INTERSPEECH 2009 BRIGHTON

A Large Greek-English Dictionary with Incorporated Speech and Language Processing Tools

Dimitrios P. Lyras [1], *George Kokkinakis* [1], *Alexandros Lazaridis* [1], *Kyriakos Sgarbas* [1], *Nikos Fakotakis* [1]

[1] Speech & Language Processing Group, Wire Communications Laboratory, Department of Electrical and Computer Engineering, University of Patras, Patras, Greece, GR-26500

dimlyras@upatras.gr, gkokkin@wcl.ee.upatras.gr, alaza@upatras.gr, sgarbas@upatras.gr, fakotaki@upatras.gr

10.21437/Interspeech.2009-548

Abstract

A large Greek-English Dictionary with 81,515 entries, 192,592 translations into English and 50,106 usage examples with their translation has been developed in combined printed and electronic (DVD) form. The electronic dictionary features unique facilities for searching the entire or any part of the Greek and English section, and has incorporated a series of speech and language processing tools which may efficiently assist learners of Greek and English. This paper presents the human-machine interface of the dictionary and the most important tools, i.e. the TTS-synthesizers for Greek and English, the lemmatizers for Greek and English, the Grapheme-to-Phoneme converter for Greek and the syllabification system for Greek.

Index Terms: Dictionary, TTS Synthesizer, Grapheme-to-Phoneme Converter, Syllabification System, Lemmatizer

1. Introduction

The advent of computers and multimedia technologies have changed the traditional dictionaries into their new CD- and DVD-form providing them with unique features for presenting, retrieving and handling the required information. Then, the huge expansion of the World Wide Web allowed the access of dictionaries installed everywhere in the network [1].

Nowadays, the advancement of speech and language technology gives the possibility to further improve the capabilities of electronic dictionaries by incorporating speech and language processing tools such as speech synthesizers, lemmatizers, speech recognizers etc. In this case, elaborate dictionaries result, which efficiently assist language learners in several aspects [2].

In particular, bilingual dictionaries are the most important aid in foreign language learning. Thus, their improvement in content, its access and its presentation to the user has always been a major endeavor by lexicographers and publishers. Currently, the above mentioned technological advancements have made the development of electronic bilingual dictionaries with exceptional features possible.

This paper presents the electronic version of a large Greek-English dictionary incorporating a series of novel speech and language processing tools. The dictionary, named KORAIS in honor of the great Greek man of letters Adamantios Korais (1748-1833), has been based on resources of the Speech and Language Processing Group of the Wire Communications Laboratory (WCL) and was developed in two phases [3]: Initially, by a consortium of Universities and companies in the framework of the EPET II R&D Programme of the Greek General Secretariat for Research and Development - GSRT (1999-2001), and then at WCL (2001-2008) by members of the Speech and Language Processing Group and specialist external collaborators.

The main features of both the printed and electronic versions of the dictionary are:

- the large number of entries (81,515), translations into English (192,592), and usage examples with translation (50,106),
- the presentation of each entry's pronunciation with the computer phonetic alphabet (CPA) symbols in the printed version and prerecorded speech in the electronic one,
- the broad grammatical analysis with inclusion of declension forms,
- the detailed clarification of each entry's meaning through the use of synonyms, collocates, or phrasal description, and, in particular, illustrative examples of usage in the form of complete sentences,
- the complete translation into English of all meanings and examples of usage.

Furthermore, the electronic dictionary in DVD-form offers:

- a user-friendly human-machine interface with easy selection and on-screen presentation of the desired information,
- a flexible search of each coded information item, i.e. entry, pronunciation, example of usage, or a combination of codes,
- an "intelligent search", using specific conditions,
- the possibility of omitting stress signs or the distinction between capital and lower case letters,
- the display of probable results according to criteria set by the user.

In addition to the above, a series of language technology tools have been incorporated into the electronic dictionary which may efficiently assist learners of the Greek and English language:

- Text-to-Speech (TTS) synthesizers for Greek and English,
- lemmatizers for Greek and English,
- a grapheme-to-phoneme converter for Greek,

- a syllabification system for Greek,
- other tools.

