ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΕΛΟΠΟΝΝΗΣΟΥ

ΥΠΟΛΟΓΙΣΤΙΚΗ ΓΛΩΣΣΟΛΟΓΙΑ
11η ΔΙΑΛΕΞΗ

ΠΑΝΑΓΙΩΤΗΣ ΓΑΚΗΣ

gakis@sch.gr

Τι είναι ένα σώμα κειμένων (text corpus);

- Corpus = σώμα στα Λατινικά
- Σώμα κειμένων είναι ένα **σώμα εμφανίσεων γλωσσολογικών στοιχείων** που προκύπτουν με φυσικό τρόπο
- Συνήθως συλλέγεται με κάποιο συγκεκριμένο σκοπό και είναι αντιπροσωπευτικό μιας γλώσσας
- Χρησιμοποιείται για να
  1. Επαληθεύσει υπάρχουσες θεωρίες και υποθέσεις Γλωσσολογίας
  2. Να δημιουργήσει καινούριες γλωσσολογικές υποθέσεις
  3. Εκτός Γλωσσολογίας, να παράσχει στοιχεία κειμενικά σε θέματα που αφορούν σε ανθρωπιστικές και κοινωνικές επιστήμες

Ο πιο διαδεδομένος ορισμός

*"Σώμα κειμένων θεωρείται **κάθε συλλογή τμημάτων μιας συγκεκριμένης γλώσσας**, τα οποία επιλέγονται και διατάσσονται σύμφωνα με συγκεκριμένα γλωσσολογικά κριτήρια, έτσι ώστε να μπορούν να χρησιμοποιηθούν ως αντιπροσωπευτικό δείγμα της γλώσσας αυτής" (Sinclair, 1996)*

Σωμάτα Κειμένων (Corpora)

- Συλλογές τεράστιου αριθμού αυθεντικών κειμένων, αποθηκευμένων σε ηλεκτρονική μορφή, επεξεργάσιμων και προσπελάσιμων με υπολογιστικά εργαλεία.
- Συγκροτημένα σύμφωνα με κριτήρια και αρχές, για να εξυπηρετήσουν συγκεκριμένους ερευνητικούς σκοπούς.
- Επιτρέπουν την αποθήκευση, την ταχύτατη ανάκληση και την επεξεργασία τεράστιου όγκου γλωσσικών πληροφοριών

Είναι ένα σώμα κειμένων (text corpus);

- Μια λίστα λέξεων (λεξικό)
- Ένα μεμονωμένο κείμενο
- Μια ΤΥΧΑΙΑ συλλογή κειμένων

Τι ΔΕΝ είναι ένα σώμα κειμένων (text corpus);

- Μια λίστα λέξεων (λεξικό)
- Ένα μεμονωμένο κείμενο
- Μια ΤΥΧΑΙΑ συλλογή κειμένων

Ηλεκτρονικά Σώματα Κειμένων (ΗΣΚ)

- Ορισμοί:

1. «**πεπερασμένη συλλογή μηχανο-αναγνώσιμων κειμένων**, επιλεγμένων έτσι ώστε να είναι όσο το δυνατό πιο αντιπροσωπευτική μιας γλώσσας ή γλωσσικής ποικιλίας» (McEnery & Wilson 1996: 177).
2. «συλλογή τμημάτων γλώσσας, τα οποία επιλέγονται και διατάσσονται σύμφωνα με συγκεκριμένα γλωσσολογικά κριτήρια, έτσι ώστε να μπορούν να χρησιμοποιηθούν ως αντιπροσωπευτικό δείγμα μιας συγκεκριμένης γλώσσας» (Sinclair 1996).

