Supporting specific styles and terminology & boosting the Spanish module

David Méndez (hello@davidemdot.com)

(This GSoC proposal has been previously discussed in the LanguageTool forum.)

Synopsis

I will add support for **different style guides and specific terminology**. The user will be able to activate a set of rules designed for writing scientific papers. In the same way, I will develop a tool for allowing the implementation of a particular terminology (e.g., if the user is writing something related to microbiology).

The core of the proposal is a **wide improvement of the Spanish module**, which will **place this language at the same level of the most active**. I am going to work in the same line as the current maintainer, so my work will be an extension of yours and my deliveries will have continuity in the future, trying to **increase the amount of contributions from the community**.

Benefits

Spanish has 567 million of potential speakers: 472 million of natives, 73 million with basic competence and 21 million of students[1]. Last year, the web version of the dictionary of the Royal Spanish Academy handled 750 million of queries[2].

As a comparison, WordReference is a popular multilingual translation website, and Spain is the top 1 country of the visitors (Spain, 15.3%; United States, 12.7%; France, 11.6%; Italy, 10.4%; Mexico, 5.6%)[3]. On the other hand, LanguageTool is less attractive for Spaniards: Poland, 19.2%; Germany, 16.2%; Russia, 8.4%; China, 6.2%; Spain, 5.9%[4].

Even though there is a high interest from Spanish speakers in language services, **LanguageTool has a low rate of Spanish visitors** in comparison with other popular language tool as WordReference.

To give the final boost to Spanish module of LanguageTool, I contacted Juan Martorell (Spanish maintainer) so I could understand what the state-of-the-art was for Spanish language. I think that **it is essential to work in the same line as the current maintainer**, so my work will be an enhancement of yours and there will not be conflicts between both ways of working. It will also give continuity to my deliveries. He explained to me the roadmap for Spanish. Also, he shared with me a repository that contains some tools for analysing the new rules and check whether they improve the user experience.

---

[1] http://www.cervantes.es/imagenes/File/prensa/EspanolLenguaViva16.pdf
[2] http://www.rae.es/noticias/el-dle-en-linea-recibe-750-millones-de-consultas-en-2017
[3] https://www.alexa.com/siteinfo/wordreference.com
[4] https://www.alexa.com/siteinfo/languagetool.org

Deliverables

My contribution will be a Spanish module that offers a high level of grammatical correctness and highly reliable style guidance. When **version 4.3** of the project is released **on 26 September, LanguageTool will have Spanish as one of its main languages**.

Also, the user will be able choose between different **style guides and terminologies**. These features **can be extended to other languages**.

All the developments will be accompanied by user documentation and developers' documentation. It is a priority to give a **proper guide to the Spanish users who want to collaborate with the project**.

Project Details

Final boost to Spanish module

Recently, Juan Martorell's committed a larger and improved **POS dictionary** and its synthesizer counterpart for Spanish. Now it is time to design a good **disambiguator**, following the strategy of starting with the longest expressions and continuing towards the shorter ones (composed of two tokens). This could include complex constructions, such as conditional sentences (to evaluate the verb tenses used) and subordinates as well. And face the big challenge of attributive sentences.

Regarding the **rules**, it is necessary to review all of them, simplifying them as far as possible taking advantage of a good disambiguator, reorganizing them and adding those that are necessary. Also, completing the rules that do not have correct examples. It would be beneficial for the Spanish module to reduce technical debt as much as possible.

