ΚΟΡΑΗΣ: Έντυπο και Ηλεκτρονικό Ελληνο-Αγγλικό Λεξικό

Γεώργιος Κοκκινάκης, Ελένη Κουτσογεωργοπούλου, Δημήτριος Λύρας, Κυριάκος Σγάρμπας

ΠΕΡΙΛΗΨΗ

Στην παρούσα ανακοίνωση παρουσιάζεται το έντυπο και ηλεκτρονικό λεξικό ΚΟΡΑΗΣ, το οποίο αναπτύχθηκε στα πλαίσια του ερευνητικού/τεχνολογικού προγράμματος ΕΠΕΤ ΙΙ της ΓΓΕΤ από κοινοπραξία Πανεπιστημίων κ.α. με ανάδοχο το Πανεπιστήμιο Πατρών και στη συνέχεια αναμορφώθηκε από μέλη της Ομάδας Επεξεργασίας Ομιλίας και Γλώσσας του Εργαστηρίου Ενσύρματης Τηλεπικοινωνίας (ΕΕΤ) και ειδικούς εξωτερικούς συνεργάτες. Στόχος ήταν η δημιουργία μεγάλου, σύγχρονου και καινοτόμου λεξικού με αξιοποίηση των δυνατοτήτων που προσφέρουν οι νέες τεχνολογίες της υπολογιστικής λεξικογραφίας, της άντλησης πληροφοριών από το διαδίκτυο, της ενταμίευσης πολυμεσικής πληροφορίας σε DVD καθώς και η ενσωμάτωση στο ηλεκτρονικό μέρος του λεξικού εργαλείων γλωσσικής τεχνολογίας.

Το λεξικό περιέχει πλούσιο γλωσσικό υλικό της καθομιλουμένης αλλά και καλλιεργημένης Ελληνικής, που περιλαμβάνει μεταξύ άλλων πλήθος όρων διαφόρων ειδικών πεδίων. Συγκεκριμένα: 81.515 λήμματα (μεταξύ των οποίων και 14.371 όρους) με διεξοδική ερμηνεία των σημασιών τους, 192.592 αποδόσεις στην Αγγλική όλων των σημασιών, 50.106 παραδείγματα χρήσεως με μετάφραση.

Επίσης περιλαμβάνει γραμματικές και υφολογικές πληροφορίες, καθώς και απόδοση της προφοράς των λημμάτων με σύμβολα του φωνητικού αλφαβήτου υπολογιστή (CPA - Computer Phonetic Alphabet) στην έντυπη μορφή και ηχογραφήσεις στην ηλεκτρονική.

Το ηλεκτρονικό λεξικό διαθέτει επί πλέον μοναδικές δυνατότητες αναζήτησης πληροφορίας σε ολόκληρο ή τμήμα του ελληνικού και του αγγλικού μέρους, καθώς και μία σειρά εργαλείων γλωσσικής τεχνολογίας, πολύ χρήσιμων στον σπουδαστή της ελληνικής και αγγλικής γλώσσας: λημματοποιητές της Ελληνικής και Αγγλικής, συνθέτες ομιλίας από κείμενο της Ελληνικής και Αγγλικής, μετατροπέα γραπτού κειμένου της Ελληνικής σε φωνητική μορφή (CPA), συλλαβιστή της Ελληνικής, κ.α.

Το έντυπο λεξικό απευθύνεται, λόγω μεγέθους, σε προχωρημένους Έλληνες σπουδαστές και χρήστες της Αγγλικής γλώσσας και αγγλόφωνους που σπουδάζουν ή χρησιμοποιούν την ελληνική γλώσσα. Το ηλεκτρονικό λεξικό μπορεί να χρησιμοποιηθεί σε οποιοδήποτε στάδιο εκμάθησης.