Ηλεκτρονικά Σώματα Κειμένων (ΗΣΚ)

- Ορισμοί:

3. η συλλογή τμημάτων γλώσσας τα οποία **επιλέγονται** και διατάσσονται σύμφωνα με συγκεκριμένα **γλωσσολογικά κριτήρια** έτσι ώστε να χρησιμοποιηθούν ως αντιπροσωπευτικό δείγμα μιας συγκεκριμένης γλώσσας (EAGLES 1996).
4. ένα **εκτεταμένο δείγμα αυθεντικής χρήσης** της υπό εξέταση γλώσσας, που συγκροτείται και χρησιμοποιείται ως πηγή στοιχείων για τη δημιουργία ή εξέταση υποθέσεων για τη φύση της γλώσσας» (Stubbs 2001: 6).

Συλλογή κειμένων η οποία είναι κωδικοποιημένη για τυποποιημένες (standardized) και ομοιογενείς εργασίες ανάκτησης γλωσσικής πληροφορίας.

Ο ρόλος των Η/Υ στην ανάλυση των ΗΣΚ

- Δυνατότητα τεράστιας **αποθήκευσης** κειμενικών δεδομένων (πάνω από 1 δισ. λέξεις)
- Μεγάλη ταχύτητα **επεξεργασίας** γλωσσικών δεδομένων (Wordsmith: λέξεις το δ/λπτο)
- Συνεπής και «αλάνθαστη» απόδοση σε επαναληπτικές διαδικασίες (**επαναληψιμότητα** (replicability)).

Ένα σώμα κειμένων

- Είναι αντιπροσωπευτικό μιας γλώσσας
- Φανερώνει τι είναι σύνηθες σε μια γλώσσα
- Μπορεί να δώσει ακριβείς στατιστικές μετρήσεις των φαινομένων της γλώσσας
- Αποθηκεύεται και ανακαλείται οποιαδήποτε στιγμή η πληροφορία σε αυτό
- Παρέχει φυσικά/πραγματικά παραδείγματα της γλώσσας
- Αποτελεί αντικειμενικό δείγμα της χρήσης της γλώσσας
- Είναι διαθέσιμο σε όλους
- Μπορεί να επικαιροποιείται συνεχώς και αν αντανακλά τις πρόσφατες αλλαγές στην γλώσσα
- Δεν μπορεί να παρέχει αρνητική ένδειξη για μη δυνατά φαινόμενα
- Δεν μπορεί από μόνο του να εξηγήσει αυτά που παρατηρούνται σε αυτό
- Αποτελεί περιορισμό για την όποια έρευνα στηρίζεται σε αυτό

Επισημείωση/Annotation

- Το κείμενο φέρει φωνολογικές, γραμματικές, συντακτικές ή σημασιολογικές πληροφορίες.
- Μέθοδοι επισημείωσης:

1. Χειρωνακτική

```
Perdita&NN1-NP0; ,&PUN; covering&VVG; the&AT0; bottom&NN1; of&PRF;
the&AT0; lorries&NN2; with&PRP; straw&NN1; to&TO0; protect&VVI;
the&AT0; ponies&NN2; '&POS; feet&NN2; ,&PUN; suddenly&AV0; heard&VVD-
VVN; Alejandro&NN1-NP0; shouting&VVG; that&CJT; she&PNP; better&AV0;
dig&VVB; out&AVP; a&AT0; pair&NN0; of&PRF; clean&AJ0; breeches&NN2;
and&CJC; polish&VVB; her&DPS; boots&NN2; ,&PUN; as*CJS; she&PNP;
'd&VM0; be&VBI; playing&VVG; in&PRP; the&AT0; match&NN1; that&DT0;
afternoon&NN1; .&PUN;
```

Επισημείωση/Annotation

- Η επισημείωση μπορεί να πραγματοποιηθεί σε διάφορα επίπεδα γλωσσολογικής γνώσης