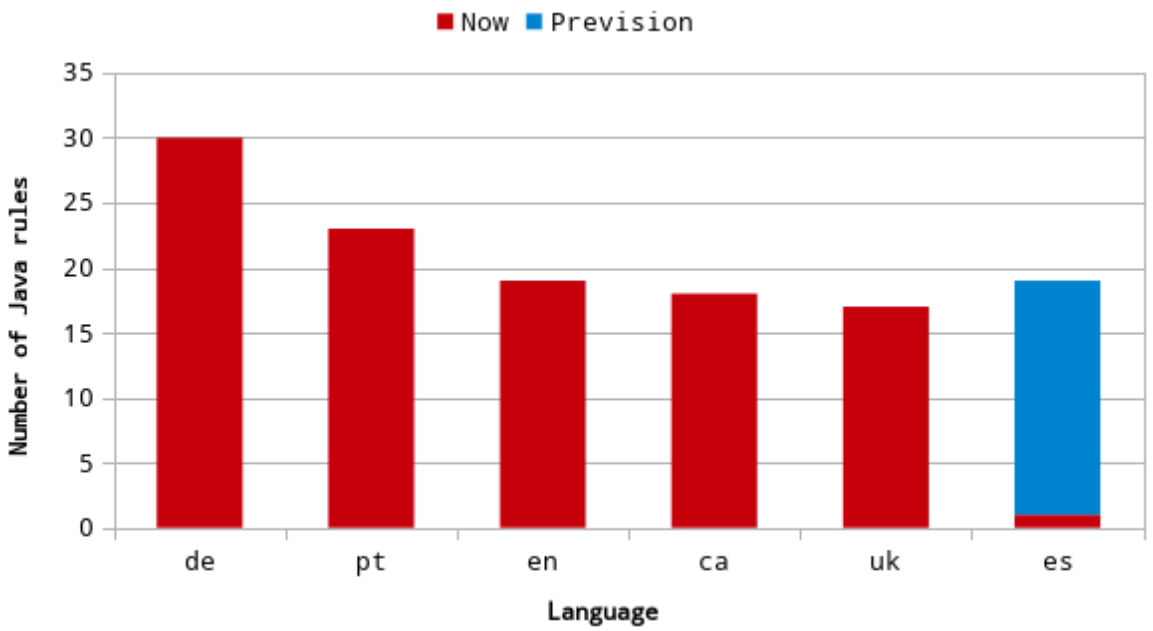
To prioritize the rules, the CAES[5] and CAELE[6] **corpus of errors**, and the papers already published around them, will be used. To reorganize them, could be helpful to have a look to the **typology** defined in some papers[7][8].

Some **Java rules** already developed in other languages that will be used or localized to Spanish:

1. BarbarismsRule
2. ClicheRule
3. CompoundRule
4. ContractionSpellingRule
5. DashRule
6. DateCheckFilter
7. EmptyLineRule
8. LongSentenceRule
9. RedundancyRule
10. ReflexiveVerbsRule
11. SentenceWhitespaceRule
12. WeaselWordsRule
13. WhiteSpaceAtBeginOfParagraph
14. WhiteSpaceBeforeParagraphEnd
15. WordCoherencyRule
16. WordinessRule
17. WordRepeatBeginningRule
18. WrongWordInContextRule

