ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

Γλωσσικές Βάσεις Δεδομένων Αιχμής: Προϋποθέσεις, Προαπαιτούμενα, Προδιαγραφές, Προοπτικές

Δημήτρης Παπαζαχαρίου
Πανεπιστήμιο Πατρών

ΕΡΕΥΝΗΤΙΚΑ ΣΕΜΙΝΑΡΙΑ ΤΟΜΕΑ ΓΛΩΣΣΟΛΟΓΙΑΣ ΕΚΠΑ 2021 - 2022

Γλωσσικές Βάσεις Δεδομένων: Τι είναι

Λογισμικά εργαλεία που επιτρέπουν την αποτελεσματική πλοήγηση και αναζήτηση μέσα σε μεγάλο όγκο γλωσσικών δεδομένων.

Γλωσσικές Βάσεις Δεδομένων: Τι είναι

Λογισμικά εργαλεία που επιτρέπουν την αποτελεσματική πλοήγηση και αναζήτηση μέσα σε μεγάλο όγκο γλωσσικών δεδομένων.

Αποτελούνται από τρεις πυλώνες:

Γλωσσικές Βάσεις Δεδομένων: Τι είναι

Λογισμικά εργαλεία που επιτρέπουν την αποτελεσματική πλοήγηση και αναζήτηση μέσα σε μεγάλο όγκο γλωσσικών δεδομένων.

Αποτελούνται από τρεις πυλώνες:

Δεδομένα

Γλωσσικές Βάσεις Δεδομένων: Τι είναι

Λογισμικά εργαλεία που επιτρέπουν την αποτελεσματική πλοήγηση και αναζήτηση μέσα σε μεγάλο όγκο γλωσσικών δεδομένων.

Αποτελούνται από τρεις πυλώνες:

Δεδομένα

Μεταδεδομένα

Γλωσσικές Βάσεις Δεδομένων: Τι είναι

Λογισμικά εργαλεία που επιτρέπουν την αποτελεσματική πλοήγηση και αναζήτηση μέσα σε μεγάλο όγκο γλωσσικών δεδομένων.

Αποτελούνται από τρεις πυλώνες:

Δεδομένα

Μεταδεδομένα

Εργαλεία αναζήτησης συνδυάζοντας δεδομένα και μεταδεδομένα

Γλωσσικές Βάσεις Δεδομένων: Προϊστορία

Προϊόν της ψηφιακής εποχής και της επανάστασης των υπολογιστών

Γλωσσικές Βάσεις Δεδομένων: Προϊστορία

Προϊόν της ψηφιακής εποχής και της επανάστασης των υπολογιστών

Εφικτές εξαιτίας

- της απόλυτης ταύτισης ανάμεσα στα πρωτότυπα δεδομένα και στα πιθανά πολλαπλά ψηφιακά αντίγραφα τους

Γλωσσικές Βάσεις Δεδομένων: Προϋποθέσεις

Προϊόν της ψηφιακής εποχής και της επανάστασης των υπολογιστών

Εφικτές εξαιτίας

- της απόλυτης ταύτισης ανάμεσα στα πρωτότυπα δεδομένα και στα πιθανά πολλαπλά ψηφιακά αντίγραφα τους
- της ταχύτητας δημιουργίας των ψηφιακών αντιγράφων

Γλωσσικές Βάσεις Δεδομένων: Προϋποθέσεις

Προϊόν της ψηφιακής εποχής και της επανάστασης των υπολογιστών

Εφικτές εξαιτίας

- της απόλυτης ταύτισης ανάμεσα στα πρωτότυπα δεδομένα και στα πιθανά πολλαπλά ψηφιακά αντίγραφα τους
- της ταχύτητας δημιουργίας των ψηφιακών αντιγράφων
- της ταχύτητας διαχείρισης τους με τη βοήθεια λογισμικών σε έναν υπολογιστή

Γλωσσικές Βάσεις Δεδομένων: Προϋποθέσεις

Πέρα από τις τεχνολογικές προϋποθέσεις,

Η επιθυμία αυτών που έχουν Σώματα Κειμένων να τα μοιραστούν με την ακαδημαϊκή κοινότητα, και όχι μόνο (ΚΑΘΟΡΙΣΤΙΚΗ ΠΡΟΫΠΟΘΕΣΗ)