1. **Φωνολογικό επίπεδο**
   - Όρια φωνητικών συλλαβών
   - Στοιχεία προσωδίας (ο τόνος ή ο επιτονισμός της φωνής, ή άλλα ποικίλα χαρακτηριστικά του ομιλητή ή του εκφωνήματος: την συναισθηματική κατάσταση του ομιλητή, την μορφή του εκφωνήματος (δήλωση, ερώτηση ή εντολή), την παρουσία ειρωνείας ή σαρκασμού, την έμφαση, αντίθεση ή εστίαση ή άλλα στοιχεία της γλώσσας που δεν κωδικοποιούνται από την γραμματική ή από την επιλογή του λεξιλογίο)
2. **Μορφολογικό επίπεδο**
   - Προθέματα
   - Επιθέματα
   - Λήμματα/Θέματα
   - Επισημείωση μορφολογικής πληροφορίας
   - Μέρη του Λόγου

Επισημείωση/Annotation

3. **Συντακτικό επίπεδο**
   - Treebanks
   - Ομαδοποίηση λέξεων σε φράσεις
4. **Σημασιολογικό επίπεδο**
   - Έννοιες λέξεων
5. **Πραγματολογικό επίπεδο**
   - Αντωνυμικές αναφορές (Αρχικά το κείμενο παραθέτει το όνομα και στη συνέχεια αναφέρεται σε αυτό μέσω (προσωπικών και δεικτικών) αντωνυμιών)
   - Επισημείωση υφολογικών στοιχείων

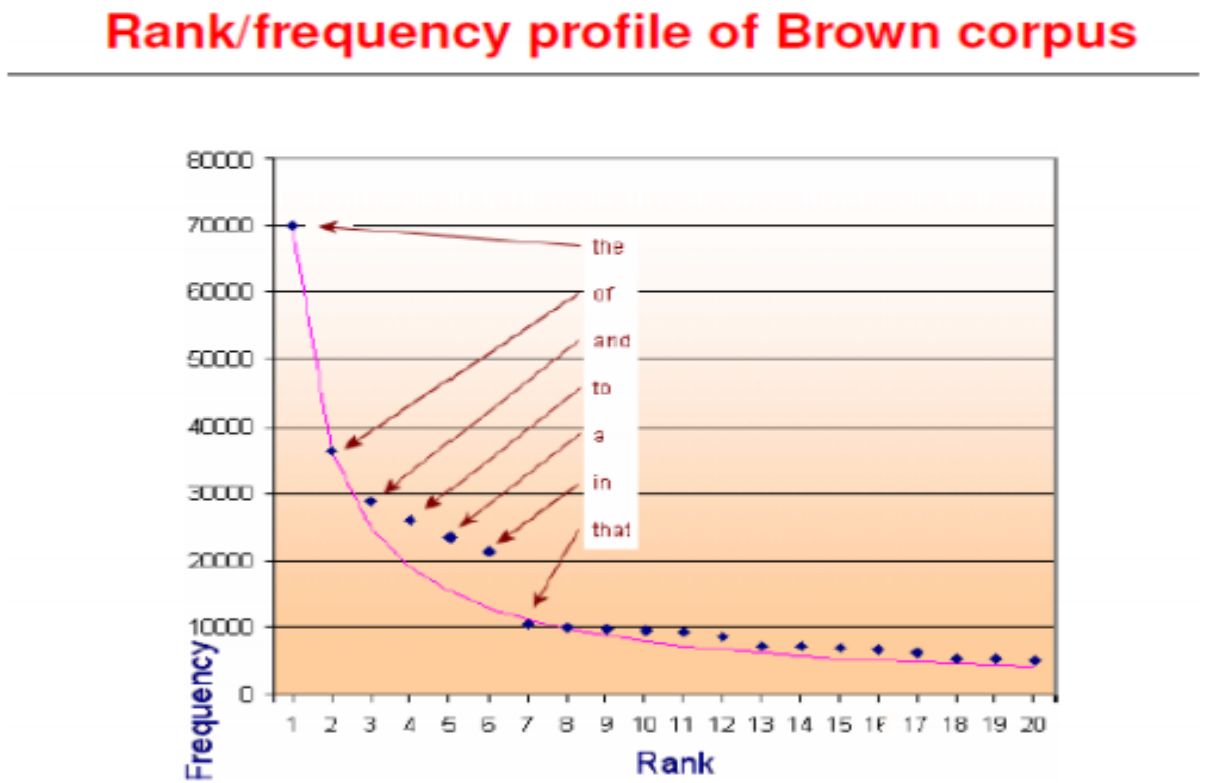
Τι μπορώ να μετρήσω με ένα ΣΚ; **Ωμές συχνότητες** (Raw frequencies)

