooX -- <http://oootools.free.fr/fooox/>
- OpenOffice Indexer --
<http://www.jerger.org/en/indexer.html>
- o3find -- <http://web.tiscali.it/fanelia/sw/o3find/>
(Windows-only)
- Google Desktop Search plugin --
<http://desktop.google.com/plugins/indextheopenoffice.html>
- OOo 2.0: plugin for Windows search



OOo 2.0 and Beyond

- OOo 2.0: much improved thesaurus
 - English
 - less synonyms, but better quality (WordNet-based)
 - en_GB and en_US (both in one thesaurus)
 - German + English + ...: words categorized by meaning
- Planned for OOo 2.0.1: Hunspell replaces MySpell
 - English: no difference (but Hunspell is actively maintained)
 - German + Hungarian: compound word support
 - several non-Western languages: Unicode support
- Later: grammar checking interface



Conclusion

- Good spell checkers available for many languages
- Thesauri available for several languages, but not enough
- Grammar checkers still in beta
- Grammar checker integration still missing
- Testing and more feedback needed in all areas
- StarOffice is different, and that's not a help
- See <http://lingucomponent.openoffice.org> and join the mailing list
dev@lingucomponent.openoffice.org



Thanks for your attention!



Questions?