Γλωσσικές Βάσεις Δεδομένων: Προϋποθέσεις

Πέρα από τις τεχνολογικές προϋποθέσεις,

Η επιθυμία αυτών που έχουν Σώματα Κειμένων να τα μοιραστούν με την ακαδημαϊκή κοινότητα, και όχι μόνο (ΚΑΘΟΡΙΣΤΙΚΗ ΠΡΟΫΠΟΘΕΣΗ)

(Διαφορετική πρόσληψη μεταξύ γενεών / διαφορετική αντίληψη «εκμετάλλευσης» των δεδομένων)

Γλωσσικές Βάσεις Δεδομένων: Προϋποθέσεις

Πέρα από τις τεχνολογικές προϋποθέσεις,

Η επιθυμία αυτών που έχουν Σώματα Κειμένων να τα μοιραστούν με την ακαδημαϊκή κοινότητα, (ΚΑΘΟΡΙΣΤΙΚΗ ΠΡΟΫΠΟΘΕΣΗ)

Για να γίνει κάτι τέτοιο εφικτό, αυτονόητη προϋπόθεση είναι η διασφάλιση των πνευματικών δικαιωμάτων αυτών που έχουν τα Σώματα Κειμένων

Γλωσσικές Βάσεις Δεδομένων: Προαπαιτούμενα

Ύπαρξη Σωμάτων Κειμένων (όχι απλά δεδομένα, αλλά συγκεντρωμένα και κωδικοποιημένα με συγκεκριμένες προδιαγραφές)

Γλωσσικές Βάσεις Δεδομένων: Προαπαιτούμενα

Ύπαρξη Σωμάτων Κειμένων (όχι απλά δεδομένα, αλλά συγκεντρωμένα και κωδικοποιημένα με συγκεκριμένες προδιαγραφές)

Συστηματική και συνεχής τεχνική υποστήριξη του λογισμικού από ομάδα προγραμματιστών

Γλωσσικές Βάσεις Δεδομένων: Προαπαιτούμενα

Ύπαρξη Σωμάτων Κειμένων (όχι απλά δεδομένα, αλλά συγκεντρωμένα και κωδικοποιημένα με συγκεκριμένες προδιαγραφές)

Συστηματική και συνεχής τεχνική υποστήριξη του λογισμικού από ομάδα προγραμματιστών

(Συνηθισμένη ‘παιδική αρρώστια’ διαφόρων Βάσεων Δεδομένων της δεκαετίας του 2000, η υποστήριξη να διαρκεί όσο και η χρηματοδότηση ενός ερευνητικού προγράμματος)

Γλωσσικές Βάσεις Δεδομένων: Προαπαιτούμενα

Ύπαρξη Σωμάτων Κειμένων (όχι απλά δεδομένα, αλλά συγκεντρωμένα και κωδικοποιημένα με συγκεκριμένες προδιαγραφές)

Συστηματική και συνεχής τεχνική υποστήριξη του λογισμικού από ομάδα προγραμματιστών

Σεβασμός του νομικού πλαισίου για τη διαχείριση προσωπικών δεδομένων, και απόκτηση ειδικών αδειών, ειδικά για ΒΔ που επιτρέπουν διαδικτυακή πρόσβαση

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΔΕΔΟΜΕΝΑ

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΔΕΔΟΜΕΝΑ

(Πρωτογενή & Επεξεργασμένα)

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΔΕΔΟΜΕΝΑ

(Πρωτογενή & Επεξεργασμένα)

Πρωτογενή = ψηφιακή εκδοχή των πρωτογενών δεδομένων

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΔΕΔΟΜΕΝΑ

(Πρωτογενή & Επεξεργασμένα)

Πρωτογενή = ψηφιακή εκδοχή των πρωτογενών δεδομένων

ΔΙΑΦΟΡΩΝ ΕΙΔΩΝ: Κείμενο, ήχος, εικόνα = Πολυτροπική Βάση Δεδομένων

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΠΡΩΤΟΓΕΝΗ ΔΕΔΟΜΕΝΑ

Κείμενο = jpeg ή pdf format, ειδικά αν το πρωτότυπο κείμενο είναι χειρόγραφο. Στις μέρες μας το πρωτότυπο κείμενο θα μπορούσε να είναι απευθείας κείμενο ψηφιακής μορφής (π.χ. τα κείμενα από το διαδίκτυο)