KORAIS: A printed & electronic Greek-English dictionary

George Kokkinakis, Helen Coutsogeorgopoulos, Dimitrios P. Lyras, Kyriakos Sgarbas

ABSTRACT

In the present paper a large Greek-English dictionary named KORAIS is presented. This dictionary, available in both printed and electronic form, is intended for advanced Greek learners and users of

English, as well as for English speakers who are learning or using the Greek language. It contains a wealth of language material of both common use and erudite Greek. More specifically, it contains: 81,515 entries with detailed clarification of their meaning, 192,592 translations into English of all meanings, 50,106 usage examples with translation. It also includes grammatical and usage information as well as the pronunciation of each entry with computer phonetic alphabet (CPA) symbols in the printed version and prerecorded speech in the electronic version. Furthermore, 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 tools which are very useful for learners of the Greek and English language: lemmatizers for Greek and English, Text-to-Speech systems for Greek and English, a transcriber for conversion of any Greek written text into CPA-text, a syllabification system for Greek, and other tools.

0 Εισαγωγή

Επί αιώνες, η λεξικογραφία ήταν συνυφασμένη με επίπονη, χρονοβόρα και πολυδάπανη συλλογή και αποδελτίωση λημμάτων και κατόπιν χειρωνακτική επεξεργασία του υλικού.

Η εμφάνιση των ηλεκτρονικών υπολογιστών άλλαξε ριζικά το τοπίο δίνοντας τη δυνατότητα επεξεργασίας τεράστιου όγκου κειμένων σε σύντομο διάστημα και εξαγωγής λεπτομερειακών και αντικειμενικών στοιχείων σχετικά με την εμφάνιση και χρήση λέξεων, τη γραμματική, τη σύνταξη, τη σημασία τους κ.ο.κ. Κατόπιν, η δημιουργία ψηφιακών δίσκων (CDs και DVDs) και προσωπικών υπολογιστών (PCs) υψηλής ταχύτητας, άνοιξε το δρόμο δημιουργίας ηλεκτρονικών λεξικών με πολυμεσική πληροφορία – κείμενο, ήχο, γραφικά, εικόνα – και παρουσίασή της σε PC [1], ενώ η επέκταση του διαδικτύου (internet) επέτρεψε λεξικά ενταμιευμένα σε οποιαδήποτε βάση ανά τον κόσμο να είναι προσπελάσιμα από οποιονδήποτε χρήστη του δικτύου. Τέλος, η πρόσφατη ανάπτυξη της γλωσσικής τεχνολογίας έδωσε τη δυνατότητα ενσωμάτωσης στα ηλεκτρονικά λεξικά σειράς γλωσσικών εργαλείων, όπως συνθετών ομιλίας, λημματοποιητών κ.α., που μπορούν να βοηθήσουν σημαντικά τον χρήστη [2].

Το έντυπο και ηλεκτρονικό Ελληνο-Αγγλικό λεξικό με την ονομασία ΚΟΡΑΗΣ προς τιμή του μεγάλου Έλληνα λόγιου Αδαμάντιου Κοραή (1748-1833) δημιουργήθηκε με στόχο την αξιοποίηση των καινοτομιών που αναφέρθηκαν στην λεξικογραφία και των εργαλείων γλωσσικής τεχνολογίας του Εργαστηρίου Ενσύρματης Τηλεπικοινωνίας (ΕΕΤ), ώστε να δοθεί αποτελεσματική βοήθεια στους Έλληνες σπουδαστές και χρήστες της Αγγλικής καθώς και, ιδιαίτερα, στους αγγλόφωνους που σπουδάζουν ή χρησιμοποιούν την ελληνική γλώσσα.

Έτσι, όλη η λεξικογραφική εργασία πραγματοποιήθηκε σε PCs από εκπαιδευμένους λεξικογράφους και μεταφραστές, με χρήση ειδικών προγραμμάτων. Το βασικό λημματολόγιο (40.000 περίπου λήμματα με συχνότερη χρήση από το σύνολο των 81.515 λημμάτων του λεξικού) προήλθε από ανάλυση μεγάλου σώματος κειμένων του ΕΕΤ, ενώ τόσο στην κατασκευή του ελληνικού μέρους (πλαισίου), όσο και στη μετάφραση χρησιμοποιήθηκαν, επιβοηθητικά, λεξικογραφικές πηγές του διαδικτύου. Επί πλέον χρησιμοποιήθηκαν εξελιγμένες μέθοδοι δημιουργίας του ηλεκτρονικού λεξικού σε DVD με πολυμεσική πληροφορία.