In the following sections 2 to 6 the human-machine interface and the tools for Greek and English TTS synthesis, lemmatization for Greek and English, Grapheme-to-Phoneme conversion for Greek, and syllabification for Greek are presented. Section 7 gives a brief conclusion

2. Human-Machine Interface

The KORAIS dictionary features a user friendly human-machine interface which provides easy access to the stored information and a functional presentation to the user.

Concerning the ease of access to all information, great efforts have been made to supply the user with intelligent and adaptable search options. To this end, the dictionary's framework containing the codified information of the specific fields, i.e. headword, pronunciation, part of speech, translation, examples, etc, was transformed into an SQL database. Via this transformation, the possibility of performing a quicker and more thorough search in the contents was made possible. By using different types of SQL queries, the user is enabled to search for entries in the dictionary not only the usual way (i.e. entries that ***start with*** the typed keyword) but also with the options of ***ends with*** (i.e. retrieve entries of the dictionary that end with the typed keyword, e.g. in compound words), ***equals*** (i.e. retrieve only entries that match exactly the typed keyword), and ***contains*** (i.e. retrieve entries where the typed keyword is a part of the entry). Furthermore, since through the transformation of the framework to an SQL database the information included in each entry of the dictionary was split into its different fields (translation, pronunciation, examples, synonyms etc), the search options were further enhanced. The user is able to retrieve all the examples, synonyms, pronunciations and so forth in the dictionary that fulfill his/hers searching criteria.

As far as the working environment of KORAIS is concerned, it consists of a Windows Graphical User Interface (GUI) with four panels: Search Panel, Word List Panel, Result Panel, and Options Panel. It also includes a menu bar and a tool bar. The Search Panel accepts the keyword to be searched as well as some definitions concerning the type of search that has to be performed. The Word List Panel allows a quick browse and transfer to the search panel of any entry included in the dictionary. The Result Panel displays the results of the search procedure. Finally, the Options Panel serves for defining some options regarding the search type that will be performed and the formatting/presentation of the returned results before they appear on the Result Panel.

In order to further facilitate the familiarization of the user with the interface, the KORAIS – GUI is adapted each time to the selected theme of the user's operating system. In addition, options concerning modifications of the appearance of the user interface, i.e. background color, font size/style/color of the presented search results etc, are also provided to the user.

3. Text-To-Speech Synthesizers

The Festival speech synthesis framework [4] was used for the implementation of the Text-to-Speech (TTS) systems in KORAIS dictionary. Festival is a multilingual speech synthesis framework incorporating various methods of speech synthesis, e.g. concatenate (diphone, unit selection) synthesis, statistical parametric speech synthesis based on hidden Markov models (HMM) etc. In our case, a diphone based TTS synthesis was used. This choice gives the advantage of implementing a reliable and computationally not very demanding TTS system. In addition, modifications on the diphone voices can be easily effected by the user.

The diphone synthesis is based on the concatenation of diphone units. A diphone is the speech unit extending from the middle of one phone to the middle of the next one. Consequently when concatenating diphones instead of phones the joining takes place in a more stable part of the signal, the middle, rather than on the start or the end of the phone where the co-articulation phenomenon is taking place during the production of speech [5]. In diphone synthesis a small database containing all the diphones occurring in a language is needed. Only one instance of each diphone is required. In Festival the diphone based synthesis is utilized using the Residual Exited LPC synthesis technique [6].

Two diphone voices have been implemented in the TTS systems in KORAIS, one American English male diphone voice (KAL) [4] using the CMU Lexicon [7] and one Greek female diphone voice (ZETA), accordingly. Both voices use LPC of 16^{th} order and synthesize speech at 16 kHz. The prosodic (duration and intonation) models for the American English voice have been trained on the Boston University FM Radio corpus [8] using Classification and Regression Trees (CART) [9]. For the Greek voice the prosodically rich WCL-1 database [10] was used for training the duration [11] and intonation models employing also CART. In addition prosodic phrasing is provided by using part of speech and local distribution of breaks in both voices. The phone set of the American English voice consists of 40 phones plus silence and approximately 1600 diphones. For the Greek voice the phone set consists of 39 phones plus silence and 700 diphones. Listening tests with both voices provided good intelligibility and acceptable naturalness. Work is continued though, to further improve the quality of the Greek TTS system.

