Nuspell: version 3 of the new spell checker

FOSS spell checker implemented in C++17 with aid of Mozilla

Sander van Geloven

FOSDEM, Brussels

February 1, 2020

Nuspell

Workings

Technologies

Upcomming

Nuspell

Nuspell is

- spell checker
- free and open source software with LGPL
- library and command-line tool
- written in C++17

Nuspell – Team

Our team currently consists of

- Dimitrij Mijoski
  - lead software developer
  - github.com/dimztimz
- Sander van Geloven
  - information analyst
  - hellebaard.nl
  - linkedin.com/in/svgeloven
  - github.com/PanderMusubi

Nuspell – Spell Checking

Spell checking is not trivial

- much more than searching a long flat word list
- dependent of language, character encoding and locale
- involves case conversion, affixing, compounding, etc.
- suggestions for spelling, typing and phonetic errors
- long history of decades with `spell`, `ispell`, `aspell`, `myspell`, `hunspell` and now `nuspell`

See also my talks at FOSDEM 2019 and FOSDEM 2016 at archive.fosdem.org/2019/schedule/event/nuspell and archive.fosdem.org/2016/schedule/event/integrating_spell_and_grammar_checking

Nuspell – Goals

Nuspell's goals are

- a drop-in replacement for browsers, office suites, etc.
- backwards compatibility MySpell and Hunspell format
- improved maintainability
- minimal dependencies
- maximum portability
- improved performance
- suitable for further development and optimizations

Implemented in object-oriented, templated and modern C++.

Nuspell – Features

Nuspell supports

- many character encodings
- compounding
- affixing
- complex morphology
- suggestions
- personal dictionaries
- 167 (regional) languages via 89 existing dictionaries

Nuspell – Support

Mozilla Open Source Support (MOSS) funded in 2018 the creation of Nuspell. Thanks to Gerv Markham† and Mehan Jayasuriya. See mozilla.org/moss for more information. Mozilla Open Source Support (MOSS) funded in 209/2020 the development of Nuspell version 3. Thanks to Mehan Jayasuriya and Bas Schouten. See mozilla.org/moss for more information.

Verification Hunspell has a mean precision of 1.000 and accuracy of 0.997. Perfect match 70% of tested languages. On average checking 30% faster and suggestions 8× faster.

Workings – Spell Checking

Spell checking is highly complex and unfortunately not suitable for a lightning talk. It mainly concerns

- searching strings
- using simple regular expressions
- locale-dependent case detection and conversion
- finding and using break patterns
- performing input and output conversions
- matching, stripping and adding (multiple) affixes, mostly in reverse
- compounding in several ways, mostly in reverse
- locale-dependent tokenization of plain text

Workings – Case Conversion

Examples of non-trivial case detection and conversion

▶ `to_title("istanbul")` →	English `"Istanbul"` Turkish `"İstanbul"`
`to_upper("Diyarbakır")` →	English `"DIYARBAKIR"` Turkish `"DİYARBAKIR"`
▶ `to_upper("σίγμα")` →	Greek `"ΣΙΓΜΑ"`
`to_upper("ςίγμα")` →	Greek `"ΣΙΓΜΑ"`
`to_lower("ΣΙΓΜΑ")` →	Greek `"ςίγμα"`
▶ `to_upper("Straße"` →	English `Straße"`
`to_upper("Straße"` →	German `STRASSE"`
▶ `to_title("ijsselmeeer")` →	English `"Ijsselmeer"`
`to_title("ijsselmeeer")` →	Dutch `"IJsselmeer"`

Workings – Suggestions

1.	upper case*	TODOh[ëê]llo → TODOhello
2.	replacement table	h[ëê]llo → hello
3.	mapping table	hełło$ → hello
4.	adjacent swap*	TODOhhello → TODOhello
5.	distant swap*	TODOhhello → TODOhello
6.	keyboard layout	hrllo → hello
7.	extra character	hhello → hello
8.	forgotten character	hllo → hello
9.	move character*	TODOhellø → TODOhello
10.	bad character	hellø → hello
11.	doubled two characters*	TODOhellø → TODOhello
12.	two words suggest*	TODOhellø → TODOhello
13.	phonetic mapping	ˆello → hello

