MODERN GREEK DIALECT LEXICOGRAPHY:
AN ONLINE LEXICAL DATABASE FOR CYPRIOT GREEK

MARIANNA KATSOYANNOU, GEORGIOS KOUROUPETROGLOU
SPYROS ARMOSTI, KYRIAKI CHRISTODOULOU, GERASSIMOS XYDAS
University of Cyprus, University of Athens

'Syntychies' is the first online lexical database for Cypriot Greek dialect, with enhanced sorting and searching functionalities and a text-to-speech feature for listening to the pronunciation of the words. The research goal of the project focuses on the study of Cypriot Greek vocabulary and its written representation. Three main principles guide the structure of the website: accessibility, efficiency, and user friendliness.

'Syntychies' is a lexicographic research project for the production of linguistic resources focusing on the study of Cypriot Greek (henceforth CG) vocabulary, its pronunciation and its orthographic representation. This project was undertaken at the University of Cyprus during 2006-2010. An online web service website has been created in order to allow access to the 'Syntychies' lexical database (http://lexcy.library. ucy.ac.cy/).

Three main principles guide the structure and the form of the website: the accessibility, the efficiency and the user friendliness. In this article the graphical user interface will be presented, with a short description of the data shown on screen; a presentation of the sorting and searching capabilities follows and then the text to speech features are described. The Graphical User Interface is shown in Figure 1.

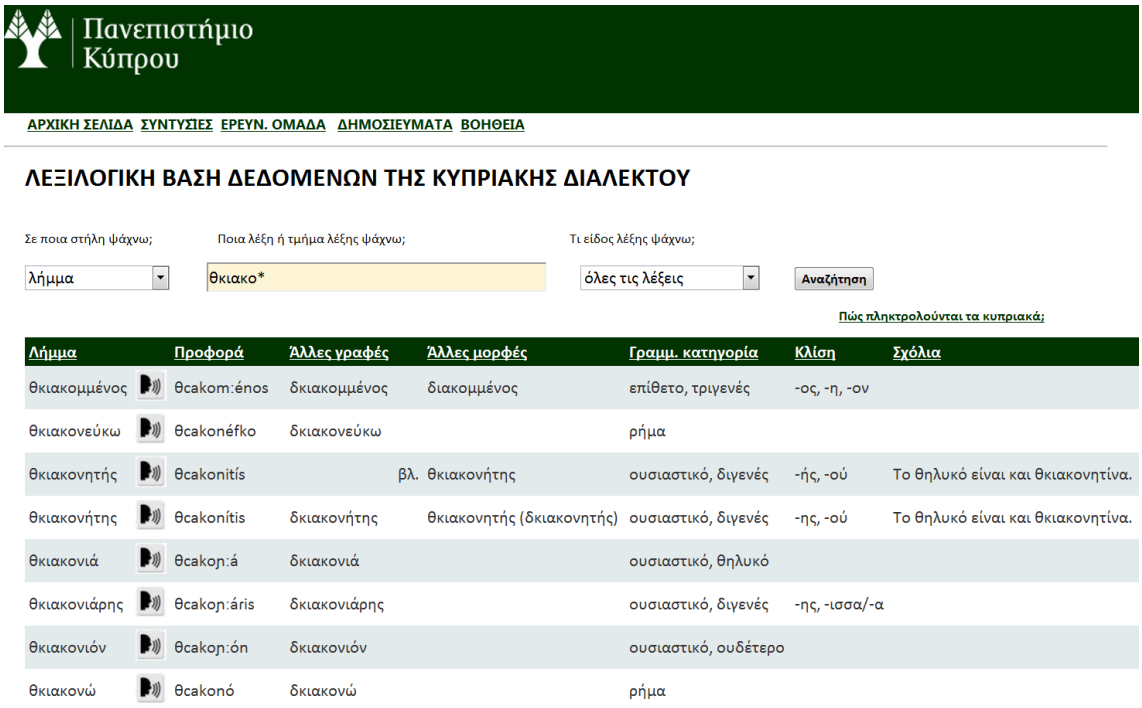

Figure 1: The 'Syntychies' graphical user interface.

The graphical user interface of the web service has minimal design and constitutes the homepage of the 'Syntychies' website. The fields currently are in Modern Greek but there are plans for extended language support. The search bar is predominant at the top of the screen; at the left and at the right of the search bar there searching tools -combo boxes- that allow the user to make sophisticated research on the database. The user exploits the lexical database mainly through this tool, specifying the combinations of search criteria (such as grammatical categories or parts of the words) that are to be met. A table follows with the data.

