



# ΓΛΩΣΣΙΚΟΙ ΠΟΡΟΙ & ΤΕΧΝΟΛΟΓΙΕΣ: Η ΣΗΜΕΡΙΝΗ ΕΛΛΗΝΙΚΗ ΠΡΑΓΜΑΤΙΚΟΤΗΤΑ

Ημερίδα παρουσίασης CLARIN-EL

1/10/2010

Πένυ Λαμπροπούλου

Ινστιτούτο Επεξεργασίας Λόγου / Ε.Κ. "Αθηνά"

# ΧΑΡΤΟΓΡΑΦΗΣΗ ΧΩΡΟΥ ΓΤ ΓΙΑ ΚΑΕ

---

- Στο πλαίσιο της μελέτης σκοπιμότητας της προπαρασκευαστικής φάσης
- Δύο παράλληλες έρευνες
  - Καταγραφή ΓΠΤ, με σκοπό τη διερεύνηση της δυνατότητας κατασκευής της ΕΥ
  - Μελέτη για τη χρήση της ΓΤ στις ΚΑΕ, με σκοπό την καταγραφή των απαιτήσεων των χρηστών για την ΕΥ
- Διαδικτυακά: [www.clarin.gr/survey](http://www.clarin.gr/survey)
- Με ερωτηματολόγια και συνεντεύξεις
- Μαζική αποστολή σε στοχευμένους αποδέκτες (πανεπιστήμια, ερευνητικά κέντρα, ιδιώτες) & σχετικές ηλεκτρονικές λίστες

# ΚΑΤΑΓΡΑΦΗ ΓΠΤ

---

- Δομημένη σε τέσσερα ερωτηματολόγια ανάλογα με την κατηγορία πόρου
  - Πόροι αναφοράς
  - Πόροι πρωτογενούς υλικού
  - Πόροι επεξεργασμένου υλικού
  - Εργαλεία / Εφαρμογές ΓΤ
- Παρουσίαση και ανάλυση των μέχρι στιγμής αποτελεσμάτων για κάθε κατηγορία πόρου
- Γενικά συμπεράσματα

# ΠΟΡΟΙ ΑΝΑΦΟΡΑΣ

---

- περιέχουν **κωδικοποιημένη γνώση** για τη γλώσσα και μπορούν να χρησιμοποιηθούν για την καλύτερη **οργάνωση, επεξεργασία και μελέτη των πόρων περιεχομένου**
- υποκατηγορίες
  - Λεξικοί/Εννοιολογικοί πόροι (π.χ. λεξικά, γλωσσάρια, ορολογικοί πόροι, οντολογίες)
  - Πόροι γλωσσικής περιγραφής (π.χ. γραμματικές, γλωσσικά μοντέλα, εκπαιδευτικό λογισμικό)

# ΑΠΟΤΕΛΕΣΜΑΤΑ ΕΡΕΥΝΑΣ - ΠΑ

---

- 23 λεξικοί/εννοιολογικοί πόροι & 1 εκπαιδευτικό λογισμικό – κυρίως:
  - υπολογιστικά λεξικά γενικής γλώσσας
  - ορολογικά λεξικά
  - οντολογίες
  - λίστες λέξεων (π.χ. ακρωνυμίων, με ποσοτικές ιδιότητες)
- **Περιεχόμενο λεξικών πόρων:** κυρίως μορφολογική και σημασιολογική αλλά και πολυμεσική πληροφορία
- **Μέγεθος:** κυρίως 5000 – 20000 εγγραφές, αλλά και 2 λεξικοί πόροι με >65000 εγγραφές
- **Γλώσσες:** Μονόγλωσσοι (κυρίως της ΝΕ) αλλά και δί-/πολύγλωσσοι (ειδικών θεματικών πεδίων αλλά και γενικής γλώσσας)

