Bypassing Greeklish!

A. Chalamandaris, P. Tsiakoulis, S. Raptis, G. Giannopoulos and G. Carayannis

Institute for Language and Speech Processing
Artemidos 6 and Epidavrou, GR 151 25, Maroussi, Greece
{achalam, ptsiak, spy, ggia, gcara}@ilsp.gr

Abstract

The present paper describes a new algorithm for addressing a significant issue: "Greeklish" (or "Greenglish"), which arose by the fact that the Greek language is not fully supported by computer programs and operating systems. In the first section of the paper we describe the "Greeklish" phenomenon and the current situation, in reference also with recent research on it. Some examples are provided in order to depict the complexity and the size of the problem at hand. In the second part we describe our approach and the implemented algorithm, together with examples to demonstrate its efficiency and robustness. In the third section of the paper we will present the results obtained after thorough experiments and tests of the system, together with a reference on future plans for further improvement.

Introduction

The wide spread of information and communication media has had a large impact on the communication patterns used in our every-day lives. Email and SMS are only a few of the textual communication channels that have come up to complement (or in specific cases to replace) more typical means such as the voice and the telephone.

However, the urge for their fast adoption of these new technologies has often left important issues insufficiently resolved such as issues related to language support, which presented a barrier for exploiting the full potential of electronic communication means. This is especially true for languages that represent smaller sized groups and markets, such as Greek.

"Greeklish" represent an ad hoc sideway approach for surpassing the problems caused by the lack of a formal and standardized support for the Greek language. Greeklish is not a dialect as often is referred to, but a set of transliteration patterns of Greek using the Latin alphabet. Since this transliteration is mainly based on rules of thumb, a main feature of Greeklish is their inconsistency and variety; some Greek characters may be mapped to different Latin characters, or even be mapped to a combination of Latin characters and vice versa. Nevertheless, Greeklish have managed to provide an effective solution and are thus extensively used today in e-mail communication within Greece and abroad.

Although the Greek standardization body has made an attempt to provide a standard for it (ELOT, 1982), its inconsistence and variety is so wide that it is been said that there are as many different types of "Greeklish" as the Greek-speaking computer users are. This is not accurate, but it gives an idea about the level of it. For example, the Greek word for "address" can be transliterated into more than 20 different and almost equiprobable representations in "Greeklish".

Four dominant approaches can be identified when transliterating from Greek to Greeklish (Androutsopoulos, 1999) , i.e. when mapping from Greek to Latin characters:

a) Based on the ELOT-743 ISO-843 transliteration standard provided by the Greek standardization body.
b) Based on their visual semblance.
c) Based on their phonetic semblance.
d) Based on their location on the keyboard layout.

Nevertheless, it is rather rare for a user to be consistent with only one type of them, even within the context of a single word. This is evident from the fact that common Greek words can be transliterated into more than 20 or 30 alternative representations (with comparable probability) in Greeklish, deriving from a combination of all the above types of Greeklish. Due to this, the reading and understanding of a long text in "Greeklish" can prove to be a demanding task. A recent research has shown that reading and understanding a sentence in Greeklish is in average more time consuming by over 40% and hence requires more effort, than reading and understanding the same sentence in plain Greek (Tseliga, 2003). This fact is generally independent from the Greeklish type used and it depicts the size of the problem if one considers that more time consuming indicates more tiring.

Another research (Androutsopoulos, 1999) has proven that all people above 35 years old consider "Greeklish" as "necessary evil", while almost 50% of all people agree that "reading Greeklish is a hard and tiring task".

Greeklish to Greek Conversion

Background

Since the appearance of Greeklish, there has been a number of attempts to develop tools for the automatic transliteration of Greeklish back to Greek. These have been mainly based on specific sets of rules that would map each Latin character to the respective Greek one, or a set of Latin characters to another set of Greek ones and vice versa. Examples of this approach can be found in (Converters, 2004). This method, although it provides a rough representation of long texts, it is not in any case accurate or orthographically correct. Moreover, it fails to address the problem of accentuation, which is quite important for Greek. In addition although it is easy to switch from one type of Greeklish to another within these applications, it is impossible for them to cope with different Greeklish types within a word or even a sentence unless the user requests it. Consequently, it becomes clear that this approach cannot efficiently address the problem but

just to provide a quick approach of roughly converting from consistently written Greeklish to Greek-like texts.

Our approach

This paper proposes a different approach that is based on statistical models and lexicons acquired from large corpora, in order to effectively address the problems of inconsistency and variety of Greeklish forms. Its design and function allows it to cope effectively with virtually any type of Greeklish, based on statistical data.