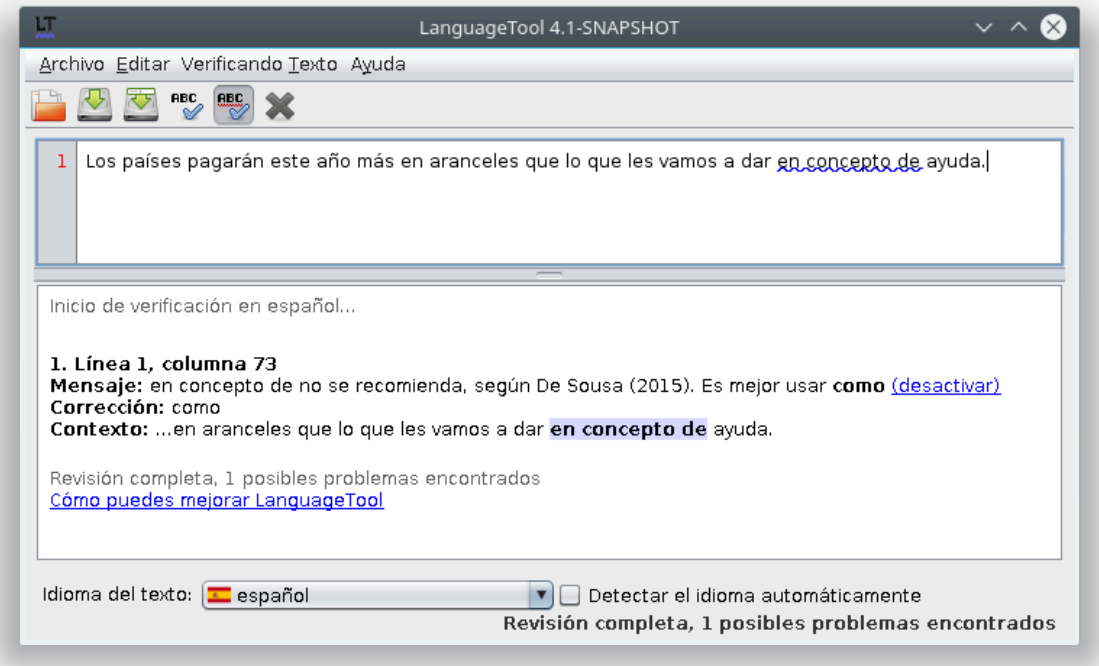
---

[5] http://galvan.usc.es/caes
[6] http://www.scielo.br/pdf/rbla/v17n3/1984-6398-rbla-201710927.pdf
[7] https://pdfs.semanticscholar.org/0ac5/846968718590d218317f2aaada0c41e04a04.pdf
[8] https://scielo.conicyt.cl/pdf/signos/v47n86/a03.pdf



(All of these stages include the use of the tools of the laboratory, **to ensure that regressions and false positives do not increase**.)

Also, could be a good idea implementing a **StyleSuggestionsRule** that advices to the user about some constructions that, even not being erroneous, could be improved. It will show a recommendation and its source.



I will study and compare if some errors have to be corrected by a Java rule or an XML one. I will use as bibliography some books and concerning documents about Spanish correctness:

- ➔ RAE (Royal Academy of Spanish) Dictionary[9]
- ➔ RAE Orthography[10]
- ➔ RAE Pan-Hispanic Dictionary of Doubts[11]
- ➔ Fundéu[12]
- ➔ Wikilengua[13]
- ➔ Style manuals: a relevant book[14] and some online resources[15]
- ➔ Other papers and corpus

For the detection of new possible rules for grammar.xml and disambiguator.xml, I will use **external taggers** (I have been testing the NLTK Brill Tagger[16], but I am going to explore other alternatives until May 14), trying to automate the process to the maximum. This could be extensible to other languages.

Specific styles and terminology

In addition, I will add support for using different style rules, so **the user is able to activate them in specific cases**, such as when writing a scientific paper.

Attending to **agility, workload and usability**, it is not practical to design rules for covering huge mistakes. Here is a text[17] that shows some basic errors that are currently not detected because of reasons explained above:

