All Greek to me! An automatic Greeklish to Greek transliteration system

Aimilios Chalamandaris, Athanassios Protopapas, Pirros Tsiakoulis, Spyros Raptis

Institute for Language and Speech Processing
Epidavrou & Artemidos 6, 15125 Maroussi, Greece
{achalam, protopap, ptsiak, spy}@ilsp.gr

Abstract

This paper presents research on “Greeklish,” that is, a transliteration of Greek using the Latin alphabet, which is used frequently in Greek e-mail communication. Greeklish is not standardized and there are a number of competing conventions co-existing in communication, based on personal preferences regarding similarities between Greek and Latin letters in shape, sound, or keyboard position. Our research has led to the development of “All Greek to me!,” the first automatic transliteration system that can cope with any type of Greeklish. In this paper we first present previous research on Greeklish, describing other approaches that have attempted to deal with the same problems. We then provide a brief description of our approach, illustrating the functional flowchart of our system and the main ideas that underlie it. We present measures of system performance, based on about a year’s worth of usage as a public web service, and preliminary research, based on the same corpus, on the use of Greeklish and the trends in preferred Latin-Greek letter mapping. We evaluate the consistency of different transliteration patterns among users as well as the within-user consistency based on coherent principles. Finally we outline planned future research to further understand the use of Greeklish and improve “*All Greek to me!*” to function reliably embedded in integrated communication platforms bridging e-mail to mobile telephony and ubiquitous connectivity.

1. Introduction and background

The word “Greeklish” stands for a combination of the Greek and the English language (Greek-lish) and it refers to transliteration of Greek using the Latin alphabet. This Romanization is used frequently in e-mail communication among Greek-speaking computer users, and its main characteristic is the lack of a standardized table of transliteration mappings. More specifically, Greeklish is a significantly inconsistent manner of transliterating Greek with the Latin alphabet, based on alternative co-existing conventions, which mainly depend on personal preferences regarding similarities between Greek and Latin letters’ shape, sound or even keyboard layout. Before full compatibility of operational systems with the Greek alphabet, Greeklish was the main means for communicating amongst users. Nowadays, even though most operational systems and programs support Greek character set, Greeklish still remains one of the main tools for safe communication via e-mail.

Several studies of Greeklish [3,4] have shown that nearly all Greek-spoken computer users have used Greeklish at least once as a means of communication via e-mail; at the same time, more than 50% of the users over 35 years old consider Greeklish to be a necessary evil in everyday computer use. Another important aspect of this Romanization is their difficulty: It has been found that reading a text written in Greeklish demands at least 40% more time and effort than reading the same text in plain Greek, even for experienced users of Greeklish [11]. Greeklish has been an apple of discord in the past [1] and its impact in the actual quality of the content they deliver is also a subject of research by linguists [11,13].

1.1. Types of Greeklish

One of the earliest studies [1] of the Greeklish phenomenon classified the basis for transliteration into three distinct categories:

1. Based on sound resemblance aiming to represent phonetically the respective Greek text, i.e. the Greek letter /θ/ yields /th/ and the diphthong /αι/ yields /e/
2. Based on similarities between Greek and Latin letter shapes, i.e. /8/ for the letter /θ/ and /w/ for the letter /ω/
3. Based on similarities in the keyboard layout, i.e. /u/ for the letter /θ/ and /c/ for the letter /ψ/.

Several other researchers [10,11,12] have agreed with this classification, nevertheless the validity of this hypothesis has not so far been tested empirically based on usage data. In this paper we present a first approach to this question in section 3.

1.2. Approaches to automatic transliteration

Since the appearance of Greeklish, several attempts have been presented in the literature, either as ad hoc approaches for automatic transliteration [14] or as more complete applications with advanced features such as email client services etc. Most of these applications are distributed freely and are based on a specific, fixed set of transliteration rules, simply replacing every Latin letter into a corresponding Greek letter. Few of these applications make use of regular expressions techniques in order to better cope with different context-dependent patterns [9]. One application particularly worth mentioning is aspell [5], an open source spell checker for Linux environment and OpenOffice suite. Aspell first maps all Latin characters to Greek ones via a specific mapping set and then applies its conventional method of spell checking and correction.

Our approach, apart from the incorporation of dictionaries, differs from all aforementioned ones on three important aspects. First, we use an intermediate stage of phonetic representation of all Greeklish words, which provides faster and more robust results than passing directly to Greek characters. Second, we use probabilistic models for the decision of the optimal mapping from Latin to Greek characters, as well as for the decision of the most probable word. And third, our system can handle very efficiently mixed texts with Greeklish and non-Greeklish words, using a language identification algorithm.