# ΠΟΡΟΙ ΠΡΩΤΟΓΕΝΟΥΣ ΥΛΙΚΟΥ

---

- κάθε πόρος **ψηφιακού / ψηφιοποιημένου λόγου** (π.χ. ψηφιοποιημένα βιβλία, άρθρα από εφημερίδες στο διαδίκτυο, τηλεοπτικές εκπομπές, ψηφιακές πολιτιστικές συλλογές κτλ.)
- κάθε **μέσο**: κείμενο, ήχος, εικόνα, βίντεο
- **χωρίς να έχουν γίνει αντικείμενο επεξεργασίας** με εργαλεία ΓΤ εκτός:
  - ψηφιοποίησης
  - κανονικοποίησης μορφής
  - εξωτερικής τεκμηρίωσης πόρου ή/και αντικειμένων που τον απαρτίζουν

# ΑΠΟΤΕΛΕΣΜΑΤΑ ΕΡΕΥΝΑΣ - ΠΠΥ

---

- 28 πόροι
  - σώματα κειμένων, συλλογές προφορικού λόγου & πολυμεσικοί πόροι για γλωσσική μελέτη ή/και ανάπτυξη εφαρμογών ΓΤ
  - ψηφιοποιημένες πολυμεσικές συλλογές από αρχειακούς οργανισμούς και βιβλιοθήκες
  - συλλογές πρωτότυπων διαδικτυακών κειμένων
- **Μέγεθος**: ανάλογα με τον σκοπό χρήσης του πόρου & την πηγή προέλευσης
- **Γλώσσες**: κυρίως ΝΕ (αλλά και παλαιότερες μορφές της ελληνικής) & δίγλωσσοι πόροι
- **Τεκμηρίωση**: σε επίπεδο πόρου και σε επίπεδο αντικειμένων, βιβλιογραφικού τύπου & θεματική ταξινόμηση
- **Μειονέκτημα**: ψηφιοποίηση σε επίπεδο "εικόνας" αλλά αξιοποίηση τεκμηρίωσης

# ΠΟΡΟΙ ΕΠΕΞΕΡΓΑΣΜΕΝΟΥ ΥΛΙΚΟΥ

---

- πόροι περιεχομένου που έχουν γίνει **αντικείμενο επεξεργασίας με εργαλεία ΓΤ**
- "επεξεργασία"
  - γλωσσολογική επισημείωση (π.χ. μέρος του λόγου, λήμμα, μορφοσυντακτική πληροφορία, σημασιολογική κατηγορία κτλ.)
  - πολυμεσική επισημείωση (π.χ. μεταγραφή προφορικού λόγου, κινήσεις, εκφράσεις προσώπου, αναγνώριση ομιλητή κτλ.)



# ΑΠΟΤΕΛΕΣΜΑΤΑ ΕΡΕΥΝΑΣ - ΠΕΥ

---

- 37 πόροι
  - σώματα κειμένων με γλωσσολογική επισημείωση
  - πολυμεσικοί πόροι με διάφορες επισημειώσεις
  - πολυμεσικοί πόροι με διάφορες επισημειώσεις & γλωσσολογική επισημείωση των κειμενικών στοιχείων
- **Είδη επισημειώσεων:**
  - γλωσσολογικές: κυρίως μορφοσυντακτική πληροφορία & λήμμα αλλά και συντακτικές & σημασιολογικές
  - πολυτροπικές: μεταγραφή, επισημείωση ομιλητή, κινήσεων, χειρονομιών, σημασιολογικών σχέσεων μεταξύ τροπικοτήτων
- **Τρόπος επισημείωσης:**
  - διαδικασία: ημι-αυτόματη για τα κατώτερα επίπεδα ανάλυσης & χειρωνακτική για τα ανώτερα
  - ανάπτυξη εργαλείων αυτόματης επισημείωσης & χρήση κυρίως ελευθερών εργαλείων για τη χειρωνακτική επισημείωση
  - υιοθέτηση διεθνών προτύπων/καλών πρακτικών για τα σχήματα μεταδεδομένων επισημείωσης

