9ο Συνέδριο «Ελληνική Γλώσσα και Ορολογία», Αθήνα, 7-9 Νοεμβρίου 2013

Κατασκευή βάσης δεδομένων ελληνικών ακρωνυμίων σε ελληνικά νομικά κείμενα

Τσιμπούρης Χαράλαμπος

Υπ. Διδάκτορας
Εργαστήριο Ενσύρματης Τηλ.
Τμήμα Ηλ. Μηχ. & Τεχ. Υπολ.
Πανεπιστήμιο Πατρών

xtsimpouris [at] upatras [dot] gr

Κυριάκος Σγάρμπας

Επίκουρος Καθηγητής
Εργαστήριο Ενσύρματης Τηλ.
Τμήμα Ηλ. Μηχ. & Τεχ. Υπολ.
Πανεπιστήμιο Πατρών

sgarbas [at] upatras [dot] gr

Υπό την αιγίδα της Επιτροπής Ερευνών του Πανεπιστημίου Πατρών, στα πλαίσια της βασικής έρευνας “Καραθεοδωρής”

Δομή Παρουσίασης

- Εισαγωγή
  - Ανάλυση & αντιμετώπιση προβλήματος
- Μελέτη
- Βάση δεδομένων e-Themis
- Μεθοδολογία
- Αποτελέσματα
- Συμπεράσματα

Ανάλυση & αντιμετώπιση προβλήματος 1..

Βασικές κατηγορίες εκτεταμένης χρήσης ακρωνυμίων:

- Τεχνικά εγχειρίδια (RFC, εγχειρίδια χρήσεως, κ.α.)
- Νομικά κείμενα
- Ιατρικά κείμενα

Επεξήγηση ορισμού:

- Περιεχόμενα, Ευρετήριο όρων (διατριβής, ή άλλου ακαδημαϊκού χαρακτήρα)
- Μέσα στην πρόταση, κατά την πρώτη αναφορά

Στατιστικά:

- 1% λεξικό Κοραής (857 λήμματα σε σύνολο 81.000)
- ~0.6% κείμενο γενικού περιεχομένου αυξάνεται σε 1.5% στα νομικά κείμενα

Ανάλυση & αντιμετώπιση προβλήματος ..2

Προβλήματα που πρέπει να αντιμετωπιστούν για αυτόματη εξαγωγή:

- Απαραίτητη υποδομή
  - Βάση εκπαίδευσης & ελέγχου
- Εκπαίδευση με αλγόριθμους μάθησης
- Προαιρετικά
  - *"Διαχείριση"* ποικιλομορφίας γλώσσας
    - Εξαγωγή μερών λόγου και σύνταξης

Μελέτη

Πώς
- Χωρίς επίσημη τυποποίηση (ΕΛΟΤ, Ιούνιος 2013)
- Κανονικές εκφράσεις
- Αυτόματα πεπερασμένων καταστάσεων
- Ταξινομητές Naive Bayes

Δημιουργείται;
Αναφέρεται την πρώτη φορά;
Αναφέρεται στην συνέχεια;
Αναγνωρίζεται;

Πού

Καταχωρείται;

Ανεπίσημες ιστοσελίδες με λίστες Ακρωνυμίων όπως
- asas.gr
- lexeis.gr

Πότε

“Παύει να υπάρχει”;

Ζει εφ' όρου ζωής στην επίσημη λίστα
ή πρέπει να συνοδεύεται και από κάποιο
χρονικό προσδιορισμό;

Βάση δεδομένων e-Themis

e-Themis.gov.gr

- Υποστηρίζεται από το Τμήμα Λογιστικής και Γραμματειακής Υποστήριξης του Υπουργείου εξωτερικών
  - Τελευταία προσπέλαση, Ιούνιος 2012
  - Μη διαθέσιμο πλέον

- Δωρεάν πρόσβαση σε 40 θεματικές ενότητες κώδικα νομοθεσίας
  - Συνταγματικό
  - Διοικητικό
  - Νομαρχίες
  - ...

- 105 Έγγραφα σε μορφή Microsoft Word
  - Κωδικοποίηση UTF-8
  - 410 Mbytes
  - ~ 32 εκ. λεκτικές μονάδες (tokens)

Μεθοδολογία, στατιστική ανάλυση..

