# ΑΝΑΠΤΥΞΗ ΗΛΕΚΤΡΟΝΙΚΩΝ ΛΕΞΙΚΩΝ ΤΗΣ ΕΛΛΗΝΙΚΗΣ ΓΛΩΣΣΑΣ

**Γ. Κοκκινάκης, Ε. Δερματάς, Ε. Κουτσογεωργοπούλου**

## ΠΕΡΙΛΗΨΗ

Η ομάδα γλωσσικής τεχνολογίας του ΕΕΤ ανέπτυξε πρόσφατα τρία ηλεκτρονικά λεξικά της Ελληνικής γλώσσας στα πλαίσια των Ευρωπαικών προγραμμάτων SOCRATES/LINGUA και του προγράμματος ΕΠΕΤ ΙΙ «Γλωσσικής Τεχνολογίας» της ΓΓΕΤ.

- Ηλεκτρονικό λεξικό προφοράς και χρήσης της σύγχρονης Ελληνικής Γλώσσας για Ξένους.
- Ηλεκτρονικό Λεξικό προφοράς και χρήσης της Ελληνικής (Γραικανικής) διαλέκτου της Νότιας Ιταλίας.
- ΚΟΡΑΗΣ: Καινοτόμο Ελληνο-Αγγλικό Ηλεκτρονικό λεξικό 100.000 λημμάτων.

Η ανάπτυξη των ανωτέρω λεξικών στηρίχθηκε στην λεξικογραφική υποδομή του ΕΕΤ που έχει δημιουργηθεί τα τελευταία 10 χρόνια και περιλαμβάνει: Ηλεκτρονικά κείμενα και ηχογραφήσεις προφορικού λόγου, λογισμικό επεξεργασίας ηλεκτρονικών κειμένων φυσικής γλώσσας, βάσεις ηλεκτρονικών λεξικών.

# DEVELOPMENT OF ELECTRONIC DICTIONARIES OF THE GREEK LANGUAGE

**G. Kokkinankis, E. Dermatas, E. Coutsogeorgopoulos**

## SUMMARY

The Speech & Language Technology group of the Wire Communcations Laboratory (WCL) recently developed three electronic dictionaries of the Greek Language within the framework of the European programmes SOCRATES /LINGUA and the programme “Language Technology” of the Greek General Secretariat of Research and Technology:

- Electronic Dictionary of Pronunciation and Usage of the Modern Greek Language
- Electronic Dictionary of Pronunciation and Usage of the Graecanic Dialect of South Italy
- KOPAIS – Innovative Greek-English Electronic Dictionary of 100.000 Lemmas

The development of the above dictionaries was based on the lexicographic infrastructure of WCL which has been created in the last ten years and consists of electronic corpora, speech databases, databases of electronic dictionaries and software for the processing of large corpora.

The above dictionaries have a series of innovative features which greatly facilitate their use.

## 1 ΕΙΣΑΓΩΓΗ

Το εργαστήριο Ενσύρματης Τηλεπικοινωνίας (ΕΕΤ) του Τμήματος Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών του Πανεπιστημίου Πατρών, εργάζεται από ετών στην περιοχή Γλωσσικής Τεχνολογίας (Επεξεργασία Ομιλίας και Φυσικής Γλώσσας) με οργανωμένες ερευνητικές ομάδες, παράλληλα με τη δραστηριότητά του στις περιοχές Τηλεπικοινωνιών και Τεχνολογίας Ήχου. Στο διάστημα των τελευταίων έξι ετών, η ομάδα ηλεκτρονικής λεξικογραφίας του ΕΕΤ, σε συνεργασία με εξωτερικούς ειδικούς επιστήμονες στην Ελλάδα, Αγγλία και Ιταλία, ανέπτυξε τρία ηλεκτρονικά λεξικά της Ελληνικής γλώσσας για διάφορες χρήσεις. Στην συνέχεια δίνονται συνοπτικά στοιχεία για τα λεξικά αυτά.