# ΕΦΑΡΜΟΓΕΣ ΓΤ

---

- εργαλεία και ολοκληρωμένες εφαρμογές που επιτελούν
  - **γλωσσική επεξεργασία** (π.χ. στοίχιση πολύγλωσσων κειμένων, μορφολογική επισημείωση, λημματοποίηση, συντακτική ανάλυση, εξόρυξη γνώσης κτλ.),
  - **παρουσίαση/προβολή των δεδομένων** (π.χ. ολοκληρωμένα περιβάλλοντα προβολής κειμένων, συλλογών πολυμεσικών δεδομένων κτλ.)

# ΑΠΟΤΕΛΕΣΜΑΤΑ - ΕΦΑΡΜΟΓΕΣ ΓΤ

---

- 51 εφαρμογές/εργαλεία
  - κυρίως εφαρμογές που χειρίζονται κείμενα, και σε μικρότερο βαθμό ήχο/σήμα και πολυμεσικούς πόρους
  - υποστηρικτικές εφαρμογές για ανάπτυξη εργαλείων ΓΤ
- **Χρήστες εργαλείων:**
  - βασικά εργαλεία που απευθύνονται στους ειδικούς (π.χ. εργαλεία μορφοσυντακτικής & σημασιολογικής επισημείωσης)
  - εργαλεία που δεν απαιτούν ειδικές γνώσεις και μπορούν να χρησιμοποιηθούν από ερευνητές ΚΑΕ για τις εργασίες τους (π.χ. εξαγωγή ορολογίας, εξαγωγή στατιστικών στοιχείων, αναγνώριση ονοματικών οντοτήτων κτλ.)
- **Γλώσσες:** κυρίως ΝΕ & αγγλική αλλά και εργαλεία ανεξάρτητα από γλώσσες

# ΣΥΜΠΕΡΑΣΜΑΤΑ – ΠΑΡΑΤΗΡΗΣΕΙΣ (1)

---

- Η καταγραφή συνεχίζεται
  - μέχρι τώρα:
    - καταχώρηση κυρίως από φορείς ΓΤ
    - λείπουν: αποτελέσματα έργων ψηφιοποίησης, ορολογικά λεξικά, συλλογές και εργαλεία για την ΑΕ, ...
    - δεν υπάρχουν, δεν καταχωρήθηκαν ή δεν διατίθενται;
  - στο μέλλον:
    - ανάρτηση των αποτελεσμάτων στον ιστότοπο
    - καταλογογράφηση των πόρων & εργαλείων
    - συγκομιδή μεταδεδομένων περιγραφής πόρων από το διαδίκτυο
- Διαθεσιμότητα:
  - δεν δηλώνεται για πολλούς ΓΠΤ
  - κυρίως για έρευνα από ερευνητικούς/ακαδημαϊκούς οργανισμούς ή/και για εκπαίδευση
  - πιο σαφής η κατάσταση στους ΠΠΥ

## ΣΥΜΠΕΡΑΣΜΑΤΑ – ΠΑΡΑΤΗΡΗΣΕΙΣ (2)

---

- **Μορφότυπα κωδικοποίησης:** παρατηρείται ποικιλία αλλά κυρίως εμφανίζονται τα ευρέως διαδεδομένα
- **Πρότυπα κωδικοποίησης:** υιοθέτηση διεθνών προτύπων και καλών πρακτικών, κυρίως για τις επισημειώσεις
- **Έλεγχος/αξιολόγηση:** έχει γίνει για πολλούς πόρους, χειρωνακτικά από ειδικούς στο σύνολο του υλικού