The columns of the data-table contain the following information:

1. Lemma (Λήμμα): The lemma is the head of the record. A standard search executes a search on this column.
2. Voice (Φωνή): Contains buttons that allow the user to hear the pronunciation of the word in the Lemma field.
3. Allomorphs (Άλλες μορφές): Morphological and phonetic variants are represented in this field.
4. Allographs (Άλλες γραφές): Since there is no standardized orthography for Cypriot Greek, alternative spellings are presented.
5. Grammatical Category (Γραμματική κατηγορία): Indicates the grammatical category to which the lemma belongs.
6. Pronunciation (Προφορά): Phonetic transcription of the lemma in the International Phonetic Alphabet.
7. Inflection (Κλίση): Description of nouns and adjectives' morphological class.
8. Comments (Σχόλια): Further information, generally morphological, concerning selected lemmas.

The web-service allows the user to sort each table in ascending or in descending order, just by double clicking each column header. The user can search for any given word just by typing the word in the search bar and pressing enter or clicking at the button labeled 'Αναζήτηση' (search). The build-in search capabilities of the web-service allow finer search such as:

1. By using expressions: The star key < * > stands for one or more characters while the question mark < ? > represents one character; so the search results for *τος, are all the existing CG words ending in < τος >, while a search for ?τος will provide one result: έτος ("year").
2. Search in other fields: Searching is also enabled for the fields [Other forms (Άλλες Μορφές)], [Other spellings (Άλλες Γραφές)] and [Pronunciation (Προφορά)] by making the appropriate selection in the left combo-box. This capability allows the user to find information that is not part of the main lemma. The user can even search for different pronunciations in IPA; this functionality is rear in e-dictionaries but it is of utmost importance for researchers, linguists, phoneticians and speech pathologists.
3. Constraining the search: The user may choose to reduce the search results to certain grammatical categories by making the appropriate selection in the right combo-box.

The text to speech component provides an auditory presentation of the words. It is based on triphone selection speech synthesis trained from a phonetically balanced subset of the lexicon corpus. The training data set consists of 2092 isolated words, which have been chosen to offer a balanced distribution of all the observed intonational phenomena for reading words and some small phrases in a lexicon application domain. Due to this application domain's nature, prosody is highly predictable with limited variance. On the other hand, the segmental quality is of major importance, to pronounce lexicon entries accurately. Thus, we chose to perform unit selection based on the segmental content and the lexical stress information, leading to high quality and phonetically accurate synthesized speech. The DEMOSTHeNES text-to-speech system (Xydas and Kouroupetroglou, 2001) has been used as the development platform and the synthesis module chain is currently capable of performing the phoneme-to-speech conversion. For the letter-to-phoneme task, we developed an off-line set of transformation rules, which resolves the ambiguity occurring with homographs (cf. Jurafski and Martin, 2000: 791) and informs the CG native and non-native speakers about the pronunciations of a lemma.

In a nutshell, 'Syntychies' is the first electronic lexical database for CG, with an online web-service and use of modern technologies. Th project aims to provide free online lexicographic recourses not only to academics, researchers and scholars but to anyone interested in CG.

References

Armosti S., Christodoulou K., Katsoyannou, M. and Themistocleous, C. (to appear) «Writing in Cypriot Greek: the need for standardisation and its importance for dialectal lexicography», *In Dialogue on dialect standardization*. Cambridge: Cambridge Scholars Publishing.

Jurafski, D. and J. Martin. 2000. *Speech and Language Processing*. New Jersey: Prentice Hall.

Themistocleous C. 2011. "Computational Greek Phonology: IPAGreek." *Proceedings of 10th International Conference of Greek Linguistics*.

Themistocleous Ch., Katsoyannou M., Armosti S. and Christodoulou K. «Cypriot Greek Lexicography: An online lexical database», *Proceedings of the XV EURALEX International Congress (Oslo, Norway), University of Oslo*, p. 889-891.

Xydas G. and G. Kouroupetroglou. 2001. 'The DEMOSTHeNES Speech Composer. *Proceedings of the 4th ISCA Tutorial and Workshop on Speech Synthesis*, SSW4: 167–172.