2. Our Approach

In this section we present the ideas that underlie our approach as implemented in *All Greek to me!* developed at ILSP [6]. *All Greek to me!* is the first automatic transliteration system that can cope with virtually any type of Greeklish and provide orthographically correct Greek text. In the following figure one can see the general flowchart of the system.

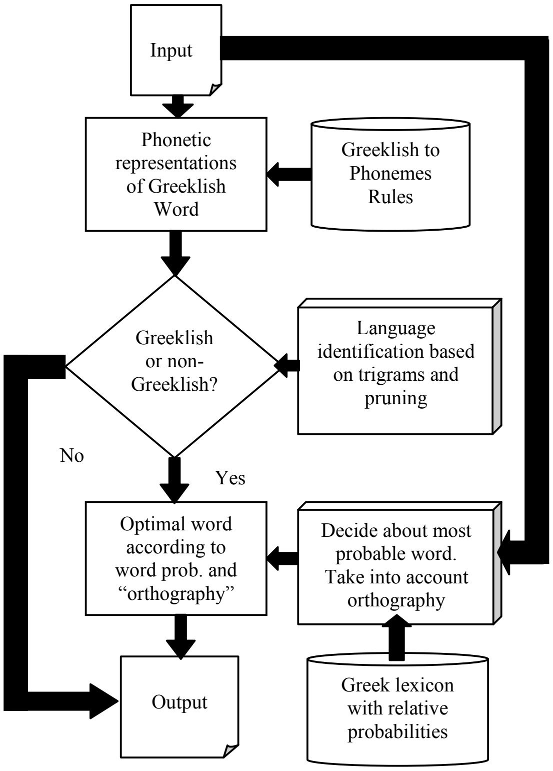
Input
Phonetic representations of Greeklish Word
Greeklish to Phonemes Rules
Greeklish or non-Greeklish?
Language identification based on trigrams and pruning
No
Yes
Optimal word according to word prob. and "orthography"
Decide about most probable word. Take into account orthography
Output
Greek lexicon with relative probabilities

Figure 1: Flowchart of *All Greek to me!* system.

The first step of its operation is to transcribe from Greeklish into all possible phonetic representations using a set of manually defined rules (enriched after initial testing [6]). This intermediate stage helps prune alternatives employing a phonotactic model for Greek, which at the same time performs language identification [8]. By using trigram probabilities, every phonetic word produces a score according to its constituent phonetic sequence. The instances that produce a score below a specific threshold are considered to be non-Greeklish words (that is, foreign words) and therefore are left intact in Latin characters. Instances scoring above the threshold are passed on to a next level and projected onto a large lexicon that contains relative probabilities of appearance in general purpose Greek texts, such as newspapers or news broadcasts. The projection is based on the phonetic representation; hence for every word in the lexicon we have also stored the corresponding phonetic sequence. The detailed structure and function of *All Greek to me!* has been presented in [6].

3. Usage data

In this section we present analyses of system usage, provided as a free web-based service at ILSP's official web site [2].

3.1. User distribution

Analysis of the performance is based on data acquired in the ten month period from January 2005 through October 2005 via the online demo version of our application [2], which limits each conversion request to 255 characters. This sample is very important because it constitutes a large corpus of real-life unbiased Greeklish, and as such it allows us to derive objective conclusions about our system.

The total size of the corpus is 2,095,037 words, including 145,601 unique words. The total number of entries (conversion requests) was 171,698, received from 18,868 unique IP addresses. The latter number does not represent the unique users because an estimated 41% of the users do not have static IP addresses and therefore may be represented in the corpus with alternative IP identities. Users originated in 83 different countries, of which the most frequent are listed in the Table 1.

No	COUNTRY	Request %	Unique IPs %
1	GREECE	47,28%	53,75%
2	GERMANY	15,85%	16,89%
3	UNITED_KINGDOM	11,90%	4,85%
4	UNITED_STATES	7,18%	5,36%
5	AUSTRALIA	3,61%	2,03%
6	FRANCE	1,98%	2,01%
7	NETHERLANDS	1,86%	0,43%
8	ITALY	1,62%	1,22%
9	CYPRUS	1,39%	1,68%
10	BELGIUM	1,09%	1,15%
	REST	6,24%	5,31%

Table 1: Countries of origin of the conversion requests making up the corpus.

With the use of cookies, we estimate that in average 62.3% of the users are frequent users. Until the day this paper was written, the use of our web service was doubled within a five-month period, exceeding 38.000 different users, and having converted more than 5,200,000 words.

