Adopting the unreleased Romanian-Catalan pair and upgrading other outdated pairs to the monolingual module system

Google Summer of Code 2018

Contact information

Name: Marc Riera Irigoyen
Location: Barcelona, Spain
E-mail: marc.riera.irigoyen@gmail.com
IRC: mriera_trad
GitHub: MarcRiera
Timezone: UTC+02:00

Why is it you are interested in machine translation?

As a Traslation and Interpreting student and will-be professional translator, machine translation is interesting due to its dramatic improvement over the last few years and its increasing prevalence in society. Therefore, I am very interested in how translators can make the most out of it and use it responsibly.

Why is it that you are interested in the Apertium project?

Apertium is not only the organization behind a great open source project; it is also a very welcoming family of collaborators and language enthusiasts. After successfully participating in GSoC 2017 with Apertium and completing my project, I felt motivated to keep contributing regularly. Now, taking part of GSoC 2018 with Apertium again is the best way to boost development, make it gain even more importance and reach new users.

Which of the published tasks are you interested in? What do you plan to do?

I am interested in upgrading several language pairs to ease future development and bring one of them (Romanian-Catalan) to release status. Development of the Romanian-Catalan pair will take place during the first two thirds of the programme, and the upgrade of the other pairs will take place during the last third.
Apertium currently uses independent language modules for pairs, so monolingual data is shared between pairs. However, this was different in the beginning, when pairs are self-contained and included monolingual data. Lots of pairs have been upgraded to the new system, which is more efficient and allows users to easily share work, but there are still a few that have not been upgraded. Consequently, potential language developers avoid them due to the extra difficulty and the pairs quickly become out of date.
One of these language pairs, Romanian-Catalan, was upgraded recently to use the new system. Despite still being unreleased, it contains a basic but decent bilingual dictionary and transfer rules for the Romanian > Catalan direction re-used from a very similar pair, Romanian > Spanish. The Apertium wiki also contains several documentation pages related this other pair providing very useful information. While many entries in the bilingual dictionary

are broken as an effect of the upgrade, the two languages are close to each other and with some intense development the pair would be error-free and ready for release. A working direct Romanian-Catalan pair would also be unique to Apertium, as other proprietary machine translation platforms (such as Google and Yandex) use English as a pivot language and the results could be much improved.
The other pairs will be first upgraded to use monolingual modules and then cleaned until they are testvoc-clear.

Title

Adopting the unreleased Romanian-Catalan pair and upgrading other outdated pairs to the monolingual module system

Reasons why Google and Apertium should sponsor it

Currently there are no machine translation systems offering direct translation between Romanian and Catalan available to the general public. English is commonly used as a pivot language, and the results are sometimes worse than what could be achievable with direct translation, because the two languages have a lot of common (both being Romance languages). A release of such a pair could easily offer good results, and the work would be freely available for anyone to make use of it.
The upgrade of several other pairs would be very positive for Apertium and would attract language developers to the project. This upgrade will need to be done sooner or later, and Google Summer of Code is an easy and fast way to get closer to having all the old pairs upgraded.

How and who it will benefit in society

On the one hand, the development of a direct Romanian-Catalan language pair will create an opportunity to strengthen the relations and cultural exchange between the two languages. Despite being quite distant geographically, migration of Romanian speakers to Catalan-speaking areas has connected both communities, and the development of such a machine translation pair would possibly create a stronger bond between both communities and hopefully create more interest into each other from both sides.
On the other hand, the upgrade of several Apertium language pairs to the new monolingual module system will hopefully attract language developers who currently have not considered contributing due to the difficulty of getting the old pairs to work and update them.

List your skills and give evidence of your qualifications

My mother languages are Catalan and Spanish, and I can also speak English, Japanese and Romanian. I am a Linux and open source software user since many years ago and I have contributed to the translation of some programs, such as the openBVE railway simulator. I know the basics of C# and I am very familiar with XML and most of the usual Apertium language data formats.