## 2. ΗΛΕΚΤΡΟΝΙΚΑ ΛΕΞΙΚΑ

### Ηλεκτρονικό Λεξικό Προφοράς και Χρήσης της Σύγχρονης Ελληνικής Γλώσσας για Ξένους

- Αναπτύχθηκε στα πλαίσια του Ευρωπαϊκού προγράμματος SOCRATES/LINGUA 1995-1-GR-30 (1995-1997) σε συνεργασία με το εκπαιδευτικό κέντρο THE BRASSHOUSE CENTRE του Δήμου Birmingham (GB).
- Χαρακτηριστικά του λεξικού:
  - 45.000 λήμματα + 10.000 παραδείγματα χρήσης.
  - Ανάπτυξη με βάση τη συλλογή κειμένων του ΕΕΤ (Corpus based lexicography) και τις συμβουλές του Brasshouse.
  - Φωνητική απόδοση των λημμάτων και μέρους των παραδειγμάτων (Ηχογραφήσεις).
  - Φωνητική γραφή κάθε λήμματος (Διεθνές φων. αλφάβητο).
  - Γραμματική κατηγορία με παραπομπή σε κλιτικό παράδειγμα.
  - Πρόσβαση σε λήμμα από μορφή λήμματος.
  - Πρόσβαση με ορθή και εσφαλμένη ορθογραφική γραφή λήμματος.
  - Πρόσβαση με φωνητική γραφή (Παροχή εναλλακτικών ορθογραφικών μορφών από φωνητικό λεξικό).
  - Αντίστροφο λεξικό (ταξινόμηση λημμάτων με βάση την κατάληξη).
  - Πίνακες τοπονυμιών και ονομάτων.

**Ηλεκτρονικό Λεξικό της Ελληνικής (Γραικανικής) Διαλέκτου της Νότιας Ιταλίας**

- Αναπτύχθηκε στα πλαίσια του Ευρωπαϊκού προγράμματος SOCRATES/LINGUA 96-06-MDD-013700 (1996-1998) σε συνεργασία με ειδικούς επιστήμονες στην Ελλάδα και Ιταλία (Δρ. Α. Μποτίνης, S. Minuto, A. Rocco, κ.ά.)
- Χαρακτηριστικά του λεξικού:
  - 10.0000 περίπου λήμματα και παραδείγματα χρήσης με μετάφραση στα Ελληνικά και Ιταλικά και αντίστροφα (τρίγλωσσο λεξικό).
  - Πολλαπλές ηχογραφήσεις κάθε λήμματος και μέρους των παραδειγμάτων σύμφωνα με τις τοπικές διαφορές προφοράς.
  - Φωνητική γραφή κάθε λήμματος (Διεθνές φωνητικό αλφάβητο).
  - Γραμματική κατηγορία και πληροφορίες σχετικές με το λήμμα (τόπος χρήσης) και τον ομιλητή (γένος, ηλικία, επάγγελμα, καταγωγή, κλπ).
  - Αξιοποίηση υπάρχοντος υλικού (λεξικά, κείμενα, ηχογραφήσεις, κλπ).
  - Ηχογραφήσεις επί τόπου.
  - Χρήση λεξικού όπως στο ηλεκτρονικό λεξικό για ξένους (πρόσβαση, κλπ).

**ΚΟΡΑΗΣ - Καινότομο Έλληνο-Αγγλικό Ηλεκτρονικό Λεξικό 100.000 Λημμάτων**

- Αναπτύχθηκε στα πλαίσια του προγράμματος «Γλωσσική Τεχνολογία» της ΓΓΕΤ 98 ΓΤ 45 (2000-2001) σε συνεργασία με ειδικούς λεξικογράφους και μεταφραστές.
- Χαρακτηριστικά του λεξικού:
  - Ηχητική απόδοση της προφοράς (Ελληνικής και Αγγλικής) κάθε λήμματος από ενταμιευμένη ηχογράφηση
  - Μετατροπή οποιασδήποτε πληκτρολογούμενης λέξης ή πρότασης της ελληνικής από την ορθογραφική της μορφή (η οποία μπορεί να είναι και λανθασμένη) στην αντίστοιχη φωνητική μορφή (IPA).
  - Μετατροπή οποιουδήποτε λήμματος του λεξικού από την πληκτρολογούμενη φωνητική του μορφή (η οποία μπορεί να είναι και λανθασμένη) στην αντίστοιχη ή αντίστοιχες ορθογραφικές μορφές (προσδιορισμός ομοήχων)
  - Εντοπισμός του λήμματος από οποιονδήποτε πληκτρολογούμενο τύπο του μέσω μορφολογικού αναλυτή (π.χ. από τον τύπο «είδες», εντοπισμός του λήμματος «βλέπω» και εμφάνιση στην οθόνη όλης της κλίσης του λήμματος).