Επιλογή αρχικού σώματος κειμένων

- Στατιστική ανάλυση με χρήση κανονικής έκφρασης

$$R = ((([\backslash A - \Omega]\{1,4\}\backslash.)?)+)([\hat{}\, A - \Omega][\hat{}.][\hat{}\ ])$$

- Ταξινόμηση κειμένων ως προς την πυκνότητα αποτελεσμάτων
- Επιλογή τριών πρώτων κειμένων

Α/Α	Λεκτ. Μονάδες	Ακρ.	Παράδειγμα ακρ.	Αρχείο μελέτης
1	**1672173**	**249**	**Υ.Ε.Ο., Υ.Ε.Ε.Π.Π.**	***13β_βιομηχανία_ανάπτυξη_αλιεία***
2	**1048518**	**154**	**Β.Π., Υ.Ε.Α., Υ.Α.**	***37_πολεμικό ναυτικό***
3	**2367870**	**288**	**Ο.Ε.Σ.Β., Γ.Α.Κ.**	***32_εκπαιδευτική_νομοθεσία_αθλητισμός***
4	1560683	185	Υ.Ν., Σ.Ν.Δ., Α.Ν.Σ.	*37α_πολεμικό_ναυτικό*
5	1674019	198	Γ.Ε.Α., Ε.Β.Α.	*38_πολεμική_αεροπορία_πολιτική_αεροπορία*

Μεθοδολογία, εξαγωγή ακρωνυμίων..

Ημιαυτόματη διαδικασία εξαγωγής αποτελεσμάτων

- 533 Διαφορετικά ακρωνύμια
  - ➔ Διαφορετικά == Εφόσον έχουν διαφορετικό ορισμό

- 339 Με επεξήγηση ορισμού (~63%)
  - ➔ Εφόσον αναφερόταν μέσα στην ίδια πρόταση

Εξαγωγή χαρακτηριστικών (features):

- Επεξήγηση σε επόμενη διαφάνεια

Εξαγωγή επισημείωσης:

- Ενσωμάτωση επισημείωσης XML στα κείμενα μελέτης

Μεθοδολογία, εξαγωγή χαρακτηριστικών & δημιουργία λεξικού/corpus

Εξαγωγή των παρακάτω χαρακτηριστικών:

- Ακρωνύμιο (σε *κανονική* μορφή)
- Ακρωνύμιο (όπως ανιχνεύτηκε)
- Επεξήγηση/Ορισμός (εφόσον έχει δηλωθεί)
- Θέση μέσα στο αρχείο
- Παράθυρο 200 χαρακτήρων πριν την εμφάνιση ακρωνυμίου
- Παράθυρο 200 χαρακτήρων μετά την εμφάνιση ακρωνυμίου
- 0/1 αν η επεξήγηση βρίσκεται εντός του παραθύρου χαρακτήρων *πριν* το ακρωνύμιο
- 0/1 αν η επεξήγηση βρίσκεται εντός του παραθύρου χαρακτήρων *μετά* το ακρωνύμιο
- 0/1 αν το ακρωνύμιο βρίσκεται μέσα σε παρενθέσεις

Αποτελέσματα, προβλήματα αναγνώρισης ακρωνυμίων

Αδυναμία χρήσης “απλοϊκής” κανονικής έκφρασης, τύπου

“([Α-Ω].[ ])+” ή “([Α-Ω])+”

Δηλαδή, κεφαλαία γράμματα που ακολουθούνται από τελεία ή κενό

για αναγνώριση ακρωνυμίων.

- Ανάμεικτη χρήση πεζών & κεφαλαίων χαρακτήρων
- Λέξεις/Προτάσεις με μόνο κεφαλαία γράμματα (Τίτλοι, κ.α.)
- Υπερβολική χρήση σημείων στίξης τελείας & παύλας, και πιθανού κενού -> προβλήματα στον αυτόματο διαχωριστή προτάσεων (sentence tokeniser) και λέξεων (word tokeniser)
- ✓ Ελάχιστη χρήση αριθμών

Αποτελέσματα, προβλήματα αναγνώρισης ορισμού (1/3)

- Διαφορετική κλίση ή τρόπος γραφής

Ακρωνύμιο: *Α.Ε.*
Επεξήγηση, γραφή 1: Ανώνυμος Εταιρ**εία**
Επεξήγηση, γραφή 2: Ανωνύμων Εταιρ**ειών**
Επεξήγηση, γραφή 3: Ανωνύμων Εταιρ**ιών**

- Δυσκολία συσχέτισης μεταξύ γραμμάτων ακρωνυμίου και λέξεων επεξήγησης

Ακρωνύμιο: *ΚΥΣΕΠ*
Επεξήγηση: ***Κεντρικόν Υπηρεσιακόν και Πειθαρχικόν Συμβούλιον Εποπτικού Προσωπικού Μέσης Τεχνικής και Επαγγελματικής Εκπαιδεύσεως***

- Επαναλαμβανόμενη χρήση λέξης

Ακρωνύμιο: *Ε.Ε.Κ.*
Επεξήγηση, γραφή 1: **επαγγελματική εκπαίδευση και κατάρτιση**
Επεξήγηση, γραφή 2: **επαγγελματική εκπαίδευση και επαγγελματική κατάρτιση**