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΠΡΩΤΟΓΕΝΗ ΔΕΔΟΜΕΝΑ

Κείμενο = jpeg ή pdf format, ειδικά αν το πρωτότυπο κείμενο είναι χειρόγραφο. Στις μέρες μας το πρωτότυπο κείμενο θα μπορούσε να είναι κατευθείαν ψηφιακής μορ΄φης (π.χ. τα κείμενα από το διαδίκτυο)

΄**Ηχος** = ηχητικά αρχεία, μη απολεστικής συμπίεσης (**.wav** και ΌΧΙ .mp3 ή ακόμη χειρότερα, mp4)

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΠΡΩΤΟΓΕΝΗ ΔΕΔΟΜΕΝΑ

Κείμενο = jpeg ή pdf format, ειδικά αν το πρωτότυπο κείμενο είναι χειρόγραφο. Στις μέρες μας το πρωτότυπο κείμενο θα μπορούσε να είναι κατευθείαν ψηφιακής μορφής (π.χ. τα κείμενα από το διαδίκτυο)

Ηχος = ηχητικά αρχεία, μη απολεστικής συμπίεσης (.wav και ΌΧΙ .mp3 ή ακόμη χειρότερα, mp4)

Βίντεο = .mov, .mpg, και κάθε format που υποστηρίζεται από τη Java

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΕΠΕΞΕΡΓΑΣΜΕΝΑ ΔΕΔΟΜΕΝΑ

Αρχεία που να αναφέρονται στα πρωτότυπα δεδομένα και τα προσδιορίζουν

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΕΠΕΞΕΡΓΑΣΜΕΝΑ ΔΕΔΟΜΕΝΑ

Αρχεία που να αναφέρονται στα πρωτότυπα δεδομένα και τα προσδιορίζουν

- Μεταγραφή

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΕΠΕΞΕΡΓΑΣΜΕΝΑ ΔΕΔΟΜΕΝΑ

Αρχεία που να αναφέρονται στα πρωτότυπα δεδομένα και τα προσδιορίζουν

- Μεταγραφή
- Επισημειώσεις

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΕΠΕΞΕΡΓΑΣΜΕΝΑ ΔΕΔΟΜΕΝΑ

Μεταγραφή

Διαφόρων ειδών: Ορθογραφική

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΕΠΕΞΕΡΓΑΣΜΕΝΑ ΔΕΔΟΜΕΝΑ

Μεταγραφή

Διαφόρων ειδών: Ορθογραφική

Φωνητική

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΕΠΕΞΕΡΓΑΣΜΕΝΑ ΔΕΔΟΜΕΝΑ

Μεταγραφή

Διαφόρων ειδών: Ορθογραφική

Φωνητική

Φωνολογική

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΕΠΕΞΕΡΓΑΣΜΕΝΑ ΔΕΔΟΜΕΝΑ

Μεταγραφή

Διαφόρων ειδών: Ορθογραφική

Φωνητική

Φωνολογική

Ανάλυσης Συνομιλίας (CA)

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΕΠΕΞΕΡΓΑΣΜΕΝΑ ΔΕΔΟΜΕΝΑ

Επισημειώσεις

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΕΠΕΞΕΡΓΑΣΜΕΝΑ ΔΕΔΟΜΕΝΑ

Επισημειώσεις

Εντοπισμός και προσδιορισμός μονάδων ή/και φαινομένων οποιουδήποτε γλωσσικού επιπέδου (Φωνητικής, Φωνολογίας, Μορφολογίας, Σύνταξης, Σημασιολογίας, Πραγματολογίας, Ανάλυσης Λόγου κλπ.)

