Multilingual October/November 2016, Community Lives
Apertium
By Jeannette Stewart

We all know what it is to learn a language. We, of course, acquire our native language as we grow up. Then comes the joy, challenge or chore (depending on our capabilities) of learning languages in education. There are many ways in which languages are taught and learned, but the end result is the same in giving us the ability to communicate beyond our native language group members. It enlarges our community. Now we live in the digital age and are grappling with ways to best facilitate multilingual global communications. Languages and computers are the building blocks that are raising a new edifice of translation, also known as localization. Since one of the main uses of computers is to automate tasks performed by humans, it's a logical step to use computers for this. But how do we teach computers natural languages? Certainly we can load source and target dictionaries into computers’ memories, but how do we make the leap from facilitating machines to form the equivalences that trained professional translators would? How can we effectively enable machine learning? With natural languages, two main strategies have been devised. One uses statistical data and immensely fast processing power to translate texts. The other uses a rule-based approach which adheres to syntactical, morphological and semantic forms and word pairs to enable lexical transformations. There are a number of impressive rule-based systems in use, but the undisputed leader in the Free and Open Source field is Apertium.
The first thing I noticed when learning about Apertium is the vibrant and committed community that facilitates its function. The second thing is that it is Free and Open Source. Even if a developer wishes to customize aspects of the platform for their own distribution, they must agree to make any extensions to the base-release open source too. A third is that they are as dedicated to the lesser stars in the language firmament as to the big ones, maybe even more dedicated. Commitment, passion and a commanding grasp of natural language processing technologies underpin their impressive achievements and are propelling them on a path to even greater success.
Apertium has grown out of work done at the University of Alacant (Catalan for Alicante) on the Costa Blanca in Spain. In 2003 Mikel Forcada issued a call for support to a number of machine learning research groups in Spain with the objective of lobbying the government for support in building a Free and Open Source (FOS) machine learning application. One of computing’s less widely known strategies is reuse and that’s how Apertium came into being. It is, in fact, a FOS rewriting of two Machine Translation (MT) precursors, interNOSTRUM and Traductor Universia. The result was an MT application that worked extremely well with related languages, specifically Spanish, Catalan and Galician.
As the instigator of Apertium, Mikel Forcada combines an impressive array of academic achievements with technical accomplishments that have taken him to the forefront of the MT community. He is currently a professor at the University of Alicante in Spain and President of the European Association for Machine Translation. He is a passionate advocate of Free Open Source technology and
believes strongly in teaching his translation and interpreting students to use MT efficiently in their work. His success with Apertium and his research work at Promsipit Language Engineering, which was created to respond to the demand for professional services from Apertium, have earned him a well-deserved place on the cutting edge of translation studies.

Apertium is a distinctive name and Mikel Forcada explained its ‘etymology’ in an email to me: “I invented the name, more or less: I wanted a rather unique name. I wanted "Apert-" for open, and I searched the net for Latin-sounding names (as we had had interNOSTRUM before). So "Apertum" (open, neuter gender) and "Apertius" (more open, neuter gender) could be found, but the non-Latin word "Apertium" was found only once or twice in the net (mainly as a typo in manuscript transcriptions), and I liked it.”

According to Forcada, the design principles of Apertium are conservative and the flow of work through its various stages makes great logical sense. The approach is one which renders an understandable translation which is then refined, by making it understandable for gisting and for post-editing efficiency, to its finished form. He explains that with a bilingual translation, transfer is the intermediate step in an indirect machine-translation system. This occurs just after analysis of the source-language and before target-language generation. Transfer may be "deeper" or "shallower", depending on the level of abstraction that analysis reaches. "Shallow transfer" usually means that the analysis does not construct entire parse trees, but rather identifies phrases and then carries out local transformations on them.

There is, as all linguists know, a long-standing debate on using word-for-word translation in which conveying the sense of a source text is sacrificed for rendering literal meaning. Apertium’s approach to translation is to generate a ‘reasonably intelligible’ text which does not pose much difficulty in working between related languages like Spanish and Portuguese. With such target material, it is then possible, if required, to clarify the substance of texts with additional stages of refinement like ‘lexical processing’ and so on. Apertium adopted this approach as a direct result of the experience gained from building on the work of its predecessors. Another advantage of the Apertium philosophy is that it is possible to generalize the main concepts of a word-for-word translation to deal with harder, not-so-related language pairs, so that linguistic complexity is kept as low as possible. Interestingly, Apertium is built on the idea of separating data (language-pairs) and the algorithms (translation engines) that process them. This use of a language-independent engine along with a set of tools for managing language data in an expanding database of language pairs gives great flexibility in catering for different document formats and content. One more aspect that characterizes the community involvement that Apertium is seeking to leverage is their commitment to Free Open Source software. With this approach and thanks to the GNU General Public License, high quality, collaborative project work is enabled and contentious matters like copyright and the use of license-centered applications are avoided. The result is a slick MT service that is available for use by all.