List any non-Summer-of-Code plans you have for the Summer

I have no plans other than Google Summer of Code, so I will be able to dedicate at least 30 hours a week to the project. If there were any change of plans affecting my dedication to the programme, I would make sure to compensate them with extra time.

My plan

Major goals

- Have at least four pairs upgraded to encourage future development and maintenance
- Reach clean testvoc in all the pairs

Romanian-Catalan

- Add ~2,000 new bidix stems a week to improve coverage
- Rewrite and expand the transfer rules for Romanian > Catalan to take advantage of chunks
- Write rules for Catalan > Romanian, including rules described in Apertium-es-ro documentation
- Write CG rules to improve Romanian disambiguation

Workplan

Week	Dates	Goals	Bidix	WER / PER	Coverage
Post-application period	28 March - 13 May	• Find at least four pairs that need an upgrade • Build frequency lists for Romanian and Catalan • Begin fixing broken bidix entries	~13,000	~36% (ron > cat) ~61% (cat > ron)	79% (Romanian) 78% (Catalan)
1	14 May - 20 May	• Expand bilingual dictionary • Rewrite transfer rules (Romanian > Catalan)	~15,000	~34% (ron > cat) ~60% (cat > ron)	80.1% (Romanian) 79.1% (Catalan)
2	21 May - 27 May	• Expand bilingual dictionary • Rewrite transfer rules (Romanian > Catalan) • Testvoc: adj	~17,000	~32% (ron > cat) ~59% (cat > ron)	81.1% (Romanian) 80% (Catalan)
3	28 May - 3 June	• Expand bilingual dictionary • Rewrite transfer rules (Romanian > Catalan)	~19,000	~30% (ron > cat) ~58% (cat > ron)	81.9% (Romanian) 80.9% (Catalan)
4	4 June - 10 June	• Expand bilingual dictionary • Add transfer rules (Romanian > Catalan) • Testvoc: n	~21,000	~28% (ron > cat) ~57% (cat > ron)	82.7% (Romanian) 81.7% (Catalan)
5	11 June - 17 June	• Expand bilingual dictionary • Add transfer rules (Romanian > Catalan) • Write documentation **First evaluation**	~23,000	~26% (ron > cat) ~56% (cat > ron)	83.4% (Romanian) 82.4% (Catalan)
6	18 June - 24 June	• Expand bilingual dictionary • Add transfer rules (Catalan > Romanian) • Testvoc: vblex	~25,000	~25% (ron > cat) ~53% (cat > ron)	84.1% (Romanian) 83% (Catalan)
7	25 June - 1 July	• Expand bilingual dictionary • Add transfer rules (Catalan > Romanian)	~27,000	~24% (ron > cat) ~50% (cat > ron)	84.7% (Romanian) 83.6% (Catalan)
8	2 July - 8 July	• Expand bilingual dictionary	~29,000	~23% (ron > cat)	85.3% (Romanian

		• Add transfer rules (Catalan > Romanian) • Testvoc: others		~47% (cat > ron)	) 84.2% (Catalan)
9	9 July - 15 July	• Expand bilingual dictionary • Add transfer rules (Catalan > Romanian) • Write documentation **Second evaluation**	~31,000	~22% (ron > cat) ~45% (cat > ron)	85.8% (Romanian) 84.7% (Catalan)
10	16 July - 22 July	• Upgrade first pair • Make the upgraded pair testvoc-clean			
11	23 July - 29 July	• Upgrade first pair • Make the upgraded pair testvoc-clean			
12	30 July - 5 August	• Upgrade first pair • Make the upgraded pair testvoc-clean			
13	6 August - 14 August	• Upgrade first pair • Make the upgraded pair testvoc-clean **Final evaluation**			

Coding Challenge

As a coding challenge, I created a new Romanian-Catalan pair with monolingual modules but reusing existing data. It was later merged into the old one to keep commit history. In addition, I have already been testing the pair and working to make it as testvoc-clean as possible. The work is distributed in several commits in the apertium-ron and apertium-ron-cat repositories in Github, but the testvoc improvements and most of the work regarding transfer rules can be seen here: [1]