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΕΠΕΞΕΡΓΑΣΜΕΝΑ ΔΕΔΟΜΕΝΑ

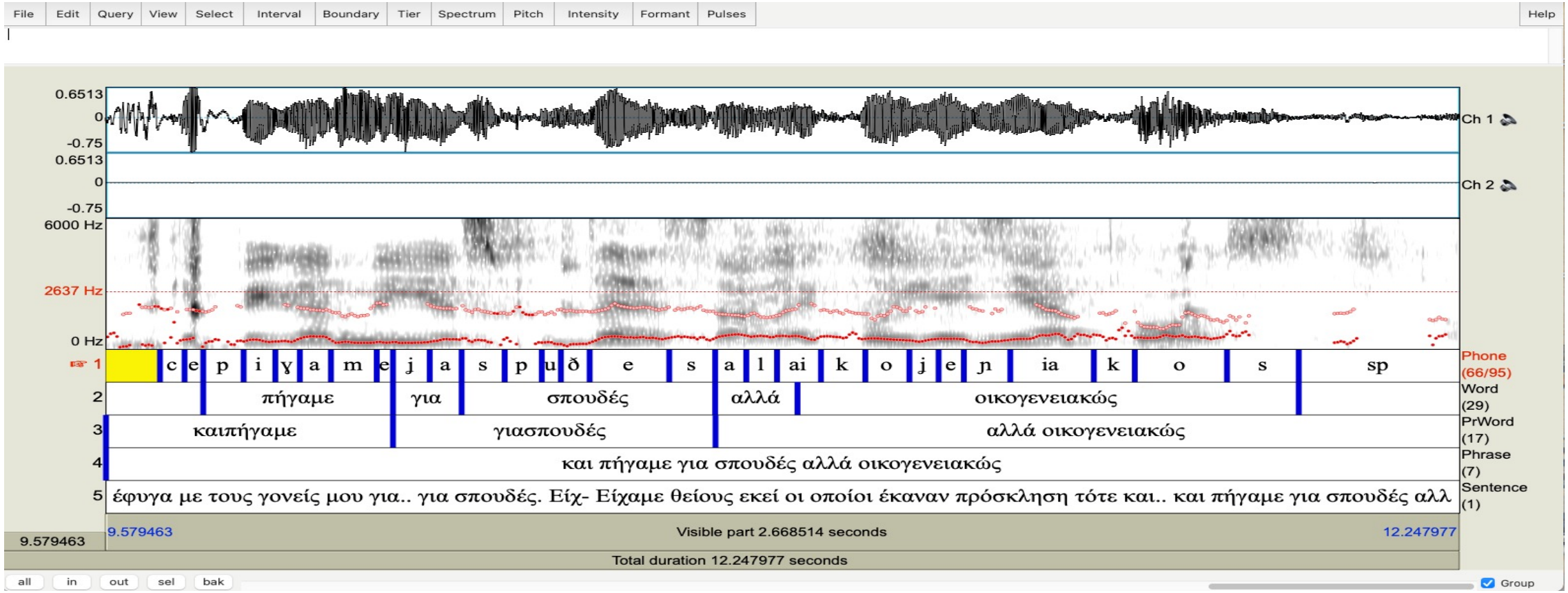
Επισημειώσεις

Εντοπισμός και προσδιορισμός μονάδων ή/και φαινομένων οποιουδήποτε γλωσσικού επιπέδου (Φωνητικής, Φωνολογίας, Μορφολογίας, Σύνταξης, Σημασιολογίας, Πραγματολογίας, Ανάλυσης Λόγου κλπ.)

Αναγνώριση όσο το δυνατόν περισσότερων formats, ειδικά των formats των λογισμικών ανοιχτού κώδικα (όπως Praat, Elan, CLAN, CHAT, κλπ.) που χρησιμοποιούνται διεθνώς για γλωσσολογική επισημείωση και ανάλυση

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΕΠΕΞΕΡΓΑΣΜΕΝΑ ΔΕΔΟΜΕΝΑ: Μεταγραφή και επισημειώσεις



Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΜΕΤΑΔΕΔΟΜΕΝΑ

Τα μεταδεδομένα θα πρέπει να προκύψουν από διεθνώς αναγνωρισμένα schemata, όπως:

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΜΕΤΑΔΕΔΟΜΕΝΑ

Τα μεταδεδομένα θα πρέπει να προκύψουν από διεθνώς αναγνωρισμένα schemata, όπως:

- CIDOC-CRM
- TEI
- FOF
- Dublin CORE
- CLARIN

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΜΕΤΑΔΕΔΟΜΕΝΑ

(Πιθανές πληροφορίες: i. Το γλωσσικό σύστημα, ii. το ερευνητικό πρόγραμμα, iii. Το είδος κειμένων, iv. Πληροφορίες για τους ομιλητές ή τους συγγραφείς, v. Πληροφορίες για τη γλωσσική κοινότητα, vi. Τεχνικά χαρακτηριστικά)

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

ΜΕΤΑΔΕΔΟΜΕΝΑ

(Πιθανές πληροφορίες: i. Το γλωσσικό σύστημα, ii. Πληροφορίες για το ερευνητικό πρόγραμμα, iii. Το είδος κειμένων, iv. Πληροφορίες για τους ομιλητές ή τους συγγραφείς, v. Πληροφορίες για τη γλωσσική κοινότητα, vi. Τεχνικά χαρακτηριστικά)