Η ανάπτυξη του ΚΟΡΑΗΣ έγινε αρχικά (1999 – 2001) στο πλαίσιο του προγράμματος ΕΠΕΤ ΙΙ της ΓΓΕΤ από κοινοπραξία στην οποία συμμετείχαν μεταξύ άλλων το Πανεπιστήμιο Πατρών ως ανάδοχος του έργου, το Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης και το Ιόνιο Πανεπιστήμιο [3]. Στη συνέχεια (2001-2008) το λεξικό ελέγχθηκε συστηματικά και εμπλουτίστηκε από την Ομάδα Επεξεργασίας Ομιλίας και Γλώσσας του ΕΕΤ και ειδικούς εξωτερικούς συνεργάτες. Στις παραγράφους 1 ως 4 που ακολουθούν παρουσιάζονται αντίστοιχα η μεθοδολογία ανάπτυξης, η δομή του λεξικού, το ηλεκτρονικό λεξικό και το συμπέρασμα.

1 Μεθοδολογία Ανάπτυξης

Η ανάπτυξη του λεξικού στηρίχτηκε στη χρήση [4]:

α) ηλεκτρονικής λεξικογραφίας σε προσωπικούς υπολογιστές (PCs), με ειδικά προγράμματα. Για το σκοπό αυτό αφενός χρησιμοποιήθηκαν λεξικογράφοι που ήδη διέθεταν τις απαιτούμενες γνώσεις, αφετέρου εκπαιδεύτηκαν νέοι λεξικογράφοι.

β) ισορροπημένου σώματος κειμένων του ΕΕΤ 60 εκατομ. λέξεων της ελληνικής γλώσσας.

γ) υπαρχόντων γλωσσικών πόρων (λεξικών κ.λ.π.) στους εταίρους.

δ) μηχανισμού ελέγχου και διόρθωσης λαθών από όλους τους εταίρους.

ε) αναγνωρισμένων μεθόδων επιλογής λημματολογίου, γραμματικής και μορφολογικής ανάλυσης, καθορισμού της σημασίας, κ.λ.π.

- Για ένα μεγάλο τμήμα του λεξικού χρησιμοποιήθηκε ο “Λεξιθήρας”, ένα σύστημα υπολογιστικής λεξικογραφίας που αναπτύχθηκε στο ΕΕΤ. *Ο λεξιθήρας* πραγματοποιεί ευρεία στατιστική ανάλυση των κειμένων, λημματοποίηση, γραμματική ανάλυση λέξεων κ.α. και παρέχει στο λεξικογράφο την δυνατότητα να επιλέξει κατάλληλα παραδείγματα χρήσεως. Μπορεί να επεξεργαστεί πολύ μεγάλα σώματα κειμένων φυσικής γλώσσας (εκατοντάδες εκατομμύρια λέξεις) [5].

- Χρησιμοποιήθηκε το Ηλεκτρονικό Λεξικό της Νέας Ελληνικής για Ξένους, 40.000 λημμάτων, που είχε αναπτυχθεί από το ΕΕΤ στα πλαίσια του προγράμματος Socrates Lingua [6]. Επίσης, για τη συμπλήρωση του λημματολογίου πέραν των βασικών 40.000 λημμάτων χρησιμοποιήθηκαν γλωσσικοί πόροι του ΑΠΘ-Τομέας Γλωσσολογίας, δόκιμα μονόγλωσσα και δίγλωσσα λεξικά της Νεοελληνικής, καθώς και λεξικά εγκατεστημένα στο διαδίκτυο [3].

- Για την επιλογή των λημμάτων χρησιμοποιήθηκαν ιεραρχικά τα εξής κριτήρια: α) συχνότητα εμφάνισης στο σώμα κειμένων, β) χρησιμότητα για τον χρήστη, γ) βαθμός δυσκολίας στο επίπεδο της μορφολογίας ή της σύνταξης. Σε κάθε περίπτωση η επιλογή γινόταν αρχικά από τον λεξικογράφο, ενώ η τελική απόφαση λαμβανόταν κατόπιν από τον υπεύθυνο της σύνταξης.

- Η ανάλυση των λημμάτων έγινε συστηματικά σύμφωνα με ***πεδία πληροφορίας*** που ορίστηκαν προκαταβολικά και ***κωδικοποιήθηκαν***. Κάθε πληροφορία, γραμματική, σημασιολογική ή άλλη καταχωρίσθηκε στο αντίστοιχο πεδίο με τον συγκεκριμένο κωδικό. Με αυτόν τον τρόπο ήταν εύκολος ο εντοπισμός και η ανάκτηση των δεδομένων.