WorkingsTODO – Dependencies

- gspell: balsa, corebird, evince, evolution, geary, gedit, gnome-builder, gnome-recipes, gnome-software, gnote, gtranslator, latexila, osmo, polari
- ispell: bgoffice, dsdo, eo-spell, eoconv, eog, eog-plugins, espa-nol, espctag, espeak, espeak-ng, espeakedit, espeakup, esperanza, espresso, esptool, hkgerman, hkl, hl-todo-el, hlbr, hlbrw, ifrench, ifrench-gut, ifrit, ifscheme, ifstat, iftop, ifupdown, ifupdown-extra, ifupdown-multi, ifupdown2, ifuse, igal2, igdiscover, igmpproxy, ignition-cmake, ignition-cmake2, ignition-common, ignition-fuel-tools, ignition-math2, ignition-math4, ignition-msgs, ignition-transport, ignore-me, igor, igraph, igtf-policy-bundle, ii, ii-esu, iio-sensor-proxy, ippl, ippusbxd, iprange, iproute2, iprutils, ips, ipset, ipsvd, iptables, iptables-converter, iptables-netflow, iptables-optimizer, iptables-persistent, iptotal, iptraf-ng, iptstate, iptux, iputils, ipv6calc, ipv6pref, ipv6toolkit, ipvsadm, ipwatchd, ipwatchd-gnotify, ipxe, ipxe-qemu-256k-compat, irda-utils, irqbalance, irssi, isc-dhcp, isl, ispellcat, isrcsubmit, isso, istack-commons, istgt, isync, italc, itamae, iucode-tool, iw, jackd2, m17n-lib, magyarispell, norwegian, petname, pkg-perl-tools, pkg-php-tools, pkgconf, pkgdiff, pkgsel, pkgsync, pktanon, pktools, pktstat, softcatala-spell

WorkingsTODO – Dependencies

- python3-enchant: mwic, ocrfeeder, pygtkspellcheck, python-autobahn, python-avro, python-aws-requests-auth, python-aws-xray-sdk, python-axolotl, python-axolotl-curve25519, python-azure, python-azure-devtools, python-azure-storage, python-b2sdk, python-babelgladeextractor, python-backcall, python-backports-abc, python-backports-shutil-get-terminal-size, python-backports.tempfile, python-backports.weakref, python-base58, python-txaio, python-txosc, python-typeguard, python-typing, python-typing-extensions, python-tz, python-tzlocal, sphinxcontrib-spelling, translate-toolkit, translation-finder, translatoid, transmageddon, transmission, transmission-el, transmission-remote-cli, transmission-remote-gtk, transmissionrpc, transrate-tools, transtermhp, trapperkeeper-clojure, trapperkeeper-metrics-clojure, trapperkeeper-scheduler-clojure, trapperkeeper-status-clojure, trapperkeeper-webserver-jetty9-clojure, trash-cli, traverso, travis, trayer, tre, tree, tree-puzzle, tree-style-tab, virtaal

WorkingsTODO – Dependencies

- hunspell: aegisub, calibre, chromium, chromium-browser, codeblocks, codelite, dict-af, dict-nr, dict-ns, dict-ss, dict-st, dict-tn, dict-ts, dict-ve, dict-xh, dict-zu, dictd, dutch, dv4l, dvbackup, dvbcut, dvblast, dvbsnoop, dvbstream, dvbstreamer, dvbtune, dvcs-autosync, dvd+rw-tools, dvd-slideshow, dvdauthor, dvdbackup, dwz, e2fsprogs, ebtables, ec2-hibinit-agent, ed, edk2, efibootmgr, efivar, eject, elfutils, emacs, emacsen-common, enchant, enchant-2, entrypoints, eo-spell, eoconv, eog, eog-plugins, espa-nol, espctag, espeak, espeak-ng, espeakedit, espeakup, esperanza, espresso, esptool, featherpad, focuswriter, ghostwriter, goldendict, gwaei, hspell, hunspell-bo, hunspell-dz, ifrench, ifrench-gut, ifrit, ifscheme, ifstat, iftop, ifupdown, ifupdown-extra, ifupdown-multi, ifupdown2, ifuse, igaelic, igal2, igdiscover, igerman98, igmpproxy, ignition-cmake, ignition-cmake2, ignition-common, ignition-fuel-tools, ignition-math2, ignition-math4, ignition-msgs, ignition-transport, ignore-me, igor, igraph, igtf-policy-bundle, ii, ii-esu, iio-sensor-proxy, iirish, ispell-czech, ispell-fi, ispell-fo, ispell-gl, ispell-lt, ispell-tl, ispell-uk, ispell.pt, ispellcat, isrcsubmit, isso, istack-commons, istgt, isync, italc, itamae, iucode-tool, iw, jackd2, libgtk2-spell-perl, libreoffice, libtext-hunspell-perl, link-grammar, lokalize, mudlet, mythes, nixnote2, norwegian, onboard, pkg-perl-tools, pkg-php-tools, pkgconf, pkgdiff, pkgsel, pkgsync, pktanon, pktools, pktstat, plume-creator, psi-plus, pyhunspell, qtvirtualkeyboard-opensource-src, scribus, scribus-ng, sigil, sonnet, sphinxcontrib-spelling, tea, texstudio, texworks, thunderbird