3.2. Transliteration pattern preferences

A series of hierarchical log-linear models with and without a latent class were constructed in order to test the hypothesis that users of Greeklish tend to prefer one of the three main modes of transliteration (visual, phonetic, keyboard layout). For this test we used transcriptions yielding /η/, /υ/, /ω/, /θ/, and /ου/. These five graphemes are the only ones easily admitting all three modes of transcription and producing distinct outcomes (visual: /n/ /u/ /w/ /8/ /ou/; phonetic: /i/, /i/, /o/, /th/, /u/; keyboard: /h/ /y/ /v/ /u/ /oy/, respectively). Under the assumption that a unique IP address represents mainly a single user, we retained and grouped texts submitted for transcription by the 50 IPs with the highest total counts of the critical characters. There were 170,478 total instances of these letters, ranging between 1,500 and 12,050 per IP (mean 2,410, standard deviation 2,064, median 2,981 instances per IP).

The raw counts were entered in a log-linear model description for processing, hierarchically grouped under three manifest variables: L(etter), transcription G(roup), and P(erson). There were 5 levels of L (one per grapheme), 3 levels of G (visual, phonetic, keyboard), and 50 levels of P (the 50 IPs). All analyses were carried out using the lEM software program developed by J.K. Vermunt [15] at the University of Tilburg. The saturated model, naturally, fit the data completely (χ^2=0.0000, 649 parameters, dissimilarity index=0.0000, BIC log-likelihood=1910374.5), with all variable main effects and interactions statistically significant. Removal of all interactions resulted in an unacceptable model (χ^2= 251055.8, *df* = 694, *p*< 0.00005, 55 parameters, dissimilarity index = 0.5295, BIC log-likelihood = 2164382.0), indicating that the independent effects of the 3 variables were insufficient to determine the observed distribution. Therefore, even if specific IP-level trends could be discerned, they could not be uniform over the 5 letters.

By restricting the model to treat P and L as independent, model fit was reduced substantially compared to the saturated model but not very severely (χ^2= 4298.6, *df*=196, p<0.00005, 553 parameters, dissimilarity index=0.0604, BIC log-likelihood= 1912322.7). The hypothesis of interest was then tested by constructing a latent class model in which P was allowed to affect the transcription counts only via X, a 3-level construct, which was entered in the model as independent from L but allowed to affect G in interaction with it. That is, a given mode preference was allowed to be modulated by the particular letter as to how much it was expected to affect transcription, in order to account for a letter's better or worse perceived fit to each mode of transcription. This latent class model did not fit the data very well (χ^2 = 45206.1, *df* = 566, *p*<0.00005, 134 parameters, dissimilarity index = 0.1554, BIC log-likelihood = 1946795.4) and was significantly worse than the model in which P was allowed to affect G directly, without the latent variable (*p* < 0.00005 in L2 comparison). However, the latent class model was significantly better (*p* < 0.00005) than a model in which P was ignored and only L, G, and their interaction was considered (χ^2 = 234927.7, *df* =7 35, p < 0.00005, 14 parameters, dissimilarity index = 0.3732, BIC log-likelihood = 2070657.0).

In the latent-class model, the effect of X (the latent variable) was significant (Wald χ^2=489631.1, *df*=2, *p*<0.0005), as was its interaction with G (χ^2=116240.27, df=4, *p*<0.0005) and the triple interaction with LG (χ^2=1689050.2, *df*=16, *p*<0.0005). The probabilities P(X|P) indicated that most persons were probabilistically classified into the three levels of X with a clear domination of one level. The latent class output for the 3 levels of G (the transcription modes) indicated that level X=1 was associated primarily with phonetic (0.52) and somewhat less with visual (0.34) transcription, X=2 with keyboard transcription (0.73), and X=3 with a less differentiated performance over keyboard (0.23), phonetic (0.47), and visual (0.48) transcription.

Therefore, our findings indicate that there are IP-level trends in transcription mode preferences. These trends do not exhaust the IP-level variance in transcription frequencies; however, it must be taken into account that IPs are, at best, an imperfect correlate of unique person identity, because many ISP subscribers and professional intranet users will access the *All Greek To Me!* site with common IPs corresponding to their domain gateway. The fact that IP-level trends are significant anyway suggests that person-level trends are likely to be stronger, thereby validating to some extent the hypothesis that stable preferences of individual users exist and can be grouped in 3 categories.

Nevertheless, the IP-level trends discerned do not match up perfectly with the 3 transcription modes but with certain mixtures of them. Specifically one preference mode was found to produce primarily keyboard-based transcriptions, while another mixes primarily phonetic with some visual transcriptions. The statistically significant interaction with letter suggests that different letters may be more amenable to one or another of the transliteration modes. Such patters remain to be investigated in future data collection in which single-person usage should be ascertained.