Αποτελέσματα, προβλήματα αναγνώρισης ορισμού (2/3)

- Η επεξήγηση βασίζεται σε άλλες συντομεύσεις

  Ακρωνύμιο: *ΥΠΕΠΘ*
  Επεξήγηση, γραφή 1: **Υπουργείο Εθνικής Παιδείας και Θρησκευμάτων**
  Επεξήγηση, γραφή 2: **υπ. εθν. παιδείας και θρησκευμάτων**

- Ασάφεια στον επίσημο τίτλο δημόσιας υπηρεσίας

  Ακρωνύμιο: *Ε.Σ.Α.Π.*
  Επεξήγηση: **Εθνικό Συμβούλιο Ανώτατης Παιδείας**

  Ακρωνύμιο 2: *Σ.Α.Π.*
  Επεξήγηση: **Συμβούλιο Ανώτατης Παιδείας**

  Ακρωνύμιο 3: *Ε.Σ.Π.*
  Επεξήγηση: **Εθνικό Συμβούλιο Παιδείας**

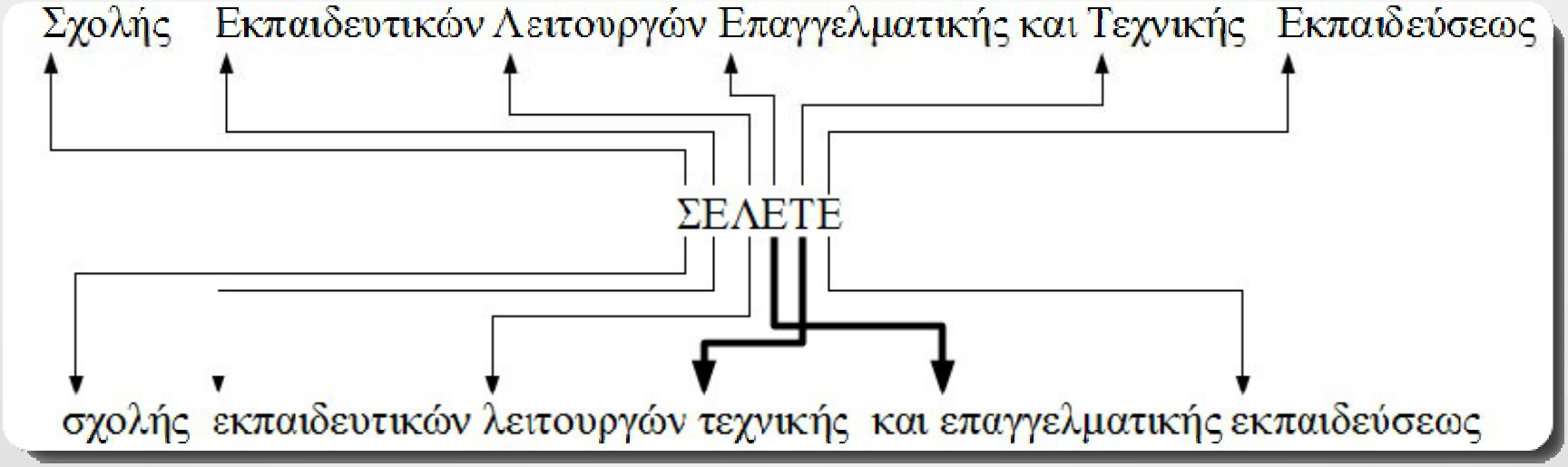
Αποτελέσματα, προβλήματα αναγνώρισης ορισμού (3/3)

- Αλλαγή σειράς λέξεων

Ακρωνύμιο: *ΣΕΛΕΤΕ*

Επεξήγηση, γραφή 1:
Σχολής Εκπαιδευτικών Λειτουργών Επαγγελματικής και Τεχνικής Εκπαιδεύσεως

Επεξήγηση, γραφή 2:
σχολής εκπαιδευτικών λειτουργών τεχνικής και επαγγελματικής εκπαιδεύσεως



Συμπεράσματα & Μελλοντικές κατευθύνσεις

- Σημαντική έλλειψη επίσημης τυποποίησης
  - Δημιουργίας ακρωνυμίων
  - Αναφοράς ορισμού
- Έλλειψη επίσημης καταγραφής σε ενιαία βάση δεδομένων
- Έλλειψη ενημέρωσης κοινού
  (*δικηγόροι, νομοτεχνικοί & δημ. υπάλληλοι, δημοσιογράφοι, πολίτες*)
- ✓ Διαδικτυακή έκδοση (υπό κατασκευή)
  - ✓ Αυτόματη αναζήτηση ακρωνυμίων
  - ✓ Σύνδεση με αντίστοιχο ορισμό

Ερωτήσεις

Σας ευχαριστώ για την προσοχή σας