- Η πιο **απλή ποσοτική προσέγγιση** στην επεξεργασία ενός σώματος κειμένων είναι η **μέτρηση των φορών** που εμφανίζεται μια λέξη/φράση μέσα στο σώμα κειμένων.
- Στο παραπάνω κείμενο:
  - Συχνότητα(«κειμένων») = ?
  - Συχνότητα («η») = ? (case-sensitive) ΠΕΖΑ - ΚΕΦΑΛΑΙΑ
  - Συχνότητα («η») = ? (όχι case-sensitive)

Ο Νόμος του Zipf (Zipf's Law)

- **Συχνότητα μιας λέξης είναι αντιστρόφως ανάλογη με τη σειρά της στον κατάλογο συχνότητας** (οι πιο συχνές λέξεις καταλαμβάνουν ένα μεγάλο μέρος του σώματος κειμένων, αφήνοντας το υπόλοιπο για πάρα πολλές λέξεις, που έχουν όμως πολύ μικρή συχνότητα (Νόμος του Zipf)
- Σε ένα σώμα κειμένων μετράω την **συχνότητα (f)** των λέξεων και τις ταξινομώ κατά φθίνουσα συχνότητα.
- **r = η θέση μιας λέξης** στην παραπάνω κατάταξη (rank)
- Ο Zipf (1949) ανακάλυψε ότι **f*r=k (σταθερό)**
- Εάν ο πιο συχνός όρος (the) εμφανίζεται f φορές τότε ο δεύτερος πιο συχνός όρος (of) εμφανίζεται f/2 φορές ο τρίτος πιο συχνός όρος (and) εμφανίζεται f/3 φορές …

Ο Νόμος του Zipf (Zipf's Law)



Τι μπορώ να μετρήσω με ένα ΣΚ; Κανονικοποιημένες συχνότητες (Normalized frequencies)

- Στο British National Corpus ομιλίας η υβριστική λέξη f**k εμφανίζεται 250 φορές
- Στο British National Corpus γραπτού λόγου η ίδια λέξη εμφανίζεται 500 φορές
- **Βρίζουν οι άνθρωποι με διπλάσια συχνότητα στον γραπτό λόγο από ότι στον προφορικό**;
- **Όχι,** το BNC γραπτού λόγου είναι 9 φορές **μεγαλύτερο σε μέγεθος** (~90 εκατ. λέξεις) από το BNC προφορικού λόγου (~10 εκατ. λέξεις)
- Κανονικοποιημένη συχνότητα στο σώμα προφορικού λόγου:
  - 250/10.000.000 = x/1.000.000 => x=25
- Κανονικοποιημένη συχνότητα στο σώμα γραπτού λόγου:
  - 500/90.000.000 = x/1.000.000 => x=5.55

Ωμές συχνότητες (Raw frequencies)