Its operation consists on the following consecutive steps:

a) Transcription of every Greeklish word into all possible phonetic representations, by taking into account all possible different types of Greeklish along with all their likely combinations.
b) Pruning of all alternative solution by using an acoustic model specifically designed for the Greek language.
c) Search for the most probable solutions within a specially designed lexicon derived from large Greek corpora.
d) Decision for the best solution according to derived probabilities and context dependent rules.

In addition to all the aforementioned steps, there is also a language identification algorithm embedded in between these steps, in order to avoid unsuccessful attempts of transliterating a non-Greek word into Greek. This is essential in an application of this scope, since it has been noted that when it comes to email written in Greeklish, the authors tend to use intercalary English words or even phrases more often (Tseliga, 2003). This is performed with the use of probabilistic rules and statistical models specially designed for the Greek language, allowing it to discriminate efficiently Greek from non-Greek words. In the following paragraphs we are going to present in more detail the function and the results of this module.

The proposed system

The flowchart of our system is depicted below.

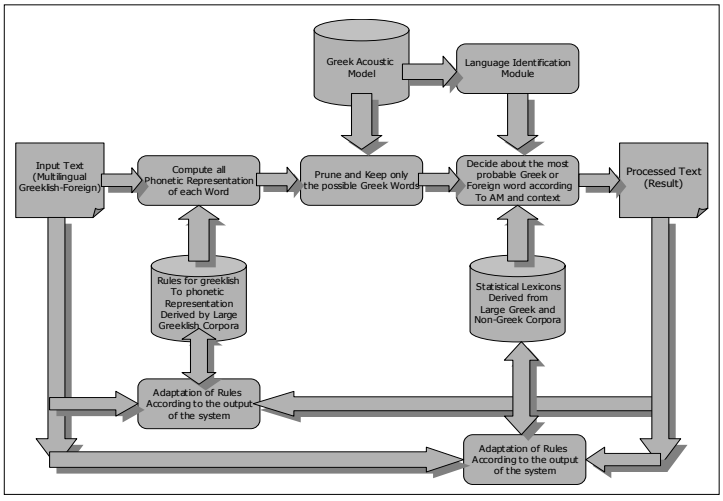

Figure 1: Flowchart of the proposed system.

As it is shown in the flowchart, we make use of an intermediate phonetic representation of every Greeklish word. This is essential for the implementation of the acoustic model for language identification, and it proved to be very efficient for online pruning of the possible outcomes a Greeklish word might have. By doing so we manage to maintain low computational complexity and high performance. The same idea applies also in the case where a Greeklish word is not listed in the dictionary; nevertheless it will reproduce the most probable phonetic representation of it, derived from the Greek acoustic model. The feature of providing the most probable phonetic version of an unknown Greeklish word is also important in cases where the output of the system is fed to a Text to Speech system synthesizer. In that case even if the word is not listed in the dictionary, it will be properly uttered by the TtS system (lacking, of course, the proper accentuation).

The probabilistic rules for mapping a Greeklish word to its corresponding possible representations arose from the assigning rules and the corresponding probabilities when transliterating from Greek to Greeklish and vice versa. The latter were computed semi-automatically from a large Greeklish corpus gathered by both mailing lists and private mails, written by more than 60 different people. The outcome of this process was very useful in order to investigate different transliteration patterns between Greek and Greeklish and relative frequencies. In the following figure one can see the results for single letter theta (‘Θ’) and the 6 different alternative ways for transliterating in Greeklish, as it was observed in our corpus.

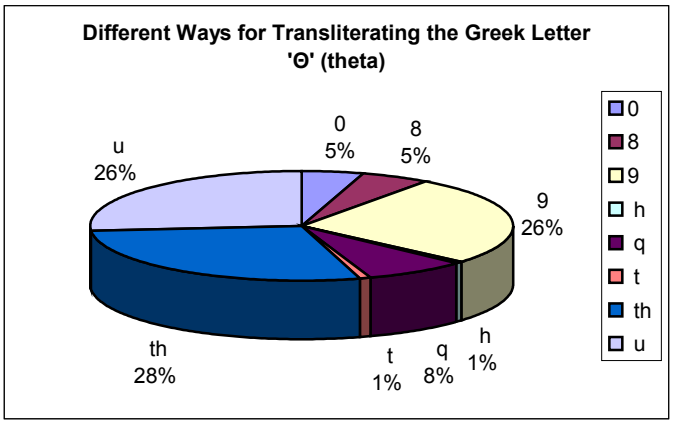

Figure 2 alternative transliterations of ‘θ’ (total number of instances 5,775). Although the transliteration patterns of ‘h’ and ‘t’ are unlikely and probably caused to spelling errors, they turned to be two pattern we needed to consider in our system’s design in order to make it more robust.