The KORAIS TTS-systems are accessible through a Graphical User Interface (GUI). Through this interface the user is able, apart from choosing the voice (KAL or ZETA) depending on the language she/he wants to synthesize, to modify the rhythm of speech and the volume of the synthesized speech.

4. Lemmatizers

The term "lemmatization" refers to the act of normalizing the form of an (inflected) word into the form used as the headword in a dictionary, glossary or index [12]. The importance of lemmatization is evident considering that any dictionary is useless without knowledge of the headword to be searched.

In the electronic version of the KORAIS dictionary the language independent lemmatizer described in [13] was employed. This lemmatizer achieves the automatic induction of the normalized form (lemma/headword) of regular and mildly irregular words with no direct supervision using language-independent algorithms. Specifically, two string distance metric models are employed and the final lemmatized form of the input word results as a combination of the output of these two models which is based on the string similarity and the most frequent inflectional suffixes of the

language in question.

The first similarity model employed is the Levenshtein Distance algorithm [14], (also known as Edit Distance algorithm), which is a simple dynamic programming algorithm that addresses the problem of character string matching based on the notion of a primitive edit operation (i.e. substitution, insertion or deletion of a symbol). The Levenshtein Distance of two strings A and B that belong to the same alphabet is the minimum number of single character insertions, deletions and substitutions required to transform A into B. The Levenshtein Distance, denoted by $d_L(A, B)$, can be easily computed using the dynamic programming scheme shown in Table 1.

Table 1. Dynamic programming scheme for the computation of the Levenshtein distance.

$$d_L[i, j] = min\begin{cases} d_L[i-1,j]+c[A_i,\varepsilon], \\ d_L[i ,j-1]+c[\varepsilon,B_j], \\ d_L[i-1,j-1]+c[A_i,B_j] \end{cases}$$

for $i \geq 1$ *and* $j \geq 1$,
with $d[0, 0] = 0$ *and*
$d_L[i,-1] = d_L[-1, j] = \infty$

where:
c *is the cost between two strings*
ε *denotes the empty word*
A_i *is the* i^{th} *element of the string A*
B_j *is the* j^{th} *element of the string B*

The second distance metric model employed by the lemmatizer is the Dice coefficient similarity measure [15], which has been applied by [16] for the automatic retrieval of the lexical similarity between two strings. The Dice coefficient association factor can be computed according to the following formula:

$$sim_{Dice} = \frac{2 \times \left| bigrams(x) \cap bigrams(y) \right|}{\left| bigrams(x) \right| + \left| bigrams(y) \right|} \qquad (1)$$

where *bigrams* is the function which reduces a word to a multi-set of character bigrams and x and y are the examined strings.

Obviously, according to formula 1, the Dice similarity coefficient is a real number representing the number of matching bigrams or consecutive pairs of characters between the examined strings. Therefore, if the words under consideration share exactly the same bigrams, then a coefficient association of 1.0 would be returned by the model, whilst if they do not have any bigrams in common, the coefficient association would be equal to zero.

The employed lemmatizers achieve the mapping of any input word to the headwords included in KORAIS by combining the aforementioned similarity measures. In particular, initially both models are executed for the input word and two different result sets are returned (one by each model). Then, a combination of the returned sets is performed and the final list containing all the results of both models hierarchically ranked is constructed according to the following conditions:

a) If a target word is contained in both returned result sets, then place it high on the final list of returned results.

b) Else, hierarchically rank the target words based on their Levenshtein distance score or their Dice coefficient value. Target words with a Dice coefficient value close to one are placed higher than target words with relatively high Levenshtein Distance and vice versa.

Moreover, in order to further increase the performance of the lemmatizers, all input words undergo an iterative inflectional suffix removal procedure before they are forwarded as input to the above described similarity algorithms.