- Ο έλεγχος του λημματολογίου, της ανάλυσης και της μετάφρασης από πλευράς συνέπειας καταχωρίσεων, ορθογραφικών κ.α. λαθών, πραγματοποιήθηκε τμηματικά από τους εταίρους της κοινοπραξίας (1999-2001). Συστηματικός, ενιαίος έλεγχος τόσο του ελληνικού μέρους όσο και της μετάφρασης έγινε στο διάστημα 2001-2008, από μέλη της Ομάδας Επεξεργασίας Ομιλίας και Γλώσσας και από ειδικούς εξωτερικούς συνεργάτες. Παράλληλα με ευρείες διορθώσεις έγινε αναμόρφωση της ύλης, ώστε το ΚΟΡΑΗΣ να πάρει τη σημερινή, σημαντικά διαφοροποιημένη ως προς την αρχική, μορφή.

2 Δομή του λεξικού

- Ως **λήμματα** έχουν καταχωρισθεί στο λεξικό, με αλφαβητική σειρά και μαύρα στοιχεία, αυτοτελείς λέξεις, π.χ. **βιβλίο**, **ωραίος**, **παίζω**, **αύριο, όταν,** περιφραστικές λέξεις, π.χ. **αθλητική εφημερίδα**, **Αιγαίο Πέλαγος** και λεξιλογικές φράσεις, π.χ. **ζω έναν εφιάλτη**. Η αλφαβητική καταχώριση έχει διατηρηθεί και στις περιπτώσεις περιφραστικών λέξεων και φράσεων με άρθρο στην αρχή, π.χ. **η Παλαιά Διαθήκη**.

- Στο ευρύτατο φάσμα του λεξιλογίου του σύγχρονου γραπτού και προφορικού λόγου της ελληνικής γλώσσας που καλύπτεται, έχουν περιληφθεί: λέξεις και φράσεις κοινής χρήσης, επιστημονικοί και τεχνικοί όροι, λέξεις της λογοτεχνίας, κύρια ονόματα σημαντικών προσώπων και τοπωνυμίων, ηχομιμητικές λέξεις, νεολογισμοί του γραπτού και προφορικού λόγου (λέξεις ξενικής προέλευσης, δάνεια, κ.α.), λέξεις και φράσεις της

αργκό, άκλιτοι ή μεμονωμένοι τύποι του λόγου, ακρωνύμια. Επί πλέον έχουν καταχωρισθεί: λόγιες λέξεις που χρησιμοποιούνται σε καλλιεργημένες μορφές λόγου (επιστημονικά συγγράμματα, υψηλή αρθρογραφία, κ.λ.π.), ιδιωματικές/ λαϊκότροπες λέξεις και φράσεις.

- Κάθε λήμμα συνοδεύεται από τη ***φωνητική μεταγραφή*** του σε αγκύλες, από ***γραμματικές πληροφορίες***, από ***ορισμό της σημασίας*** ή ***των διαφορετικών σημασιών του***, όπου θεωρείται απαραίτητο, και από την ***μετάφρασή*** τους στην Αγγλική. Τα παραδείγματα χρήσεως αποτελούν ολοκληρωμένες προτάσεις, ώστε να δείχνουν τις εννοιολογικές αποχρώσεις, αλλά και το πώς συντάσσεται το λήμμα στον ρέοντα λόγο και στις δύο γλώσσες. Η φωνητική μεταγραφή ακολουθεί το διεθνές φωνητικό αλφάβητο, όπως αυτό έχει προσαρμοστεί για χρήση της ελληνικής γλώσσας σε ηλεκτρονικό υπολογιστή με συνηθισμένο πληκτρολόγιο (CPA = Computer Phonetic Alphabet), δηλ. χωρίς τα ειδικά σύμβολα του διεθνούς φωνητικού αλφαβήτου [3]. Η προσαρμογή αυτή ήταν απαραίτητη για τους χρήστες του ηλεκτρονικού λεξικού, ώστε να είναι δυνατή η εμφάνιση στην οθόνη του υπολογιστή της φωνητικής γραφής κατά την αυτόματη μετατροπή “Gr→Ph”.