We could easily understand the level of variability in transliterating a Greek word into Greeklish and its order of complexity if we considered the case where we would like to reproduce for example all possible alternative Greeklish transliterations of a single Greek word. The aforementioned example of the Greek word for ‘address’, which we said that it could easily be found in more than 20 equiprobable

alternative Greeklish versions, if we consider all possible transliteration patterns we identified, it can reproduce more than 2000 alternative versions in Greeklish! This example gives a very good estimate of the complexity of the problem for transliterating a Greek word into all possible alternatives in Greeklish. What is also important to note here, is that the reverse process, of converting a Greeklish word into its corresponding Greek ones is of similar complexity since for example the Latin letter '*i'* can be transliterated into at least 5 different combinations of Greek letters, or the letters 'th' can be transliterated into at least 4 different patterns. One can easily calculate that a Greeklish word containing three 'i's could easily reproduce more than 5^3 alternative orthographies for the same word, only a very small number of which would be allowed though for the same word. In our system the pruning of allowed orthographies is performed online with the use of the Greek acoustic model, dropping the computational load by over 75%.

The values of the probabilistic rules for each level of our system were determined in two stages. Initially they were defined manually, according to our estimations and experience with Greeklish, and at a second stage during the automatic extraction of the above probabilities from large corpora. We observed significant deviations between the real-life values and the ones we initially "guessed"; nevertheless the first step was necessary for building automatically the aligned corpora between original Greeklish and their Greek transliteration. After the acquisition and semi-automatic correction of these corpora, we extracted the transliteration rules from Greek to Greeklish and vice versa, by using dynamic time warping (DTW) to every couple of words.

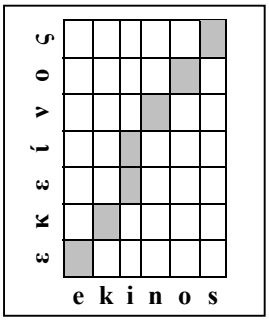

Figure 3: Example of automatic extraction of transliteration patterns between Greek and Greeklish words from aligned corpora.

By doing so we managed to extract all possible Greek to Greeklish transliteration patterns and vice versa, together with their probability weight for each one of them.

A very important finding during our research on long real-life corpora was that the derived patterns were significantly more that what we initially had estimated, not only because of people's preference to variability, but also to spelling errors that the author originally would make if he would write in plain Greek. Consequently, instead of having only three transliterating patterns for Greek letter 'ypsilon' 'υ' i.e. 'y', 'u' and 'i', as we initially expected, we found out that the same letter could be also transliterated into 'ei' 'h' and 'oi' due to spelling errors. Although the frequency of this phenomenon is not generally high, in cases of web email forums or where the author is writing an email in haste it tends to rise. In general, we found out that there are more than 100 different transliteration patterns (in contrast to the 64 patterns we initially guessed), on which we based the initial design of our system.

Language Identification Problem

The language identification problem is been addressed several years ago effectively. There are several alternative solutions to this problem, most of them though based on statistical models and data derived from large corpora. The main concept in such a system is to compute the mixed probability of a sequence of letters according to HMM models derived for every language. In the case where for example we want to identify French and English words within a text, we compute for every single word, or a set of consecutive ones, the mixed probabilities of their respective HMM model, and then decide according to the maximum value (Dunning, 1994). This method generally performs quite well and robustly when it copes with languages that the system has been trained on.

In our case the language identification problem was addressed differently since there is not a firm and consistent way of writing Greeklish, and our aim was to develop a module able to discriminate any Greeklish text from any other language. In order to surpass this problem of inconsistency in writing Greeklish, we made use of an alternative representation of every Greeklish word, namely a phonetic one. It was rather easy to build a statistical phonetic model for Greek language (i.e. acoustic model), which turned to be a very good and robust tool for language identification in our case. After having converted every word into its most likely phonetic representations according to Greeklish transliteration patterns, we used our acoustic model (in our case based on tri-grams) to decide whether this word is more likely to be in Greeklish or in any other language. By doing so our system can effectively cope with multilingual texts with high accuracy and robustness. After the first level of language identification on word level, another post-processing stage follows that takes into account the contextual information of the sentence. At the latter stage, in order to avoid erroneously transliterated non-Greek words the system filters out isolated transliterated words in pure non-Greek environments.

The performance of this module was tested with large multilingual corpora, where the initial Greek text was transliterated automatically according to 4 different sets of rules, as mentioned earlier. The output of our module was compared automatically to the initial text in order to compute the error rates for language identification. In the following table one can observe the corresponding success rates of our algorithm.