4. Transliteration Performance

Evaluation of system performance was systematically carried out as follows. First we classified all unique words into four categories affecting system behavior, namely known versus unknown (based on dictionary hit) crossed with Greeklish versus non-Greeklish (foreign; based on phonetic modeling).. A portion of all transcribed words in each of the 4 categories was manually checked for accuracy of conversion, and the resulting success rates were projected to the entire corpus according to the corresponding word frequencies. A set of subclasses were defined for each category, to help define precise criteria demarcating correct from incorrect system performance.

Distribution among the categories is shown in Table 2.

Word Category	**Unique Words**	**% of Unique Words**	**Total Words**	**% of Total Words**
Known Greeklish (Greek word in lexicon)	109.900	75,48%	1.925.797	91,92%
Non-Greeklish (Foreign)	20.330	13,96%	144.270	6,89%
Unknown Greeklish (not in lexicon)	1.470	10,11%	22.112	1,06%
Other (mostly misspelled)	661	0,45%	2.858	0,13%
Total	**145.601**	**100%**	**2.095.037**	**100%**

Table 2: Classification of the corpus words

The last category includes mainly user mistakes, which cannot be classified into any of the other categories.

For the first category, only 0.53% of the words were found to have been transliterated incorrectly, mainly due to incorrect records in the lexicon. 86.21% of the words were transliterated correctly unambiguously; the remaining 13.26% admitted more than one correct transliteration, depending on the context (and hence on the appropriate grammatical form). The latter portion cannot be considered incorrect because the output of the system, at the single-word level is orthographically acceptable. In such cases the system is designed to take into account the Greeklish orthography, when relevant, as well as the relative word probabilities.

In the second category of converted words, the error rate was 0.68%, while in the third category it was 9.11%. By projecting these error rates to the entire lexicon, the total overall error rate was estimated at 0.63%; in other words the bottom-line system performance was 99.37% correct. This estimate does not take into account words converted correctly by the system that admit alternative transliterations depending on the context in which they appear. These words can only be handled by incorporating language modeling, an addition in our future plans for system improvement.

5. Discussion

To summarize, in this paper we presented our research on Greeklish and the system we developed for automatic transliteration from Greeklish to Greek. We have presented the flowchart of the system and the corpus acquired via a free web service. Performance assessment of *All Greek to me!,* was systematically carried out on this corpus, showed extremely high system performance. Nevertheless there is still room for improvement. Integration of a language model will allow the system to cope with words that have several alternative Greek transcriptions. Finally, regarding Greeklish type preferences among the *All Greek to me!* users, our preliminary results indicate that users tend to use a mix and match approach to Greeklish type in their messages, instead of only one of the three different types mentioned in section 1. This result was not surprising since the initial discrimination of these three types was mainly attempting to rationalize the variation in Greeklish preferences and not to provide a distinct classification. The identification of potentially distinct types of Greeklish and preferences among users, depending on other factors, is an object of research we aim to further investigate.

6. Aknowledgmenets

We would like to thank Yannis Papageorgakopoulos and Alexandros Baxevanis for their valuable help in the design of the web service, all ILSP personnel that helped in testing and all the people who have contributed to this project.

7. References

[1] Academy of Athens (2001). Greeklish: An enemy of the nation. *Press release January 2001*

[2] All Greek to me! Online Demo Web Site: http://www.ilsp.gr/greeklish/greeklishdemo.asp

[3] 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.

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

[5] Aspell spell checker for Greek: http://aspel.source.gr

[6] Chalamandaris A., Tsiakoulis P., Raptis S. Giannopoulos G and Carayannis G. (2004). Bypassing Greeklish! *LREC2004 proceedings* vol. 1 pp 285-288, Lisbon 26-28 May 2004.

[7] ELOT (1982). Greek Organisation of Standardization

[8] Dunning, T. (1994). Statistical Identification of Language. *Technical report CRLMCCS-94-273*. Computing Research Lab, New Mexico State University.

[9] Karakos A. (2003). Greeklish: An experimental Interface for Automatic Transliteration. *Journal of the American Society for Information Science and Technology*. Vol. 54(11). pp1069-1074

[10] Koutsogiannis, D. & Mitsikopoulou, B. (2003). Greeklish and Greekness: Trends and Discourses of 'Glocalness'. *Journal of Computer-Mediated Communication on "The Multilingual Internet"*, Vol. 9, No. 1

[11] 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.

[12] 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.

[13] Tseliga, T. (2003). Using a 'multi-strategy research' to analyse transliteration patterns in Greeklish. 16th International Symposium On Theoretical & Applied Linguistics, Thessalonica, Greece, April 11-13, 2003

[14] Online Converters
http://home.asda.gr/active/GrLish2.asp
http://www.translatum.gr/converter/greeklishconverter.htm

[15] Vermunt, J.K. (1997). LEM: A General Program for the Analysis of Categorical Data. Department of Methodology and Statistics, Tilburg University"