- Με κεφαλαία γράμματα και συντομογραφία χαρακτηρίζεται σε πολλά λήμματα το πεδίο επιστημονικών, επαγγελματικών και άλλων κλάδων, στο οποίο αυτά χρησιμοποιούνται, π.χ. Γεωγραφία (ΓΕΩΓΡ), Ιστορία (ΙΣΤ), Ηλεκτρολογία (ΗΛΕΚΤΡΟΛ), κ.λ.π. Συνολικά γίνεται διάκριση σε 42 πεδία στα οποία έχουν καταχωρισθεί 14.371 όροι. Επί πλέον, χρησιμοποιούνται χαρακτηρισμοί επιπέδου γλώσσας ως συντομογραφίες με πεζά γράμματα σε παρένθεση, που πληροφορούν το χρήστη για το περιβάλλον χρήσεως του λήμματος και το ύφος του λήμματος: (υβρ) = υβριστικό, (χυδ) = χυδαίο, κ.λ.π.

- Για την διευκόλυνση, ιδιαίτερα του ξένου χρήστη, δεν χρησιμοποιήθηκε η συνήθης ***υπαγωγή*** και ***διάσπαση*** λημμάτων, π.χ. υποκοριστικών, μεγεθυντικών, κ.λ.π. αλλά η καταχώρισή τους ως ***αυτοτελών λημμάτων***. Έτσι, δίνεται η φωνητική μεταγραφή (και η ηχητική απόδοση στο ηλεκτρονικό λεξικό), οι γραμματικές πληροφορίες, η μετάφραση, κ.λ.π., ξεχωριστά και συγκεντρωμένα σε κάθε λήμμα.

- Από γραμματικής πλευράς, έχουν καταχωρισθεί ως λήμματα, ουσιαστικά, επίθετα, αντωνυμίες, ρήματα, μετοχές και τα άκλιτα μέρη του λόγου [3].

3 Ηλεκτρονικό Λεξικό

Το ηλεκτρονικό λεξικό ΚΟΡΑΗΣ περιέχει όλες τις πληροφορίες του έντυπου λεξικού, παρέχοντας στον χρήστη τις ακόλουθες δυνατότητες:

1. Φιλική διεπαφή για εύκολη και γρήγορη ανάκτηση των πληροφοριών.
2. Ευέλικτη αναζήτηση μέσα στους διάφορους κωδικούς που απαρτίζουν την ανάλυση των λημμάτων (π.χ. λήμμα, προφορά, παραδείγματα χρήσης).
3. "Έξυπνη Αναζήτηση" με χρήση κωδικο-λέξεων (π.χ. "περιέχεται", "αρχίζει από", "τελειώνει σε") για πιο ευρεία αναζήτηση στα περιεχόμενα του λεξικού.
4. Αναζήτηση αγνοώντας ή όχι τις διακρίσεις πεζών/κεφαλαίων χαρακτήρων καθώς και τονισμένων/ μη τονισμένων λέξεων.
5. Εμφάνιση όλων των πιθανών αποτελεσμάτων που πληρούν τα κριτήρια αναζήτησης που έχει θέσει ο χρήστης.

Επιπλέον, το ηλεκτρονικό λεξικό ΚΟΡΑΗΣ έχει ενσωματωμένα μια σειρά από γλωσσικά εργαλεία που μπορούν να φανούν πολύ χρήσιμα στους σπουδαστές της Ελληνικής και της Αγγλικής γλώσσας [7]:

6. Δύο Συνθέτες Ομιλίας από κείμενο (TTS-Synthesizers) για την ελληνική και την αγγλική γλώσσα, αντίστοιχα.
7. Δύο λημματοποιητές για την Ελληνική και την Αγγλική, αντίστοιχα.
8. Μεταγραφέα της γραφηματικής παράστασης μιας λέξης στην αντίστοιχη φωνητική.
9. Συλλαβιστή για την ελληνική γλώσσα.