## 3. ΑΝΑΠΤΥΞΗ ΤΩΝ ΛΕΞΙΚΩΝ

Η ανάπτυξη των ανωτέρω λεξικών στηρίχθηκε στην λεξικογραφική υποδομή του ΕΕΤ που έχει δημιουργηθεί τα τελευταία 10 χρόνια και περιλαμβάνει:

Ηλεκτρονικά κείμενα και ηχογραφήσεις προφορικού λόγου:

1. Συλλογή ηλεκτρονικών κειμένων 50 εκατομμυρίων λέξεων
2. Ηλεκτρονικό κείμενο 310.000 λέξεων με λεπτομερή γραμματική ανάλυση
3. Βάσεις δεδομένων προφορικού λόγου (SAM, POLYGLOT, SIEMENS)

Λογισμικό επεξεργασίας ηλεκτρονικών κειμένων φυσικής γλώσσας:

1. Λεξικογραφική επεξεργασία των κειμένων σε πραγματικό χρόνο
2. Σύστημα αυτόματης γραμματικής ανάλυσης
3. Σύστημα αυτόματης μετατροπής κειμένου από ορθογραφική σε φωνητική γραφή και αντίστροφα

Βάσεις ηλεκτρονικών λεξικών

1. Ορθογραφικό λεξικό 100.000 λημμάτων με παραδείγματα χρήσεως
2. Λεξικό 1.5 εκατομμυρίων κυρίων ονομάτων σε ορθογραφική και φωνητική γραφή

Τα ανωτέρω λεξικά παρουσιάζουν μια σειρά καινοτομιών, οι οποίες διευκολύνουν σημαντικά τον χρήστη.

## ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Dermatas E. and Kokkinakis G., "Lexithiras: Multilingual Corpus Based Lexicography On PCs", Language Engineering on the Information Highway, Santorini, Greece, 26-30, Sept. 1994

[2] Γ. Κοκκινάκης, «Γλωσσική Τεχνολογία: Ερευνητικά αποτελέσματα από τη συμμετοχή του Εργαστηρίου Ενσύρματης Τηλεπικοινωνίας του Πανεπιστημίου Πατρών σε Ευρωπαϊκά προγράμματα την τελευταία δεκαετία». Καινοτομία, Έρευνα και Τεχνολογία, Τεύχος 5, Ιαν-Μαρτ. 1997.

[3] Koutsogeogopoulos E., Giouli P., Dermatas E. and G. Kokkinakis, "Developing of Electroning Dictionaries of the Greek Language", 19th International Conference for the Greek Language, Aristotele University of Thessaloniki, Sept. 1998.

[4] Kokkinakis G. and Dermatas E., "A web-lab for demonstrating and experimenting with speech & language technology", ESCA Workshop, 1998

[5] Dermatas E. and Kokkinakis G., "LEXITHIRAS: Corpus Based Lexicography on PCs", TSD 98, Text, Speech, Dialogue, Brno, Czech Republic, 15-20, Sep. 1998

[6] G. Kokkinakis, H. Coutsogeorgopoulos, E. Dermatas, G. Kaitsas, “Electronic Dictionary of Pronunciation and Usage of the Graecanic Dialect of Southern Italy”, LREC-2000, Athens, 31 May-1June, 2000

[7] H. Coutsogeorgopoulos, E. Dermatas, G. Kokkinakis, "A Monolingual Electronic Dictionary For The Pronunciation And Usage Of Modern Greek For Foreigners", COMLEX 2000, Pyrgos, Greece, 83-88, 21-22 Sept. 2000

[8] H. Coutsogeorgopoulos, G. Kokkinakis, E. Dermatas, "KORAIS: A Large Electronic Greek-English Dictionary With Spoken Pronunciation, COMLEX 2000, Pyrgos, Greece, 127-130, 21-22 Sept. 2000

[9] George Kaitsas, Helen Coutsogeorgopoulos, Evangelos Dermatas, George Kokkinakis, "Pronunciation and Usage of the Graecanic Dialect of Southern Italy on CDROM and Web", COMLEX 2000, Pyrgos, Greece, 35-140, 21-22 Sept. 2000,

[10] Ε. Δερματάς και Γ. Κοκκινάκης, “Στατιστικά Εργαλεία Εντοπισμού και Εξαγωγής Λεξικογραφικής Πληροφορίας από Ηλεκτρονικά κείμενα πολύ μεγάλου μεγέθους για Μονόγλωσσα Λεξικά - Μετρήσεις Απόδοσης στην Ελληνική Γλώσσα”, 22 International Conference for the Greek Language, Aristotele University of Thessaloniki, Apr. 2001

[11] George Kokkinakis, “Electronic Dictionaries Integrating Multimedia and Speech Language Technologies”, SPECOM 2001, Moscow, 27-28 Oct. 2001, (to be presented).

Γ. Κοκκινάκης, Ε. Δερματάς, Ε. Κουτσογεωργοπούλου

Εργαστήριο Ενσύρματης Τηλεπικοινωνίας (ΕΕΤ)
Τμήμα Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών
Πανεπιστήμιο Πάτρας
e-mail: gkokkin@wcl.ee.upatras.gr