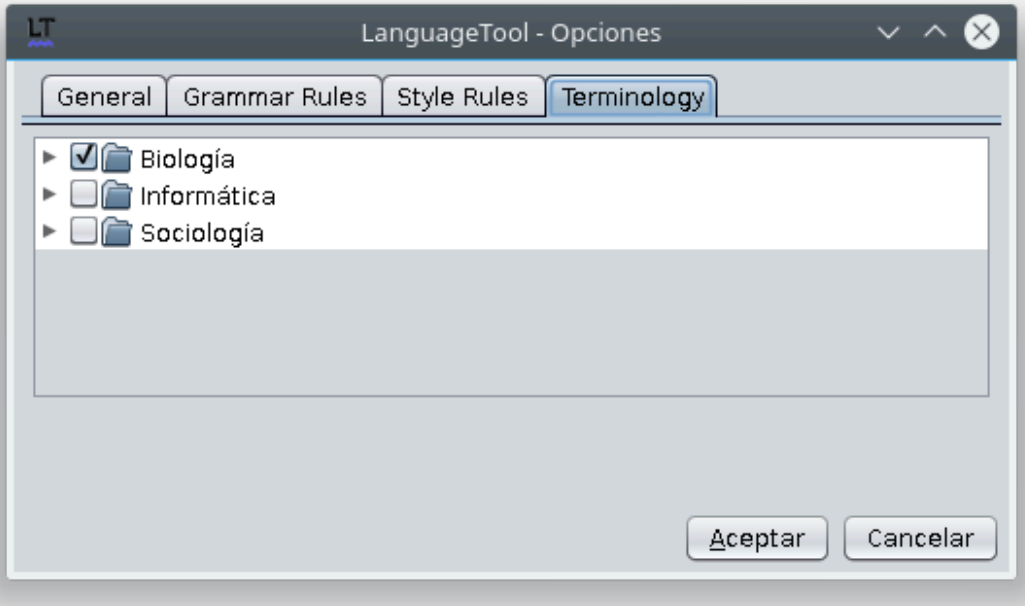
> *Me llamo Elena y estoy profesora de español en una escuela de Salamanca. Soy 29 años y así es un día normal para mí: Yo despierto más temprano, sobre las 6, pero no me levanto hasta las 6:30. Me ducha y me preparo la desayuno. Me visto y voy en la escuela en pie; la gente, en Salamanca, caminan mucho. Termino las clases a las 14 y vuelvo a la mía casa. Por la tarde, estudio y preparo las clases.*

This feature could allow loading specific rules for beginners of Spanish language. There will be a default set of rules, but it could be changed choosing the desired one from a drop-down list (in GUI versions: stand-alone, web, office suites add-ons) or including an extra parameter (in the API and the command-line version).

Also, I will develop a package for allowing implementing specific terminology (for example, to correct something related to sociology), so LanguageTool will not report errors continuously. Its development would be similar to the previous feature, but using a checklist in GUI versions, as you can see on the next picture.

---

[9] http://dle.rae.es
[10] http://www.rae.es/recursos/diccionarios/dpd
[11] http://aplica.rae.es/ortografia
[12] https://www.fundeu.es
[13] http://www.wikilengua.org
[14] De Sousa, J. M. (2015). *Manual de estilo de la lengua española*. Madrid: Trea.
[15] http://www.wikilengua.org/index.php/Lista_de_manuales_de_estilo_en_Internet
[16] http://www.nltk.org/_modules/nltk/tag/brill_trainer.html
[17] http://www.learningspanish-spain.com/correct-the-text.html



Guide for Spanish speakers

It is a priority to give a **proper guide to the Spanish users who want to collaborate with the project**. Since many potential collaborators are lost because they do not find documentation in Spanish, it will be useful to describe in Spanish all that is necessary for contributing.

Project Schedule

My availability will be about **40 hours per week** to dedicate exclusively to this job.

February 12, 2018 – March 27, 2018 (Organizations Announced – Students Apply)

Since the list of accepted organizations was published on the GSoC website, I have been **in touch with LanguageTool development community**. I wrote to the contact person, Daniel Naber, and I contacted the Spanish maintainer, Juan Martorell. I immersed myself in the code and I sent some pull requests. Then I have prepared this proposal.

April 23, 2018 – May 14, 2018 (Community Bonding period)

- ➔ Comparison of external taggers for the detection of new possible rules.
- ➔ Analysis of corpus of errors.
- ➔ Defining typology of errors.
- ➔ Exploring data sets and other documentation I have been collecting.
- ➔ Finding possible bugs.

May 14, 2018 – May 20, 2018 (Coding: stage #1.1)

- ➔ Spanish disambiguator: starting with complex constructions, such as conditional and subordinates sentences.
- ➔ Testing for ensuring that regressions and false positives do not increase.

May 21, 2018 – May 27, 2018 (Coding: stage #1.2)

- ➔ Spanish disambiguator: facing the big challenge of attributive sentences.
- ➔ Testing for ensuring that regressions and false positives do not increase.