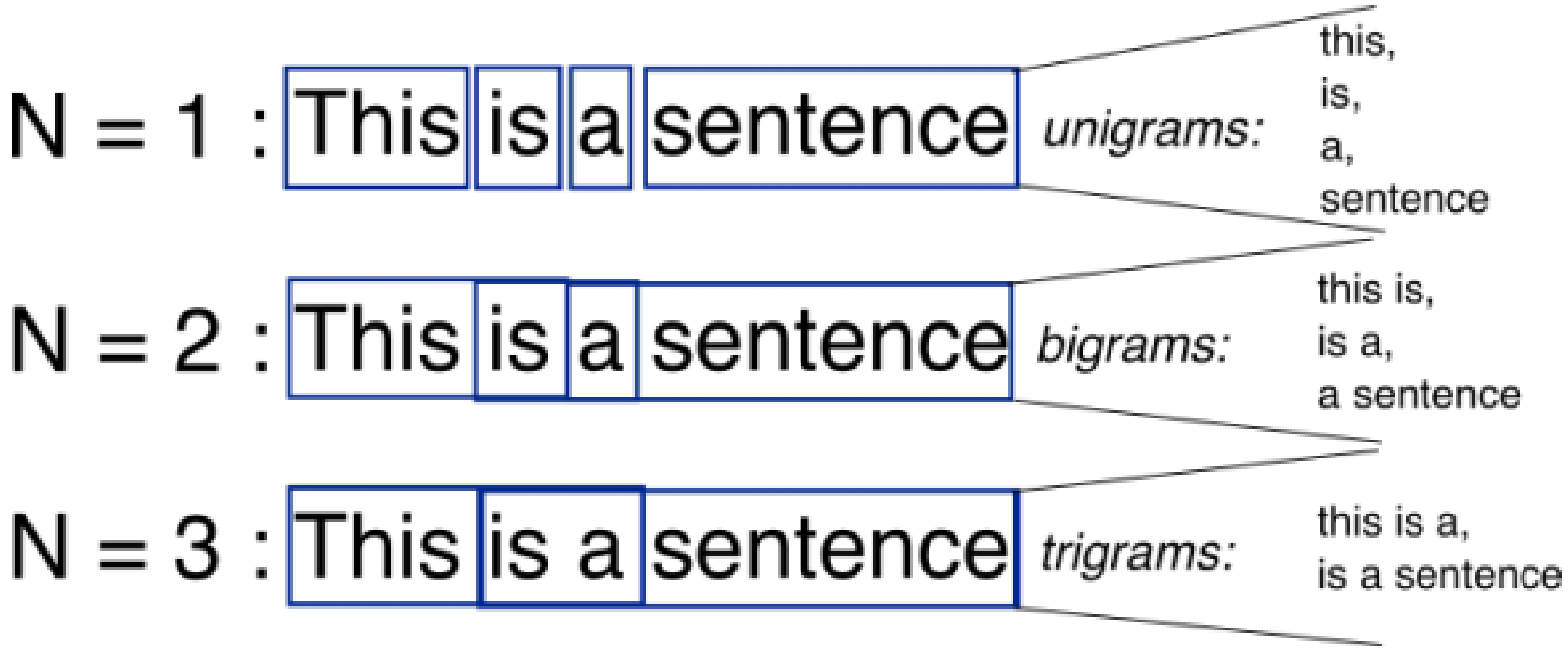
- Το πρόβλημα με τις ωμές συχνότητες είναι ότι **δεν λαμβάνουν υπόψη πώς κατανέμεται η λέξη/φράση μέσα στο σώμα κειμένων**.
- Είναι ομοιόμορφα κατανεμημένη σε όλα τα μέρη του σώματος, ή εμφανίζεται κατά κύριο λόγο σε ένα μέρος και πολύ σπάνια αλλού;
- Στο British National Corpus (BNC)
  - Οι λέξεις ***HIV, keeper, lively*** εμφανίζονται περίπου με την ίδια συχνότητα, δηλ. ~16 φορές ανα 1 εκατ. λέξεις
  - Αν χωρίσουμε το σώμα κειμένων σε 100 ισομεγέθη μέρη, τότε η λέξη **HIV** εμφανίζεται σε **62** από αυτά, ενώ οι λέξεις **keeper** & **lively** εμφανίζονται σε **97**.
  - Από αυτό συμπεραίνουμε ότι η λέξη HIV εμφανίζεται μέσα σε πιο εξειδικευμένα συμφραζόμενα.