3.1 Διεπαφή Ανθρώπου-Μηχανής

Η σχεδίαση της διεπαφής ανθρώπου-μηχανής έγινε με στόχο να παρέχει στο χρήστη εύκολους, γρήγορους, "έξυπνους" και προσαρμοζόμενους στις ανάγκες του τρόπους αναζήτησης της πληροφορίας. Για να πραγματοποιηθούν αυτά, όλα τα λήμματα του έντυπου λεξικού με τις αντίστοιχες αναλύσεις τους οργανώθηκαν σε βάση δεδομένων SQL. Έτσι έγινε δυνατή η γρήγορη και διεξοδική αναζήτηση πληροφοριών, όχι μόνο σύμφωνα με το συνήθη τρόπο (δηλαδή αναζήτηση λέξεων που αρχίζουν με τους χαρακτήρες που εισάγει ο χρήστης) αλλά και με πιό σύνθετους τρόπους (π.χ. αναζήτηση λέξεων που τελειώνουν, που περιέχουν ή απαρτίζονται εξολοκλήρου από τους συγκεκριμένους χαρακτήρες). Επιπλέον, επιτρέπει την αναζήτηση όχι μόνο στον κωδικό "ΛΗΜΜΑ" αλλά και σε οποιονδήποτε άλλο κωδικό που αποτελεί μέρος της ανάλυσης των λημμάτων του λεξικού. Έτσι, ο χρήστης μπορεί να αξιοποιήσει πλήρως τις πληροφορίες που περιέχονται στο λεξικό (π.χ. μπορεί να ανακτήσει όλες τις προφορές των λημμάτων, τα παραδείγματα, ή τα συνώνυμα του λεξικού που πληρούν τα κριτήρια αναζήτησης που θέτει).

Το περιβάλλον εργασίας στον υπολογιστή αποτελείται από συνολικά 4 επιφάνειες εργασίας: το παράθυρο εισαγωγής των κριτηρίων αναζήτησης (Search Panel), το παράθυρο εμφάνισης όλων των λημμάτων που περιέχονται στο λεξικό (Word List Panel), το παράθυρο εμφάνισης των αποτελεσμάτων που πληρούν τα κριτήρια αναζήτησης (Result Panel) και το παράθυρο Options Panel στο οποίο ο χρήστης επιλέγει τους κωδικούς στους οποίους θα γίνει η αναζήτηση, καθώς και τους κωδικούς οι οποίοι θα εμφανίζονται στο Result Panel. Επιπλέον, υπάρχει μια γραμμή με τις επιλογές του μενού (menu bar) και μια γραμμή εργαλείων (toolbar). Ο χρήστης έχει τη δυνατότητα αλλαγής τόσο του τρόπου εμφάνισης των αποτελεσμάτων (π.χ. χρώμα, γραμματοσειρά, κ.λ.π.) όσο και του χρώματος του φόντου της συγκεκριμένης εφαρμογής.

3.2 Συνθέτες Ομιλίας από κείμενο για την ελληνική και την σγγλική γλώσσα

Οι συνθέτες που έχουν ενσωματωθεί στο ΚΟΡΑΗΣ αναπτύχθηκαν στο ΕΕΤ με χρήση της πολυγλωσσικής πλατφόρμας σύνθεσης ομιλίας Festival [8], η οποία περιλαμβάνει διάφορες τεχνικές σύνθεσης (συνένωση διφώνων, επιλογή φωνητικών μονάδων, κ.λ.π.). Και οι δύο συνθέτες βασίζονται στη συνένωση διφώνων κυρίως για λόγους αξιοπιστίας του συστήματος καθώς και μικρού υπολογιστικού κόστους. Για την αγγλική γλώσσα, η σύνθεση πραγματοποιείται χρησιμοποιώντας την αμερικανική ανδρική φωνή (KAL) [8] με δίφωνα που έχουν προκύψει από τη βάση CMU Lexicon [9], ενώ για την ελληνική γλώσσα έχει χρησιμοποιηθεί η ελληνική γυναικεία φωνή (ΖΕΤΑ), με δίφωνα που προήλθαν από τη βάση WCL-1 [10].

Οι δύο συνθέτες επιτρέπουν την μετατροπή οποιουδήποτε πληκτρολογούμενου αγγλικού ή ελληνικού κειμένου σε ομιλία με καλή καταληπτότητα. Η χρήση τους είναι εύκολη, ενώ έχουν επιπρόσθετα τη δυνατότητα μεταβολής του ρυθμού (ταχύτητας) και της έντασης αναπαραγωγής της συνθετικής ομιλίας.