Generally, the language identification module performs quite well; nevertheless it is possible to improve even more if we use additional language lexicons and HMM models for every language individually.

GREEKLISH TYPE	MIXED GREEK/ENGLISH (64%-36%)		MULTILINGUAL GREEK/NON-GREEK (59%-41%)	
	Greek Words	English Words	Greek Words	Non-Greek Words
Type I (ELOT 743-ISO 843)	98.63%	98.12%	97.44%	91.16%
Type II (Phonetic)	98.31%	98.41%	98.14%	92.11%
Type III (Orthographic)	97.63%	98.41%	97.21%	91.16%
Type IV (Keyboard layout)	96.14%	98.12%	96.44%	91.16%
Type V (Mixed)	97.23%	98.12%	98.01%	92.11%

Table 1: Success rate of language identification module.

GREEKLISH TYPE	SUCCESS RATE (TOTAL WORDS)	SUCCESS RATE (UNIQUE WORDS)	SUCCESSFUL DETECTION OF GREEK WORDS	SUCCESSFUL DETECTION OF NON-GREEK WORDS
Type I (ELOT 743-ISO 843)	96.75%	98.43%	98.17%	93.45%
Type II (Phonetic)	97.32%	97.56%	98.78%	93.32%
Type III (Orthographic)	98.15%	98.65%	99.01%	94.12%
Type IV (Keyboard layout)	97.10%	97.14%	98.14%	94.12%
Type V (Mixed)	96.92%	97.43%	98.31%	94.12%

Table 2: Success rates results of our system.

Test Results

The system's performance was investigated in two different stages. During the first stage, the test data was "original" Greeklish corpora, taken by public mailing lists, private emails and web pages in Greeklish. We gathered emails and messages written by more than 60 different persons, all of them written in mixed Greeklish and English, resulting to a reference corpus of 378,108 words, the 58,256 of which were unique. The output of our system was checked manually in order to investigate the success rate, which reached 96.34% for Greeklish words, and 97.54% for successful detection of non-Greek words. (Table 1)

During the second stage of system's evaluation, we used large multilingual corpora (containing 76% Greek words) of total 1,424,456 words, the 542,453 of which were unique. The content was varying from private and public emails, to web pages, newspapers, manuals, general documents, reports and educational material for Greek high-school. The corpus was initially transliterated into Greeklish according to five different ways and then processed by our system in order to crosscheck the original text and the resultant one, word by word automatically. The results are shown below in table 2.

The success level achieved is quite high. The few identification errors could be attributed mainly to the limited ability of Greeklish to capture the Greek language's rich morphology. We also observed that our system performed slightly better in the case of the orthographic type of Greeklish. This is entirely due to the fact that this is the only Greeklish type that can preserve more information about the right spelling morphology of every word. A simple way for overcoming this obstacle is the use of advanced proofing tools that utilize syntactical and grammatical information of a sentence instead of only an orthographic dictionary.

Acknowledgments

We would like to thank Dr. Athanassios Protopapas for his helpful guidance during the design and development of the aforementioned system, Mr J. Papageorgakopoulos and Mr. A. Baxevanis, as well as whoever contributed in the development and testing of the system.

References

-Androutsopoulos, J. (1999). Latin-Greek orthography in electronic mails: use and stances]. Paper presented at the 20th Annual Meeting of the Linguistics Department, 23-25 April 1999, Aristotle University of Thessaloniki.

-Androutsopoulos, J. (2000).From dieuthinsi to diey8ynsh. Orthographic variation in Latin-alphabeted Greek]. 4th International Conference on Greek Linguistics, September 1999, University of Nicosia,

-Dunning, T. (1994). Statistical Identification of Language. Technical report CRLMCCS-94-273. Computing Research Lab, NewMexico State University.

-Koutsogiannis, D. & Mitsikopoulou, B. (2003). Greeklish and Greekness: Trends and Discourses of 'Glocalness'. Proposal submitted the forthcoming special issue of Journal of Computer-Mediated Communication on "The Multilingual Internet"

-Tseliga, T. (2003). A corpus-based study of discourse features in Roman-alphabeted Greek (i.e. Greeklish) emails. 1st International Conference on Internet and Language, Castellon, Spain, 18-20 September.

-Tseliga, T. and Marinis, T. (2003). On-line processing of Roman-alphabeted Greek: the influence of morphology in the spelling preferences of Greeklish. 6th International Conference in Greek Linguistics, Rethymno, Crete, 18-21 September, 2003.

-ELOT (1982). Greek Organisation of Standardization

-Converters, (2004).:http://home.asda.gr/active/GrLish2.asp
http://www.translatum.gr/converter/greeklishconverter.htm