Document Frequencies

- Έστω ότι το σώμα κειμένων απαρτίζεται από έγγραφα (documents).
- Document frequency: Ο αριθμός των εγγράφων στα οποία εμφανίζεται η λέξη

Τι μπορώ να μετρήσω με ένα ΣΚ; N-grams

N-gram: ακολουθία Ν λέξεων/χαρακτήρων σε ένα κείμενο



Τι μπορώ να μετρήσω με ένα ΣΚ; Συμφράσεις (Collocations)

Σύμφραση είναι μια έκφραση αποτελούμενη από δυο ή περισσότερες λέξεις, της οποίας το νόημα είναι διαφορετικό από τον συνδυασμό των νοημάτων των λέξεων που την αποτελούν:

Νέα Υόρκη (ενώ 'νέα εταιρία';)
Σκληρός δίσκος (ενώ 'σκληρό στρώμα';)

-Δεν μπορεί να αντικατασταθεί κάποια από τις λέξεις μιας σύμφρασης από άλλη (non-substitutability)
-Δεν μπορεί να συντεθεί το νόημα της σύμφρασης από το νόημα των λέξεων που την απαρτίζουν (non-compositionality)
- Δεν μπορεί να μεταφραστεί μια σύμφραση σε άλλη γλώσσα λέξη προς λέξη

Οι ιδιωματικές εκφράσεις είναι ακραία παραδείγματα συμφράσεων
- Πουλάει φύκια για μεταξωτές κορδέλες

Τι μπορώ να μετρήσω με ένα ΣΚ; Συμφράσεις (Collocations)

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Πώς υπολογίζω συμφράσεις σε ένα σώμα κειμένων;

Με συχνότητες;
Δίπλα φαίνονται τα πιο συχνά δίγραμμα (2-grams) σε ένα ΣΚ.
Εκτός από την Νέα Υόρκη, όλα τα υπόλοιπα δίγραμμα είναι ζευγάρια λειτουργικών λέξεων.

Λειτουργικές λέξεις (function words) είναι οι λέξεις σε μια γλώσσα που δεν περιέχουν σημασιολογική πληροφορία, αλλά χρησιμοποιούνται για να συνδέουν τις υπόλοιπες λέξεις του κειμένου μεταξύ τους

Λειτουργικές λέξεις είναι τα άρθρα, οι προθέσεις, οι σύνδεσμοι κλπ.

Δίλημμα

- Να χρησιμοποιήσω σώμα κειμένων που έχει δημιουργηθεί από άλλους ή να δημιουργήσω δικό μου σώμα κειμένων;;;

Ι. Γενικές αρχές σχεδιασμού ΗΣΚ

Γενικές αρχές σχεδιασμού: Αντιπροσωπευτικότητα (representativeness)

- Μπορώ να συγκροτήσω ένα πλήρες αρχείο όλων των μαρτυριών μιας γλώσσας;
- Ναι, σε μια νεκρή γλώσσα (π.χ. λατινική, αρχαία ελληνική) ή στην υπογλώσσα ενός πολύ ειδικού πεδίου (π.χ. μαρτυρικές καταθέσεις στο πλαίσιο μιας δίκης, συνδιαλέξεις με τηλεφωνικό κέντρο).
- Ο Sinclair (2008: 30) χαρακτηρίζει τη γλώσσα «πληθυσμό χωρίς όρια», ενώ ένα σώμα κειμένων είναι εξ ορισμού πεπερασμένο.
- Ένα σώμα κειμένων μιας ζωντανής γλώσσας δεν μπορεί παρά να είναι αναγκαστικά ένα υποσύνολο.
- Σύμφωνα με τον Barnbrook (1996: 24), ένα αντιπροσωπευτικό δείγμα πρέπει να διαθέτει παρόμοια χαρακτηριστικά με τον γλωσσικό πληθυσμό που στοχεύει να αντιπροσωπεύσει στην ανάλυση μιας γλώσσας.