3.3 Λημματοποιητές για την ελληνική και την αγγλική γλώσσα

Με τον όρο λημματοποίηση αναφερόμαστε στην αναγωγή των διαφορετικών τύπων ενός λήμματος στο λήμμα που είναι καταχωρημένο στο λεξικό. Ο λημματοποιητής μπορεί να διευκολύνει πολύ τον σπουδαστή διότι είναι δύσκολο να πραγματοποιηθεί αναζήτηση σε κάποιο λεξικό μιας άγνωστης λέξης, χωρίς να είναι γνωστό το λήμμα από το οποίο αυτή προέρχεται.

Η λειτουργία των λημματοποιητών για την ελληνική και την αγγλική γλώσσα βασίζεται στο συνδυασμό ενός μοντέλου στοχαστικής-επαναληπτικής αφαίρεσης καταλήξεων και ταυτόχρονη εφαρμογή δύο μοντέλων ανίχνευσης ομοιότητας μεταξύ δύο λέξεων,

προκειμένου το πρόγραμμα να αποφανθεί σε ποια λήμματα του λεξικού μοιάζει περισσότερο η λέξη εισόδου. Συγκεκριμένα, τα δύο μοντέλα ανίχνευσης ομοιότητας μεταξύ λέξεων είναι η απόσταση Levenshstein [11] και το κριτήριο ομοιότητας του συντελεστή Dice [12]. Καθένα από τα δύο μοντέλα επιστρέφει ένα σύνολο λημμάτων του λεξικού που μοιάζουν με την λέξη εισόδου καθώς και μια τιμή για καθένα από αυτά τα λήμματα, ενδεικτική του βαθμού ομοιότητάς του με τη λέξη εισόδου. Έπειτα, τα δύο αυτά σύνολα λημμάτων ενοποιούνται σε ένα ενιαίο σετ, το οποίο ταξινομείται σύμφωνα με τον αλγόριθμο που περιγράφεται στη δημοσίευση [13], και το τελικό σύνολο προτεινόμενων λημμάτων εμφανίζεται στην οθόνη του χρήστη κατά φθίνουσα τάξη ομοιότητας.

Είναι σημαντικό να τονιστεί ότι οι λημματοποιητές αυτοί δεν χρησιμοποιούν γλωσσική πληροφορία (π.χ. κανόνες κλίσης, κλιτικούς πίνακες κ.τ.λ.) και για αυτό το λόγο λειτουργούν αρκετά ικανοποιητικά για περιπτώσεις ομαλών και “ελαφρά” ανώμαλων λέξεων (π.χ. ουσιαστικών, επιθέτων, μετοχών, ομαλών ρημάτων).

3.4 Μεταγραφέας Gr→Ph της ελληνικής γλώσσας

Ο μεταγραφέας Gr→Ph της νεοελληνικής γλώσσας αποδίδει τη φωνητική γραφή της λέξης που εισάγει ο χρήστης χρησιμοποιώντας τα σύμβολα του φωνητικού αλφαβήτου υπολογιστή CPA (Computer Phonetic Alphabet) (βλ. Κεφ. 2). Ο μεταγραφέας αυτός αναπτύχθηκε στο Εργαστήριο Ενσύρματης Τηλεπικοινωνίας στα πλαίσια του ερευνητικού-αναπτυξιακού προγράμματος ESPRIT 291/860 [14].

Η αυτόματη μετατροπή γραπτού κειμένου της νεοελληνικής γλώσσας στην αντίστοιχη φωνητική γραφή πραγματοποιείται χρησιμοποιώντας ένα σετ φωνητικών κανόνων που έχουν ενσωματωθεί στο λογισμικό FONPARS1, το οποίο αναπτύχθηκε στο Πανεπιστήμιο Nijmegen [15]. Για την πλειοψηφία των ελληνικών γραφημάτων υπάρχει μία-προς-μία αντιστοίχιση με τα φωνήματα (δηλαδή κάθε γράφημα αντιστοιχεί σε μοναδικό φώνημα), ωστόσο εμφανίζονται και πιο πολύπλοκες περιπτώσεις, όπου η φωνητική αναπαράσταση ενός γραφήματος εξαρτάται από το περιβάλλον στο οποίο βρίσκεται.