Early success has enabled Apertium’s growth, including the development of an active and productive community. The community has resulted in a wiki which is used for documentation purposes and in a software repository where contributions are made. This allows all members to participate in Apertium’s work, extending across many languages allowing translators to benefit from a collective wisdom in creating new language pairs. Further
evidence of the robust health of the community can be seen in the number of externally-developed tools and functions. The Apertium community presently boasts four hundred contributing developers helped by seven administrators. Not all are active all the time, usually there is a noticeable spike of activity during summer. Over the last ten years there have been 78,000 commits (a commit being a modification of the public repository) some of them contributing just a single word, some 1,000 words. The community supports 41 stable language pairs, with another 100 in development, and is using SourceForge for the commits, a web-based platform that allows users to control and manage FOS projects, to enable linguists to contribute to dictionaries and share data. Although the Apertium community is not formally incorporated, they have managed to add a level of governance by creating an elected Project Management Committee. Its secretary is Fran Tyers. Tyers first met Forcada in 2006 when he attended a workshop in Italy on Language Technology for Minority Languages where Forcada was giving a presentation on FOS for language technology. Tyers was impressed and decided to volunteer and help make Apertium easier to use and install for linguists. Tyers at the time was contemplating studying further for a PhD and soon found himself studying at the university of Alicante under Forcada’s supervision. One of the projects that Tyers was involved in was a project on the Breton language, which is on UNESCO’s list of severely endangered languages. Tyers used to spend Christmas holidays in Brittany and was always fascinated by the Breton language, a Brythonic Celtic language dating back to the Middle Ages. On one such vacation, he decided to create a morphological analyzer based on the work of Roparzh Hemon, the distinguished Breton scholar, in order to develop a rudimentary Breton-into-French system. He adapted it from the Apertium Spanish-into-French system and the results were excellent. After he emailed the Breton language board, he was invited to a meeting with the Director, Fulup Jakez, and soon it was agreed to instigate a one-month project at Alacant University. The project was a great success and to this day Tyers is still involved and contributing to it. Tyers intends to follow an academic career and is currently in Tromso, Norway working on a postdoctoral project, but he is still deeply committed to Apertium. As he explains, Apertium’s mission is to support native speakers of marginalized languages by bringing the benefits of both free machine translation and free language technology to their languages so that their place and use are safeguarded in an increasingly connected world. A language without universally available technologies supporting it is dead in the digital world! Apertium’s success has been so great that the demand for professional services around the open source platform necessitated the formation of a support company. Prompsit was created in 2006 by eight partners, five working at the University of Alicante and three full-time employees. Gema Ramirez-Sanchez is currently the CEO, but only ten years ago, she was another of Forcada’s student and an Apertium regular. She brings a wealth of experience garnered from her years in academia and in helping Prompsit’s clients understand the benefits that MT offers in this rapidly-moving increasingly important tech sector. Prompsit was born from all the requests for customization coming from private companies requesting services and technical support. However Prompsit is also an active contributor for
the vibrant and engaged Apertium community. A Prompsit employee always participates in the Apertium Project Management Committee of governance and helps with the Google Summer Code events.

Prompsit adds client data, when they agree, to the community and this creates a never ending benefit for them. As Ramirez-Sanchez puts it, “They get what they want and much more!” Delighted customers are always a great sign that a business is delivering. Another sign is when clients show a desire really to get involved in the community themselves as is the case with the Open University of Catalonia (UOC), which has contributed to Apertium for years. They have also found that the community is the best source of linguists to work in very different languages, to teach Apertium worldwide, making it available not only as an isolated MT application but in combination with other CAT tools used by professional translators. For example, MateCat or OmegaT are win-win cases of open-source communities working together.

When a community functions in a way that produces its own momentum, the success stories it spawns can be as diverse as they are unexpected. For example, the most downloaded and used pair in Apertium is Norwegian Nynorsk - Norwegian Bokmål. Every city council in Norway has to decide which of the two Norwegians is the official one in their town and provide bilingual versions that they produce by machine translating and post-editing. Students at school learn both Norwegians and they use Apertium to check and do their homework. Apertium is really helpful for them as a tool to generate raw translations and to learn.

Prompsit and Apertium have actually helped in the standardization of languages such as Occitan which has many variants and has no widely accepted standard version. Apertium can accept any variant of any word in source language but needs to produce just one target. A group of renowned linguists of all variants of Occitan worked together for a couple of years in deciding which was the preferred form of words having discussions that lasted for hours for a single word but finally making a historical agreement.

Apertium’s strengths have resulted in it being the first machine translation system or linguistic asset for language technologies for a many languages such as the afore-mentioned Breton, Asturian, Aragonese, Chuvashian and others. It contributes significantly to enlarging content in Wikipedia. The Wikimedia Foundation uses Apertium to help the different Wikipedia linguistic communities to get raw versions of articles written in other languages, speeding up the creation of new articles in their own language. They give voice to communities and languages that are not very popular and do so with very humble resources but passion for languages.

Ramirez-Sanchez recalls the instance when developing the Afrikaans to Dutch pair often loosing contact with the person behind it for days as he had very limited internet connection, but he never gave up and Apertium proved itself once again.

Ramirez-Sanchez is justly proud of their achievements as she explains, “We started with two language pairs: Spanish-Catalan and Spanish-Galician, with a group of 6 people in our corner of the world. Apertium hosts now 41 stable pair and many other in development, 400 people registered as contributors, 7 administrators, public and private funding, opportunities for research, business and, above all, for languages and linguistic communities. We were not even dreaming of this when we first released Apertium!”
In learning about this amazing community of dedicated computational linguists and their awe-inspiring achievements, I have come to understand much better the power that language has to enable so many aspects of life here on earth. Yet they remain unsung heroes of the global economy. Their commitment to the preservation of our linguistic heritage is exemplary. They are worthy of mountains of praise and oceans of acclaim, but they remain humbly devoted to their work and making it the very best they can. I do struggle to understand why language professionals toil without much appreciation or recognition for the value they add to our vast range of commercial and cultural enterprises. Perhaps it’s simply because language is common to every one of us. Perhaps because the global community is everywhere that it just doesn’t register how special the gift of language is. But let’s not forget, Apertium derives from openness so let’s follow Apertium’s lead and spread the word.