## ΣΥΜΠΕΡΑΣΜΑΤΑ – ΠΑΡΑΤΗΡΗΣΕΙΣ (3)

---

- Ενθαρρυντικό:
  - υπάρχουν βασικά εργαλεία & πόροι για ανάπτυξη ΓΤ αλλά και για υποστήριξη ΚΑΕ
  - πόροι για την Ελληνική Νοηματική Γλώσσα
  - πολυμεσικοί πόροι
  - σημασιολογικοί λεξικοί πόροι & σημασιολογική επισημείωση
- Απαιτούνται:
  - περαιτέρω επέκταση & ανάπτυξη νέων πόρων & εργαλείων (π.χ. ορολογικών θησαυρών, ελεγχόμενων λεξιλογίων, οντολογιών, εργαλείων για παλαιότερες μορφές και διαλέκτους ελληνικής κτλ.)
  - διασφάλιση σημασιολογικής συμβατότητας μεταξύ πόρων και εργαλείων και πόρων μεταξύ τους

# ΧΡΗΣΗ ΤΩΝ ΓΠΤ ΣΤΙΣ ΚΑΕ

---

- Κυρίως από Γλωσσολογία & ΓΤ & Παιδαγωγικά
- Διαδεδομένες πρακτικές:
  - Χρήση μηχανών αναζήτησης & επίσκεψη σε συγκεκριμένους ιστότοπους (ειδικούς του τομέα & ψηφιακές βιβλιοθήκες)
  - Χρήση λεξικών πόρων για κατανόηση σημασίας & τυποποίηση ορολογίας
  - Μελέτη δευτερογενούς υλικού σε ψηφιακή μορφή
  - Χρήση πρωτογενούς ψηφιακού υλικού στο πλαίσιο έρευνας ή/και ανάπτυξης εκπαιδευτικού υλικού
  - Σε μικρότερο βαθμό, κατασκευή πρωτογενούς ψηφιακού υλικού – κυρίως με ψηφιοποίηση, σπανιότερα από υπάρχον ψηφιακό υλικό
  - Μικρή διείσδυση ΓΤ στους μη σχετικούς με τη γλωσσική μελέτη

# ΑΝΑΓΚΕΣ ΕΡΕΥΝΗΤΩΝ ΚΑΕ

---

- Επιθυμητές λειτουργίες υποδομής
  - μητρώο πόρων & εργαλείων για διευκόλυνση αναζήτησης
  - τεχνική υποστήριξη
  - νομική υποστήριξη
  - χρήση εργαλείων ΓΤ σε δικό τους υλικό
- Επιθυμητοί ΓΠΤ
  - Λεξικοί/Εννοιολογικοί πόροι
  - Εργαλεία ανάπτυξης & εμπλουτισμού πόρων
  - Επισημειωμένοι πόροι
- Τι προσφέρουν οι ίδιοι
  - Ερευνητικά αποτελέσματα
  - Παρατηρήσεις/Σχόλια
  - Όχι: νέους/εμπλουτισμένους πόρους
- Επιμόρφωση: *"Σε κάποιες από τις ερωτήσεις του ερωτηματολογίου δεν γνώριζα πολλές από τις επιλογές της απάντησης. Κατάλαβα λοιπόν ότι **υπάρχουν πολλές χρήσεις τις οποίες δεν γνωρίζω** (αλλά θα ήθελα να μάθω)."*



# ΤΙ ΧΡΕΙΑΖΕΤΑΙ ΓΙΑ ΤΗΝ ΥΛΟΠΟΙΗΣΗ ΤΗΣ ΕΥ;

---

## ○ **νομικά θέματα**

- πρότυπες άδειες πρόσβασης και χρήσης στους ΓΠΤ
- επίλυση προβλημάτων πνευματικής ιδιοκτησίας πρωτογενούς υλικού
- για έρευνα & εκπαίδευση

## ○ **τεχνικά θέματα**

- καταγραφή και περιγραφή των πόρων & εργαλείων
- διαδικτυακές υπηρεσίες
- αρχιτεκτονική συστήματος
- συντακτική & σημασιολογική συμβατότητα

## ○ **οργανωτικά θέματα**

- δικαιώματα & υποχρεώσεις μελών δικτύου

---

Ευχαριστώ πολύ!