Γενικές αρχές σχεδιασμού: Αντιπροσωπευτικότητα (representativeness)

- Άμεσα με την αντιπροσωπευτικότητα συνδέεται η δειγματοληψία.
- Για να καλυφθεί μεγαλύτερο εύρος κειμενικών ειδών μπορεί να καταφύγουμε στην επιλογή δειγμάτων από περισσότερα κείμενα αντί για λιγότερα ολόκληρα κείμενα.
- Αυτή η πρακτική ακολουθήθηκε κυρίως στην αρχή της δημιουργίας σωμάτων κειμένων όπως λ.χ. το Brown Corpus, για το οποίο συλλέχθηκαν δείγματα όμοιου μεγέθους από μια ποικιλία κειμενικών ειδών.

Γενικές αρχές σχεδιασμού: Αντιπροσωπευτικότητα (representativeness)

- Αντίθετα, ο Sinclair (2005), θεωρώντας ότι το κείμενο αποτελεί ενιαία ολότητα, είναι ρητά αντίθετος με τη δειγματοληψία και τονίζει ότι ένα σώμα κειμένων θα πρέπει να αποτελείται από κατά το δυνατόν ολόκληρα έγγραφα ή μεταγραφές ολοκληρωμένων προφορικών συμβάντων.
- Αν κατακερματιστεί η ολότητα, χάνεται πολύτιμη γλωσσική πληροφορία ή και διαστρεβλώνεται η εντύπωση για το τι μπορεί να εμφανιστεί σε κάθε κείμενο, άρα και στη γλώσσα.

Γενικές αρχές σχεδιασμού: Αντιπροσωπευτικότητα (representativeness)

- Ένα σώμα κειμένων θεωρείται αντιπροσωπευτικό δείγμα μιας γλώσσας όταν τα ευρήματα που βασίζονται στο περιεχόμενό του μπορούν να γενικευτούν για το σύνολο της γλώσσας αυτής.
- Κατά τον Kučera (2002), η αντιπροσωπευτικότητα αναφέρεται σε τρεις διαστάσεις κάθε σώματος κειμένων: το *μέγεθος*, την *αυθεντικότητα* και τις αναλογίες, τη σχετική *ισορροπία (balance)* δηλαδή μεταξύ των κειμενικών ειδών που το απαρτίζουν.

Γενικές αρχές σχεδιασμού: Μέγεθος

- Το ιδανικό μέγεθος ενός σώματος κειμένων δύσκολα προσδιορίζεται.
- Από το 1 εκατομμύριο λέξεις περίπου, που ήταν ένας επιθυμητός στόχος τη δεκαετία του 1970 για την αγγλική γλώσσα, σήμερα έχει εκτοξευτεί στο 1 δισεκατομμύριο λέξεις.
- Ας αναλογιστούμε ότι:
- μία σελίδα περιοδικού περιέχει γύρω στις 1.000 λέξεις
- μία σελίδα βιβλίου περιέχει γύρω στις 300 λέξεις
- μία ώρα ηχογραφημένου προφορικού λόγου συνήθως περιέχει περίπου 12.000-15.000 λέξεις και προϋποθέτει δύο μέρες απομαγνητοφώνησης.
- Στην πράξη, καθορίζεται ένας στόχος για το μέγεθος ενός σώματος κειμένων, ο οποίος αναπροσαρμόζεται σύμφωνα με το διαθέσιμο υλικό και την ερευνητική στοχοθεσία.

Γενικές αρχές σχεδιασμού: Αυθεντικότητα

- Η αυθεντικότητα συνδέεται με την προέλευση των κειμένων από φυσικά περιβάλλοντα επικοινωνίας, που έχουν κεντρική σημασία και δεν είναι περιθωριακά για τη συγκεκριμένη γλωσσική ποικιλία που μελετάται.
- Τα κείμενα αυτά θα πρέπει να έχουν δημιουργηθεί με αυθόρμητο, φυσικό τρόπο, και όχι κάτω από πειραματικές συνθήκες.

