KURDISH SPELL CHECKER PROJECT

▪ Brief Introduction

Using word processors in Kurdish writing is growing in a very fast pace. Currently, government ministries and departments, legal institutions, business offices, media channels, universities all use word processors in their daily work. This growth is expected to continue as computers and electronic devices become more and more prevalent in Kurdistan. Spelling is an important aspect of language writing. Poor spelling can interfere with communication between the writer and the reader. Word processer use spell checkers to suggest corrections to misspelled words. Unfortunately, existing word processors do not come with built-in spell checkers for every language. Individual nations create their own customized dictionaries and add them to the word processors for error correction. Currently, Kurdish language lacks a reliable spell checker. This limitation need to be resolved. Overcoming the lack of spell checker problem will flourish Kurdish writing and helps word spelling standardization. The objective of this project is to build a reliable and comprehensive Kurdish spell checker that can be used by public in Kurdistan.

▪ Background and statement of the problem

Word recognition and automatic correction techniques have been studied in a large spectrum of computer applications. These include word processors, machine translation, search engines, and voice recognition. While almost all modern human spoken language has one or more spell checkers, Kurdish language lacks even a very basic one. Hence, building a Kurdish spell checker will have an outstanding effect on Kurdish language processing applications.

As a first step towards building a robust, reliable and comprehensive Kurdish spell checker, we conducted a preliminary investigation to determine two aspects of the Kurdish language processing. First, we studied the latest Unicode standards [1] in order to gain understanding of the Kurdish glyph representations. Second, we checked the ability of processing Kurdish text using modern computer programming languages [2]. The initial investigation ended with positive results as we find out that there is a unique code for almost all Kurdish glyphs, and processing Kurdish text with the modern computer programming languages is possible. These findings along with some simple testing are posted on a page created for this purpose [3].

Furthermore, we conducted a brief literature review to understand the state of the problem for languages close to the Kurdish language. These include Arabic, Farsi, Urdu and Bangla languages. We find out that substantial amount of research works related to the language processing and spell checking for close languages are done [4, 5, 6, 7, 8, 9, 10, 11]. We concluded the review with understanding that building a spell checker for Kurdish language is a viable research direction and is needed urgently.

▪ Research aim and objectives

The aim of the project is to build a robust, reliable and comprehensive Kurdish spell checker that can be used by everyone in Kurdistan. These include students, government ministers and agencies, business offices, legal institutions, and Universities. The project deliverable would be either a word processer plug-in which would be released to the public as a free of charge download, or a stand-alone application that can used for correcting spelling errors. I prefer the first approach, but the difficulty we will face creating a word processer plug-in and the amount of help we will receive from the Kurdish linguistic partners will determine which road we take.

- **Data collection methods and instruments**

The main resource for collecting data would be World Wide Web. That is, words would be extracted from online web documents, books, and newspapers. Further, we might need to retype valuable books for improving data resource.

- **Mechanisms to assure the quality of the study**

To ensure the quality of the result, each extracted word needs to be checked for accuracy. This would be done using various computer testing strategies and heuristic algorithms. We will also use knowledge and expertise of Kurdish linguistics project partners for validating word spelling accuracy, understating language rules and language concepts.

- **Study period - Timetable for completion of the project**

The project would be delivered in multiple releases where each release would have one or more milestones. We expect within the first six month to collect about 50 K words and implement the first version of the spell checker. I expect the whole project to be completed within 18 months (+/- six months).

- **Participants in the study**

The spell checker project is a cross-discipline one. It requires people form computer science and linguistic department to work together. The project is also a non-commercial one; hence, the project members would be university researcher, students, and faculty members.

- **Resources required for the study**

This is an important research project, yet it does not require any special or excessive resources other than commitment from the participants. The required materials are within the limitation of any genuine university that would provide enough resources to its student for conducting proper research. These include:

1. A research lab equipped with Internet connection and personal computer or laptops for researcher (if researchers have their own laptops no additional computers required)
2. One personal computer and en external hard drive for backing up the code;
3. Peripheral objects such as a laser printer and a scanner
4. Stationery materials such as flush memories, papers, markers, black boards, etc.

- **References**

[1] http://people.scs.carleton.ca/~armyunis/projects/KAPI/KAPI.pdf

[2] http://www.unicode.org/uni2book/ch08.pdf

[3] http://people.scs.carleton.ca/~armyunis/projects/KAPI/KurdishApiProject.html

[4] Khaled Shaalan, Mohammed Attia, Pavel Pecina, Younes Samih and Josef van Genabith*, Arabic Word Generation and Modelling for Spell Checking*, In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)}, 2012.

[5] Khaled Shaalan, Amin Allam, Abdallah Gomah, *Towards automatic spell checking for Arabic*, In Proceedings of the Fourth Conference on Language Engineering, Egyptian Society of Language Engineering (ELSE), pp. 240-247, 2003.

[6]http://sourceforge.net/projects/arabic-spell/

[7] Naushad Uzzaman, Mumit Khan, *A comprehensive Bangla spelling checker,* In the Proceeding of International Conference on Computer Processing on Bangla, 2006.

[8] Tahira Naseem, *A Hybrid Approach for Urdu Spell Checking*, MS Thesis, National University of Computer & Emerging Sciences, India, 2004.

[9] Youssef Rezvan, *Towards Spell checking in FarsiTeX*, In proceeding of the 5th WSEAS international conference on Data networks, communications and computers, Pages 248-250, 2006.

[10] Ehsan, N., *Towards Grammar Checker Development for Persian Language*, In proceeding of Natural Language Processing and Knowledge Engineering (NLP-KE), 2010

[11] Mojgan Seraji, Be´ata Megyesi, Joakim Nivre, *A Basic Language Resource Kit for Persia, In LREC 2012 Proceedings, 2012.*