May 28, 2018 – June 3, 2018 (Coding: stage #1.3)

- ➔ Spanish disambiguator: developing rules for the shorter sentences.
- ➔ Testing for ensuring that regressions and false positives do not increase.

June 4, 2018 – June 10, 2018 (Coding: stage #1.4)

- ➔ Spanish disambiguator: reviewing all rules and simplifying when it is possible.
- ➔ Final testing and evaluating the performance of the rules.
- ➔ Documentation of the disambiguator.
- ➔ Preparing the **first deliverable**.

June 11, 2018 – June 17, 2018 (Coding: stage #2.1)

- ➔ Spanish rules: reviewing, simplifying and reorganizing all of them.
- ➔ Testing for ensuring that regressions and false positives do not increase.

June 18, 2018 – June 24, 2018 (Coding: stage #2.2)

- ➔ Spanish rules: localizing developed Java rules and starting to add the new ones.
- ➔ Testing for ensuring that regressions and false positives do not increase.

June 25, 2018 – July 1, 2018 (Coding: stage #2.3)

- ➔ Spanish rules: ending with Java rules and starting to write new XML rules.
- ➔ Testing for ensuring that regressions and false positives do not increase.

July 2, 2018 – July 8, 2018 (Coding: stage #2.4)

- ➔ Spanish rules: ending XML rules and reviewing the whole list of rules.
- ➔ Final testing and evaluating the performance of the rules.
- ➔ Documentation of the rules.
- ➔ Preparing the **second deliverable**.

July 9, 2018 – July 15, 2018 (Coding: stage #3.1)

- ➔ Adding the support for different style guides.
- ➔ Testing.

July 16, 2018 – July 22, 2018 (Coding: stage #3.2)

- ➔ Adding the feature for implementing specific terminologies.
- ➔ Testing.

July 23, 2018 – July 29, 2018 (Coding: stage #3.3)

- ➔ Including examples of use for these features in the Spanish module.
- ➔ Documentation of the new features.
- ➔ Writing the guide for Spanish speakers.

July 30, 2018 – August 5, 2018 (Coding: stage #3.4)

- ➔ Solving any potential error or conflict that may have appeared during the project.
- ➔ Complete reviewing of the project.
- ➔ Preparing the **third and last deliverable**.

Bio

I am a last-year **Computer Science** student and a Marketing & Communication professional. Nowadays, I am studying at **Chalmers University of Technology**, Sweden (in an exchange program of the University of A Coruña, Spain).

I have been in touch with LanguageTool community since the list of accepted organizations was published on the GSoC website. Because I have always been some kind of **language nerd** I would like to work in a project that allows me to merge both Computer Science and Communication skills.

Also, **Spanish language is my mother tongue**. I have achieved Spanish proficiency and interest on language analysis, processing texts by hand, and collecting data sets. Furthermore, I have previously studied **Communication and Language theory** and I worked as a journalist (even I have made manual sentiment analysis of news). Regarding NLP, I have used **Pattern and NLTK**.

Contributions to LanguageTool project

I have sent some **pull requests** to the LT repositories (all of them were **accepted**):

1. [es] lowercase: months, days of the week, and seasons of the year
2. [es] fix typos and outdated URLs in grammar.xml
3. Add a HTML file for allowing filtering false friends by language (issue #58)
4. [es] add Wikipedia rule
5. Detect Roman numerals (issue #816)
6. [es] add translations to index.php

Also, you can view my full CV on LinkedIn (with some recommendations from previous jobs) and my GitHub profile (where you will find samples of my programming skills).

Project-related programming skills

- ➔ **Java**
- ➔ **XML**, XSLT, XSD
- ➔ HTML5 + CSS3
- ➔ JavaScript
- ➔ Others: Python, SQL, C, OCaml...

Language skills

- ➔ **Spanish (mother tongue)**
- ➔ English (full proficiency)
- ➔ Galician (native proficiency)