The performance of the KORAIS lemmatizers was experimentally evaluated using two test files (one for Modern Greek and one for English) containing 3,000 regular and mildly irregular words in various written forms each. The results suggested that the lemmatizers can perform sufficiently well for both the Modern Greek and the English language, yielding a higher than 95% accuracy for both examined languages. Furthermore, the results indicated that the performance of the algorithms is significantly improved when the input word undergoes an inflectional suffix removal process.

5. Grapheme-to-Phoneme Converter

The Grapheme-to-Phoneme conversion system for Greek was developed at WCL in the framework of the [17]. It is based on the Computer Phonetic Alphabet (CPA) standardized in this project as adapted for the Greek language [18]. The CPA was used later as a basis for the creation of the SAMPA alphabet [19] together with the then existing two other ALVEY and COST alphabets. It differs from the SAMPA mainly in the coding of diphthongs and affricates.

The Greek Grapheme-to-Phoneme conversion algorithm consists of a set of rules for the FONPARS1 software developed at Nijmegen University [20]. For a good number of Greek sounds, there is a phoneme-to-letter correspondence, so we have rules that are not context dependent. The main difficulties appear in the cases of "iotacism", where the sound /i/ has many different orthographic forms, as well as in diphthongs and in double consonants.

The performance of the algorithm was tested in Greek newspaper texts with 10,558 different words from a corpus of 100,000 words [21]. Five subsets were built with the first one including the 1,000 most frequent words and the four others equally covering the remaining percentage from the whole set. Considering three types of errors (i.e. only in the stress position, only in the transcription, or both in the stress position and the transcription) the overall accuracy of the five tested subsets was 98.7%. The conversion accuracy was also checked in 150 town names, 31 capital names and 150 Christian names. The results are given in Table 2.

Table 2. Performance of the Grapheme-to-Phoneme Converter.

Newspapers	***National Towns***	***Capitals***	***Christian Names***
98.7%	*97.3%*	*93.5%*	*97.3%*

Additional rules incorporated after the above test, further improved the newspaper conversion accuracy to more than 99%.

6. Syllabification System

By the term syllabification we refer here to the separation of written words into syllables. Dividing Greek words into syllables highly facilitates the learning of Greek pronunciation. This, because the pronunciation of Greek bears a consistent relation to the spelling of its syllables, i.e. the pronunciation of a syllable is not changed according to its context. Thus, by having any word separated into syllables using the syllabification system the learners of Greek have no problem in pronouncing it, provided they have learned the phonetic value of the Greek letters and their combination into syllables. Since Greek is a highly inflectional and morphologically complex language, syllabification is not easily performed.

The syllabification employed in the KORAIS dictionary is performed in two steps: initially a naïve syllabification function acts on the word or phrase to be broken into syllables, and then the resulting string from the first function is fed to a strict syllabification function that fine-tunes the result. Specifically, the naïve syllabification function breaks an input word or phrase into syllables simply by retaining only one vowel per syllable, ignoring any strict grammatical rules of the Greek language. For instance, the input word *"άλλωστε"(besides, moreover),* is split into *"ά-λλω-στε" by* the naïve syllabification function.

In the second step the strict syllabification function performs a correction of the syllabification based on a series of grammatical rules described by Triantafyllidis [22]. For example

- A consonant between two vowels, is syllabized with the second vowel.
- Two consonants between two vowels are always syllabized with the second vowel, if and only if there exists a valid Greek word starting with these consonants
- The diphthongs *μπ, ντ, γκ* and *γγ* cannot be separated during the syllabification process.

By applying the rules to the output string of the naïve syllabification function, the correctly syllabized word or phrase is derived. In the above example, the input String *"ά-λλω-στε"* as produced by the naïve syllabification function is converted to *"άλ-λω-στε"*. The system has been implemented using the Java programming language. Its accuracy checked in 2000 different words proved to be higher than 99.5%

7. Conclusion

In this paper the electronic version of a large Greek-English dictionary enriched with various speech and language processing tools was presented. The KORAIS dictionary features unique facilities for presenting, retrieving and handling the stored information through a user-friendly GUI and has incorporated a series of linguistic tools: TTS synthesizers for Greek and English, lemmatizers for Greek and English, Grapheme-to-Phoneme converter for Greek, syllabification system for Greek, etc. We hope that these tools will prove to be an important aid both for Greek and English language learners.

