Creation of a Romagnol-Italian translation pair

Name: **Lorenzo Tosi**
E-mail address: **ltosi@edu.hse.ru**
Other information: IRC, GitHub: **erba994**

About me

I entered the world of computational linguistics in this academic year after graduating in translation between Italian, English and Russian in my bachelor at the University of Bologna. My interest for the preservation of my dialect (Romagnol), together with the possibility offered by text-to-speech technologies brought me into this path, and I am currently based in Moscow and enrolled in Higher School of Economics. I speak 5 languages (Italian, English, Russian, Slovak and Romagnol (Cesenaticense dialect)).

Why Apertium?

I think an open-source machine translation system, not relying on proprietary software, is the best way to achieve the goal of giving a future to endangered languages and getting people to work on them. Giving decent quality translations and "digitizing" them for NLP tasks (e.g. text generation, NER) is the only way to preserve the knowledge of them from the oblivion of social changes. It also opens the path to linguists and amateurs, with no or little knowledge of coding, for working and extending their capabilities and focus their efforts.
Why I say that? Following a meeting with Gilberto Casadio, volunteer and researcher in the Friedrich Schurr Foundation, the main institute devoted to the preservation of Romagnol language, I understood that there is no knowledge of a framework that makes the development of a usable and effective online dictionary possible, let apart harder tasks. The institution is thinking of opening a page on the italian Wiki Dictionary branch, but this framework is not powerful enough for solving the issues I will mention in the next sections, and is still not seen as an effective solution for preserving the language.
Working on the Apertium framework would make it possible to develop a system able to reach the main goal of the association, and at the same time the way of creating and updating mono and bilingual dictionaries, it is relatively simple and appealing even for amateurs and volunteers that want to dedicate time to its mantainment, with a little training.
A lot of associations working with central and northern dialects in Italy could be facing the same issue, and being able to start this project and push this knowledge could foster a positive effect, with more and more people entering the world of Apertium and developing dictionaries. The social effect of this project could be invaluable for the preservation of local italian cultures and the development of them. A lot of quantitative analysis methods previously unavailable on original texts could be finally within reach for researchers.

Challenges

The task of producing a translation system requires important efforts for two reasons: the language does not have a standardized orthography and the phonetic rendition (other from lexical choices) varies among speakers of different cities.

Luckily, there is enough written material for trying to solve these two problems (I would note down two well-written books on the topic for reference purposes: L'Ortografia Romagnola, written by Daniele Vitali and published by Il Ponte Vecchio, and Dialetti Romagnoli, written by Daniele Vitali and Davide Pioggia and published by Pazzini), but the task will need to develop an extensive ruleset to apply in the pipeline.
Romagnol is also a low-resource language. The material available nowadays is related mainly to poetry and narrative, though good Romagnol-Italian dictionaries were released, having though the same varietal issue abovementioned (Libero Ercolani 2002 and Adelmo Masotti 1996 the most recent ones). Only a very small amount of texts is available online, making very difficult to create an omogenous corpus.
Romagnol is also an interesting language for the phenomena it presents (apart from the above mentioned ones): a very big inventory of phonemes (a medium of 15 different vocalic phonemes per dialect) and metaphony for plural formation are two of them.

Project Timeline

My idea is to focus the task first on the release a model on a single dialect, namely the Santarcangiolese one as the most important circle of dialect language writers comes from there (Tonino Guerra, Raffaello Bandini, Nino Pedretti being the most famous ones). Davide Pioggia wrote some literature on the grammar and phonology of this specific variety that can be very useful to code and develop an extensive morphology. I will adapt the lexicon from dictionaries already released to this variety and use the available literature to develop a working model. This model with a working set of rules can be then easily adapted for all the varieties for future work.
I would like to create a Romagnol-Italian dictionary to help the preservation of regional material first of all, and I also consider it a more useful source-target couple than viceversa for working on the language.
My plan is to work on a very-reduced lexicon derived mainly from openly-available literature (hand-translated literature as to have a benchmark, and dialect one) to develop closed classes and main open classes in the first month, work on extending open classes and code effectively transfer rules in the second month, and work on expanding heavily the lexicon from available dictionaries, and check for rare exceptions in rules, to create an extensive model in the third month. The result should be a working translation pair by the third month, that will open the path for developing a phonetic and orthographic transducer for making it adaptable to all the varieties in the future, and also will make it possible for more maintainers to work on it as explained above.
The timeline is not casual: I expect to be a bit busier the first month as will be still exam time in my University, therefore I expect to work around 35 hours per week on the project there, but I will be completely free and working exclusively on the project in the last two months. I plan to return home at the end of the first month and stay there until the end of the project, to get more literature and missing material and to be able to discuss and work with researchers and volunteers in Friedrich Schurr Association or on the field, to better develop the system.

Why me?

I am a master student in Computational Linguistics. This year we extensively worked in developing efficient CG rules, POS taggers, phonetical transducers and coreference extractors in my classes, therefore I have knowledge of what is under the hood in MT systems. We extensively worked with bash scripting and Python libraries, and I have knowledge of ML techniques and algorithms, along with statistics and probabilities, being able to discern and analyze HMM and Bayes methods also on an algorithm point of view. The corpus creation interest comes from my bachelor experience, and having studied the main issues and techniques used by professional translators and worked extensively with dictionaries, I have experience on how to use and create corpuses for translation works. I am also a L2 speaker of my dialect, therefore having the language proficiency needed for the task.

Coding Challenge

I am creating from scratch rules and lexicon for the coding task proposed in the Apertium Wiki page. I translated the story into Italian and I am currently working on creating the Rgn monolingual dictionary and translation pair. I uploaded all in GitHub at this link.