**Apertium GSoC 2011 Proposal: Machine Translation between Dhivehi and Sinhala**

**Nalaka Buddika Warnakula**

# Name:

Nalaka Buddika Warnakula

# Email:

nalakabuddhika@yahoo.com

# Other Contact information:

IRC: Nlks (freenode)

Skype: nalakabuddhika

# Why is that you are interested in Machine Translation?

My interest in Machine Translations arose when I used to take a third language in my high school. Back then I used machine translation software to translate phrases from English to French and vice versa. Though not accurate all the time it was invaluable tool in my learning journey and made me curious about the translation technique, thus began my interest in Machine Translation. Also I have realized that machine translation is more than just a simple text translator; it conveys a way of life embedded in the translated text. The fact that the word order in French is different from English or my native language showcases the different philosophical approaches of languages. Thus it acts as a cultural catalyst that facilitates greater cultural appreciation. Similarly in this globalized world there is more and more communication and exchange of ideas when different communities interact. I feel that language barrier is a main issue that prevents more efficient exchange of ideas and I sincerely believe that machine translation can bridge the language barrier.

# Why is that you are interested in apertium project?

The main reason, apertium been an open source project. I find that Apertium project exceptionally showcases the spirit of open source; the passion and zest to share the knowledge with others. Apertium been community developed and accessible tool for all makes it have a far reaching impact on a larger user group who is unable to afford commercial software.

Another reason why I am interested in Apertium project is its flexibility and openness in the language pairs. When I had a look at the incubator page of Apertium, I saw language pairs that are not commonly present in many translator tools. I was pleasantly surprised to see my native language, Sinhala in a language pair. Even though these language pairs are still in incubator level, I felt that the initiative Apertium project has taken to include lesser known languages was a bold step in the direction of conserving and enriching particular languages and I felt that it is an avenue where I can contribute.

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

## Title

Apertium-Dh-Si: Machine Translation between Dhivehi and Sinhala

## Why should Google and Apertium sponsor it?

To begin with this will be a fresh language pair added to Apertium project. When I first thought of contributing to Apertium project I thought of enhancing the available Sinhala-English Language pair. However after community interaction through IRC with mentor, I realized Apertium does not sponsor language pairs where English is one of the languages. So I was asked whether I wanted to consider Dhivehi-Sinhala Language pair. Dhivehi is the language spoken by a population of 400 000 people in Maldives and Sinhala is spoken in Sri Lanka. Sinhala and Dhivehi are now not mutually intelligible, but they clearly form a subfamily, since the two share many features, at all levels of structure[[1]](#footnote-1). Though it will be a new language pair for the Apertium project, the similarities in the two languages would make this proposal a feasible and practical one given the scope of the GSoC duration.

While I was browsing through one of the user pages of Apertium wiki, I came across this excerpt that described the last surviving member of a particular language group. Though both Dhivehi and Sinhala are not endangered languages, they are both spoken by smaller number of people in the global context. Especially the Maldivian islands threatened by the climate change and rising water levels in the Ocean, adopting Dhivehi and Sinhala as a language pair would contribute the preservation of the language and their ways of life even though such an initiative would probably have an extremely small impact to the bigger picture.

Also I am currently involved in a three year engineering design project in my university to develop a prototype device to decrease the level of misunderstanding that occurs in communication between different languages by reducing the verbal and non-verbal communication barriers. Even though the apertium project is not entirely related to the my engineering design project, they share common ground and both these projects can mutually benefit from each other.

# How and who will benefit from the project

Neither Sinhala nor Dhivehi is a language that can be found in Google or any other main stream translating tool. So the main beneficiaries would be the language speakers and this could lead to greater cultural exchange predominantly in the field of literature.

Apart from the particular user group, it will help in the preservation of the given languages. Even though the scope of GSoC is limited to 3 months, this would be a good start to build up on the language pair and if my time permits, I would be more than happy to contribute to the language pair even after the GSoC 2011.

# Work plan

The most challenging part in this undertaking would be the fact there are very few resources in Dhivehi, even when it comes to even dictionaries. So I have to compile my own Dhivehi- Sinhala dictionary. As a rough guide I would aim for 3000 and revise accordingly.