3.5 Συλλαβιστής για την ελληνική γλώσσα

Ο συλλαβιστής αποδίδει αυτόματα το συλλαβισμό (διαχωρισμό των γραπτών λέξεων σε συλλαβές) του κειμένου που εισάγει ο χρήστης. Ο συλλαβισμός πραγματοποιείται σε 2 στάδια. Αρχικά εφαρμόζεται στο κείμενο εισόδου μια απλουστευμένη (naïve) συνάρτηση τεμαχισμού σε συλλαβές, η οποία έχει ως κριτήριο την συμπερίληψη ενός μόνο φωνήεντος σε κάθε συλλαβή. Για παράδειγμα, η λέξη εισόδου «άλλωστε» αρχικά χωρίζεται στις συλλαβές «ά-λλω-στε». Έπειτα, το κείμενο που έχει προκύψει από το προηγούμενο στάδιο,

διορθώνεται από μια συνάρτηση αυστηρού συλλαβισμού λέξεων, η οποία βασίζεται στους γραμματικούς κανόνες που περιγράφονται στην Νεοελληνική Γραμματική του Μ. Τριανταφυλλίδη [16].

4 Συμπέρασμα

Το λεξικό ΚΟΡΑΗΣ, με μια σειρά καινοτόμων χαρακτηριστικών τόσο στην έντυπη, όσο και στην ηλεκτρονική του μορφή, μπορεί να αποτελέσει σημαντικό βοήθημα στους Έλληνες σπουδαστές και χρήστες της Αγγλικής γλώσσας. Ιδιαίτερα, εκτιμάται ότι θα είναι χρήσιμο σε αγγλόφωνους που σπουδάζουν την Ελληνική στηρίζοντας την εκμάθησή της σε εποχή που η παγκοσμιοποίηση απειλεί τις λιγότερο ομιλούμενες γλώσσες.

5 Αναφορές

[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] Ελληνο-Αγγλικό Λεξικό ΚΟΡΑΗΣ, Εκδόσεις Πανεπιστημίου Πατρών, Πάτρα 2008.

[4] Coutsogeorgopoulos H., Kokkinakis G., Dermatas E., “KORAIS: A Large Electronic Greek-English Dictionary with Spoken Pronunciation”, COMLEX 2000, pp.127-130.

[5] Dermatas E., Kokkinakis G., “LEXITHIRAS: Corpus Based Lexicography on PC”, TSD-98, Prague, 1998.

[6] Coutsogeorgopoulos H., Dermatas E., Kokkinakis G., “A Monolingual Electronic Dictionary for the Pronunciation and Usage of Modern Greek for Foreigners”, COMLEX 2000, pp.83-88.

[7] Lyras D., Kokkinakis G., Lazaridis A., Sgarbas K., Fakotakis N., “A Large Greek-English Dictionary with Incorporated Speech and Language Processing Tools”, Interspeech 2009, Brighton, U.K.

[8] 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.

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

[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] 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.

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

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

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

[16] Τριανταφυλλίδης Μ., "Νεοελληνική Γραμματική", ΟΕΔΒ, Αθήνα, 1977.

Ευχαριστίες

Οι συγγραφείς ευχαριστούν θερμά όλους όσους εργάστηκαν στη δημιουργία του λεξικού ΚΟΡΑΗΣ, την ΓΓΕΤ για την υποστήριξη του έργου στην αρχική του φάση και το Ίδρυμα ΣΤΑΥΡΟΣ ΝΙΑΡΧΟΣ για την ανάληψη της δαπάνης έκδοσης.

Γεώργιος Κοκκινάκης
Ομ. Καθηγητής
gkokkin@wcl.ee.upatras.gr

Ελένη Κουτσογεωργοπούλου
Ερευνήτρια
coutsoge@wcl.ee.upatras.gr

Δημήτριος Λύρας
Υποψήφιος Διδάκτωρ
dimlyras@upatras.gr

Κυριάκος Σγάρμπας
Επικ. Καθηγητής
sgarbas@upatras.gr

Εργαστήριο Ενσύρματης Τηλεπικοινωνίας
Τμήμα Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών
Πανεπιστήμιο Πατρών
Πάτρα, Τ.Κ.26500
http://www.wcl.ece.upatras.gr