WorkingsTODO – Dependencies

- libenchant: abiword, ayttm, bibledit-gtk, bluefish, claws-mail, empathy, evolution, fcitx, fcitx5, geany-plugins, geary, gnome-builder, gnome-subtitles, gspell, gtkhtml4.0, gtkspell, gtkspell3, java-gnome, kadu, kde4libs, kvirc, lifeograph, lyx, ogmrip, php7.1, php7.2, php7.3, pluma, psi, purple-plugin-pack, pyenchant, qtspell, stardict, subtitleeditor, sylpheed, webkit, webkit2gtk, webkitgtk, xneur python3-enchant: mwic, ocrfeeder, pygtkspellcheck, python-autobahn, python-avro, python-aws-requests-auth, python-aws-xray-sdk, python-axolotl, python-axolotl-curve25519, python-azure, python-azure-devtools, python-azure-storage, python-b2sdk, python-babelgladeextractor, python-backcall, python-backports-abc, python-backports-shutil-get-terminal-size, python-backports.tempfile, python-backports.weakref, python-base58, python-txaio, python-txosc, python-typeguard, python-typing, python-typing-extensions, python-tz, python-tzlocal, sphinxcontrib-spelling, translate-toolkit, translation-finder, translatoid, transmageddon, transmission, transmission-el, transmission-remote-cli, transmission-remote-gtk, transmissionrpc, transrate-tools, transtermhp, trapperkeeper-clojure, trapperkeeper-metrics-clojure, trapperkeeper-scheduler-clojure, trapperkeeper-status-clojure, trapperkeeper-webserver-jetty9-clojure, trash-cli, traverso, travis, trayer, tre, tree, tree-puzzle, tree-style-tab, virtaal

Workings – Initialization

Initialize Nuspell in four steps in C++

- find, get and load dictionary

```
auto find = Finder::search_all_dirs_for_dicts();
auto path = find.get_dictionary_path("en_US");
auto dic = Dictionary::load_from_path(path);
```

- associate currently active locale

```
boost::locale::generator gen;
auto loc = gen("");
dic.imbue(loc);
```

These steps are more simple when using the API.

Workings – Usage

Use Nuspell by simply calling to

- check spelling

```
auto spelling = false;
spelling = dic.spell(word);
```

- find suggestions

```
auto suggestions = List_Strings();
dic.suggest(word, suggestions);
```

Technologies – Headers

Headers used in build process

- C++17 library
  e.g. GNU Standard C++ Library
  libstdc++ ≥ 7.0

- Boost.Locale
  C++ facilities for localization
  boost-locale ≥ 1.62

Technologies – Libraries

Libraries used in run-time

- C++17 library
  e.g. GNU Standard C++ Library
  libstdc++ ≥ 7.0

- Boost.Locale
  C++ facilities for localization
  boost-locale ≥ 1.62

- International Components for Unicode (ICU)
  a C++ library for Unicode and locale support
  icu ≥ 57.1

Technologies – Improvements

TODO

- C++14   C++17
- dropped
- default support for UTF8, dropping usage of imbue, locale and codecvt and Bost header file

Technologies – Compilers

Currently supported compilers to build Nuspell

- GNU GCC compiler g++ ≥ 7.0
- LLVM Clang compiler clang ≥ 5.0
- Microsoft Visual C++ compiler MSVC ≥ 2017

Upcoming supported compilers

- MinGW with MSYS mingw
- GNU GCC compiler 6.0 (backport)

Technologies – Tools

Tools used for development

- build tools such as Autoconf, Automake, Make, Libtool and pkg-config
- QtCreator for development and debugging, also possible with gdb and other command-line tools
- unit testing with Catch2
- continuous integration with Travis for GCC and Clang and coming soon AppVeyor for MinGW
- profiling with Callgrind, KCachegrind, Perf and Hotspot
- API documentation generation with Doxygen
- code coverage reporting with LCOV and genhtml

Upcoming – Next Version

Next version will have improved

- performance
- compounding
- suggestions
- API
- command-line tool
- documentation
- testing

Nuspell will then also be

- migrated to CMake
- integrated with web browsers
- offering ports and packages
- offering language bindings

Upcoming – Ports and Packages

Supported

- Ubuntu ≥ 19.10 (Eoan Ermine)
- Debian ≥ 10 (Buster)

Tested

- FreeBSD ≥ 11

Help wanted

- Android
- Arch Linux
- CentOS
- Fedora
- Gentoo
- iOS
- Linux Mint
- macOS
- NetBSD
- OpenBSD
- openSUSE
- Slackware
- Windows
- ...

Upcoming – Language Bindings

Supported

- C++
- C

Help wanted

- C#
- Go
- Java
- JavaScript
- Lua
- Objective-C
- Perl
- PHP
- Ruby
- Rust
- Python
- Scala
- ...

Upcoming – Miscellaneous

Other ways to help are

- fix bugs in dictionaries and word lists
- improve dictionaries and word lists
- contribute word lists with errors and corrections
- integrate Nuspell with IDEs, text editors and editors for HTML, XML, JSON, YAML, TeX, etc.
- integrate Nuspell with Enchant e.g. for GtkSpell
- sponsor our team
- join our team

Upcoming – Info and Contact

nuspell.github.io

twitter.com/nuspell1

facebook.com/nuspell

fosstodon.org/@nuspell

Big thank you to Dimitrij.

Contact us to support the development, porting and maintenance of Nuspell.

Thanks for your attention.