One important thing to note is that in both Dhivehi and Sinhala, the language used in written communication and oral communication are completely different. I will be concentrating only on the written communication component. Also Dhivehi (spoken in Maldives) has six predominant dialects; *Malé*, *Huvadhu*, *Mulaku*, *Addu*, *Haddhunmathee* and *Maliku (Mahl) dialects[[2]](#footnote-2)*. Māle been the standard language, I will use it in this project.

During the community bonding period, I am going to familiarize with the Thaana script, which will be the mode of input for Dhivehi. Thought it is stated that the community bonding period officially starts on the 25th, I will not be able to start until 7st May 2011 as my final semester exam will be on the 6th May. I would allocate 2 weeks for me to study the script and interact more with the Apertium community in IRC.

## Deliverables

Here I will showcase what I would like to deliver every week (if any) during the project timeline. I have made a few assumptions about dictionary entries. Since I have to manually compile the dictionary, I tried to calculate the time taken for a single entry. So it took me roughly 3 minutes to translate a Dhivehi word from Roman script to Thaana and enter the Sinhala equivalent. I am planning to work 8 hours a day and I calculated it would take me a little over a 4 weeks to complete the Dhivehi-Sinhala dictionary.

|  |  |  |
| --- | --- | --- |
| Week | Task | Deliverable(s)/ If any |
| 1 | Working on translation of Nouns considering case systems and stem types including definiteness and indefiniteness. | apertium-Dh-Si.Dh.dix, morphological Dhivehi- Sinhala dictionary |
| 2 |
| 3 | Work on the adjectives, Numerals and personal pronouns |
| 4 | Work on Verbal system |
| 5 | Basically the same Task as above but this will be in creating the Bilingual dictionary between Dhivehi and Sinhala  Preparing for midterm evaluation | apertium-Dh-Si.Dh-Si.dix  Bilingual Dictionary and midterm evaluation report. |
| 6 |
| 7 |
| 8 | Setting up Transfer rules for Nouns, adjectives and Numerals | apertium-sh-en.sh-en.t1x  Transfer rules between Dhivehi and Sinhala |
| 9 | Transfer rules for personal pronouns and verbs |
| 10 | 1 week buffer period to count for any delays due to unforeseen circumstances |  |
| 11 | Debugging and testing |  |
| 12 | Clean up and look through the dictionaries and final evaluation of the project | Properly documented/indented dictionaries handed to Apertium with final evaluation |

Should I finish before the stipulated deadlines and have sufficient time remaining. I would like to go into parts of speech tagger. However it could be a secondary objective to a post GSoC contribution should the time do not allow.

# List your skills and give evidence for your qualifications

I am a mechanical engineering student from National University of Singapore. You would wonder why a mechanical engineer would involve in a Machine translation project. As I said before, I am working on a 3 year design engineering project that has common grounds with machine technology and I feel that both projects will mutually benefit from each other.

Though I have been programming for some time, this will be my first involvement in an Open source project. I have a strong background in C and C++ programming and have taken courses pertaining at my University. Though I do not have any formal qualifications with a scripting language, I have self-taught myself python and have a good knowledge of XML. I think my knowledge in this area would be more than sufficient for my proposal given it is a new language pair and it is marked as an entry level task.

# List any non-summer of code plan you have for summer

GSoC project will have my outmost priority should I be selected for the program. I have no other plans for summer except 2 days in June, where I would be away to organize an orientation camp for the incoming freshmen at the University but it would have minimal impact on working on the project. My University will be starting in early august but it also should not be a major problem as the tutorial and workshops do not commence until the first week of September.

1. [Journal of the American Oriental Society, The](http://findarticles.com/p/articles/mi_go2081/), [July-Sept, 2007](http://findarticles.com/p/articles/mi_go2081/is_3_127/) by [James W. Gair](http://findarticles.com/p/search/?qa=James%20W.%20Gair) The Dhivehi Language: A Descriptive and Historical Grammar of Dhivehi and Its Dialects. 2 vols. By SONIA FRITZ. Beitrage zur Sudasien-Institute Universitat Heidelberg, vol. 191. Wurzburg: ERGON VERLAG, 2002. Pp. xvi 270, vi 280. [euro]75. [↑](#footnote-ref-1)
2. http://en.wikipedia.org/wiki/Dhivehi\_language [↑](#footnote-ref-2)