Το λογισμικό θα πρέπει να επιτρέπει την δυναμική εισαγωγή νέων μεταδεδομένων, τουλάχιστον εκείνων που συνοδεύουν κάθε νέο Σώμα Κειμένων που προστίθεται στη Βάση

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

Λογισμικά αναζήτησης και επεξεργασίας

Σύνθετη αναζήτηση

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

Λογισμικά αναζήτησης και επεξεργασίας

Σύνθετη αναζήτηση

(Συνδυασμός πληροφορίας από μεταδεδομένα και επεξεργασμένα δεδομένα)

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

Λογισμικά αναζήτησης και επεξεργασίας

Σύνθετη αναζήτηση

(Συνδυασμός πληροφορίας από μεταδεδομένα και επεξεργασμένα δεδομένα)

Χρήση scripts και εργαλείων AI για περαιτέρω επισημειώσεις, και εξαγωγή δεδομένων (π.χ. Forced alignment)

Γλωσσικές Βάσεις Δεδομένων: Προδιαγραφές

Λογισμικά αναζήτησης και επεξεργασίας

Παράδειγμα:

να ζητάω από το σύστημα να μου συγκεντρώσει όλες τις ερωτήσεις ολικής άγνοιας που έχουν παράξει γυναίκες χαμηλής μόρφωσης και μεγάλης ηλικίας από την Αχαΐα, και η Βάση Δεδομένων να βρίσκει αυτές τις ερωτήσεις στα ηχητικά αρχεία, να σώζει κάθε μία τέτοια ερώτηση σε νέο διαφορετικό ηχητικό αρχείο μαζί με το σχετικό κομμάτι μεταγραφής και επισημείωσης, να τις συγκεντρώνει σε ένα φάκελο και να μου επιτρέπει να το κατεβάσω στον υπολογιστή μου.

Γλωσσικές Βάσεις Δεδομένων: Προοπτικές

- Να στεγάσει / φιλοξενήσει όλο το δυνατό περισσότερα Σώματα Κειμένων

Γλωσσικές Βάσεις Δεδομένων: Προοπτικές

- Να στεγάσει / φιλοξενήσει όλο το δυνατό περισσότερα Σώματα Κειμένων
- Να τύχει συστηματικής και διαρκούς υποστήριξης από έναν φορέα ο οποίος θα μπορεί να εξασφαλίσει τέτοιου είδους υποστήριξη (π.χ. ένα πανεπιστήμιο, ένα ερευνητικό ινστιτούτο, ή ακόμη και από την Ακαδημία Αθηνών)

Γλωσσικές Βάσεις Δεδομένων: Προοπτικές

- Να στεγάσει / φιλοξενήσει όλο το δυνατό περισσότερα Σώματα Κειμένων
- Να τύχει συστηματικής και διαρκούς υποστήριξης από έναν φορέα ο οποίος θα μπορεί να εξασφαλίσει τέτοιου είδους υποστήριξη (π.χ. ένα πανεπιστήμιο, ένα ερευνητικό ινστιτούτο, ή ακόμη και από την Ακαδημία Αθηνών)
- Να συσπειρώσει έναν αριθμό γλωσσολόγων που να είναι πρόθυμες και πρόθυμοι να βοηθήσουν σε κάθε βήμα οργάνωσης ενός τέτοιου εργαλείου, ανάλογα με την ειδίκευση του καθενός και της καθεμιάς

Γλωσσικές Βάσεις Δεδομένων: Προοπτικές

- Να στεγάσει / φιλοξενήσει όλο το δυνατό περισσότερα Σώματα Κειμένων
- Να τύχει συστηματικής και διαρκούς υποστήριξης από έναν φορέα ο οποίος θα μπορεί να εξασφαλίσει τέτοιου είδους υποστήριξη (π.χ. ένα πανεπιστήμιο, ένα ερευνητικό ινστιτούτο, ή ακόμη και από την Ακαδημία Αθηνών)
- Να συσπειρώσει έναν αριθμό γλωσσολόγων που να είναι πρόθυμες και πρόθυμοι να βοηθήσουν σε κάθε βήμα οργάνωσης ενός τέτοιου εργαλείου, ανάλογα με την ειδίκευση του καθενός και της καθεμιάς
- Να γίνει εργαλείο αναφοράς στην γλωσσολογική κοινότητα που μελετά την ελληνική γλώσσα και τις ποικιλίες της

Σας ευχαριστώ πολύ!