8. Acknowledgements

The authors warmly thank all those who have worked in the creation of the KORAIS Dictionary, the GSRT for the support of the project in its initial phase and the STAVROS NIARCHOS FOUNDATION for covering the publication costs.

9. References

[1] Nesi H., *Electronic Dictionaries in Second Language Vocabulary Comprehension and Acquisition: the State of the Art*, EURALEX 2000, Aug. 8-12, Stuttgart Germany, pp. 839-847.

[2] Kokkinakis G., *Electronic Dictionaries Integrating Multimedia and Speech & Language Technologies*, SPECOM' 2001, pp. 6-7, Oct. 29-31, Moscow, Russia.

[3] Kokkinakis G., *Development of a Large Bilingual Electronic Dictionary Incorporating Speech and Language Processing Tools*, SPECOM' 2007, Vol. 1, pp. 95-98, Oct. 15-18, Moscow, Russia.

[4] Black A., Taylor P., *The Festival Speech Synthesis System, Technical Report HCRC/TR-83*, University of Edinburgh, Scotland, (1997), http://www.cstr.ed.ac.uk/projects/festival.html

[5] Dutoit T., *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, Dordrecht, 1997.

[6] Hunt, M., Zwierynski, D. and Carr, R. *Issues in high quality LPC analysis and synthesis*, Eurospeech'89, vol. 2, pp. 348-351, Paris, France. 1989.

[7] Weide L. Robert, 1998, *CMU Pronunciation Dictionary*.[online]. [cited 2002-03-01]. URL: http://www.speech.cs.cmu.edu/cgi-bin/cmudict/.

[8] Ostendorf M., Price P., and Shuttuck-Hufnagel S., *The Boston University Radio News Corpus, BU Technical report ECS-95-001*. 1995.

[9] Chung Hyunsong, and Huckvale A. Mark, *Linguistic factors affecting timing in Korean with application to speech synthesis*, Proceedings of Eurospeech 2001, Denmark.

[10] Zervas, P., Fakotakis, N. & Kokkinakis, G. (2008), *Development and evaluation of a prosodic database for Greek speech synthesis and research*, Journal of Quantitative Linguistics, 15 (2), 154-184.

[11] Lazaridis A., Zervas P., Kokkinakis G., *Segmental duration modeling for Greek Speech Synthesis*, Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Patras, Greece, 2007, pp. 518-521.

[12] C. & G. Merriam Co., *Webster's Revised Unabridged Dictionary*, (1913), http://www.thefreedictionary.com/lemmatise

[13] Lyras D., Sgarbas K., Fakotakis N.: *Applying Similarity Measures for Automatic Lemmatization: A Case Study for Modern Greek and English*, International Journal on Artificial Intelligence Tools 17(5): 1043-1064 (2008)

[14] Levenshtein V. I., *Binary codes capable of correcting deletions, insertions and reversals*, Soviet Physics Dokl., 10 February 1966, Vol. 10(8), 1966, pp. 707-710

[15] Dice, Lee R., *Measures of the amount of ecologic association between species*, Journal of Ecology, Vol. 26, 1945, pp. 297-302.

[16] Adamson G. W. and Boreham J., *The use of an association measure based on character structure to identify semantically related pairs of words and document titles*, Information Storage and Retrieval, Vol. 10, 1974, pp. 253 - 260.

[17] ESPRIT project 291/860, *Linguistic Analysis of the European Languages*, 1985-1989, Final Report, EC, 1989.

[18] KORAIS: Greek-English Dictionary, p. xviii, Patras University Press, 2008

[19] SAMPA, ESPRIT project 1541:SAM, 1989, http://www.phon.ucl.ac.uk/home/sampa/index.html

[20] FONPARS1 PHONOLOGY COMPILER, Institute of Phonetics, University of Nijmegen, The Netherlands, June 1985.

[21] Gibbon D. Moore R., Winski R., *Handbook of Standards and Resourses for Spoken Language Systems*, Mouton de Gruyer, 1997 pp.514.

[22] Triantafyllidis M., *"Modern Greek Grammar"*, OEDB, Athens, 1977.