Γενικές αρχές σχεδιασμού: Ισορροπία (balance)

Γενικές αρχές σχεδιασμού: Ισορροπία (balance)

- Για να είναι ένα σώμα κειμένων ισορροπημένο δείγμα του συνολικού πληθυσμού των κειμένων που παράγονται στην υπό μελέτη γλώσσα πρέπει να συμπεριλαμβάνει μέσω ποικίλων τεχνικών δειγματοληψίας όσον το δυνατόν ευρύτερη ποικιλία ή ακόμη και το πλήρες εύρος των κειμενικών ειδών αυτής της γλώσσας σε αναλογίες που εξασφαλίζουν τη «σχετική» ισορροπία μεταξύ τους.

Πρακτικές δυσκολίες (Atkins & Rundell 2008: 63-66):

- Δυσκολία συγκέντρωσης και μεταγραφής δεδομένων προφορικού λόγου, με αποτέλεσμα τα ΗΣΚ να περιλαμβάνουν μικρότερο ποσοστό προφορικών κειμένων ή και καθόλου προφορικά κείμενα.
- Δειγματοληψία ευχέρειας: πρόκριση εύκολα προσβάσιμων κειμενικών ειδών (π.χ. εφημερίδων, λογοτεχνικών κειμένων), με αποτέλεσμα να διαταράσσεται η αντιπροσωπευτικότητα και η ισορροπημένη κατανομή του υλικού.
- Νέα κειμενικά είδη που συνδέονται αποκλειστικά με τη χρήση του διαδικτύου και εμφανίστηκαν την τελευταία δεκαετία (π.χ. chatrooms, κοινωνικά δίκτυα, ιστολόγια) απουσιάζουν εντελώς από σώματα κειμένων που συγκροτήθηκαν πριν από το 2000.

ΙΙ. Είδη ΗΣΚ

Είδη Σωμάτων Κειμένων

- Τα σώματα κειμένων διαφέρουν μεταξύ τους ανάλογα με:

- την τροπικότητα
- το εύρος της γλωσσικής και κειμενικής ποικιλίας,
- τη δυνατότητα ανανέωσης του υλικού τους,
- τη χρονική κάλυψη των κειμένων που περιλαμβάνουν,
- τον αριθμό γλωσσών,
- το σύστημα επισημείωσης,
- τις συμπληρωματικές γλωσσικές ή πραγματολογικές πληροφορίες.

Τροπικότητα

- Προφορικό (με αντίστοιχες ηχογραφήσεις)
- Γραπτό
- Μικτό
- Πολυτροπικό

Εύρος γλωσσικής κάλυψης: γενικά vs εξειδικευμένα

- Ανάλογα με το εύρος των κειμενικών ειδών που περιλαμβάνουν, τα σώματα κειμένων διακρίνονται σε γενικά (general language corpora) και σε εξειδικευμένα (specialised corpora).
- Τα γενικά σώματα κειμένων αφορούν το σύνολο μιας γλώσσας (π.χ. της ελληνικής, της αγγλικής, της ιταλικής) ή μιας γεωγραφικής γλωσσικής ποικιλίας (π.χ. της βρετανικής ή της αμερικανικής αγγλικής), και εξού και θεωρούνται σώματα κειμένων αναφοράς (reference corpora) (Baker et al. 2006).

Γενικά Σώματα Κειμένων

- Ποικιλία τρόπων (προφορικός/γραπτός λόγος), θεματικών πεδίων και μέσων δημοσίευσης (π.χ. ραδιόφωνο, τηλεόραση, εφημερίδα, περιοδικό, βιβλίο).
- Ευρύ φάσμα γενικών κειμενικών ειδών (π.χ. τύπος, εκπαιδευτικά κείμενα, λογοτεχνία, ακαδημαϊκός λόγος) και ειδικότερων κειμενικών ειδών τόσο από τον γραπτό λόγο (π.χ. για τον τύπο: αστυνομικό ρεπορτάζ, άρθρα γνώμης, αγγελίες, δελτίο καιρού) όσο από τον προφορικό λόγο (π.χ. συνομιλίες, συνεντεύξεις, διαλέξεις).
- Συχνά περιέχουν υπο-σώματα (π.χ. σώμα προφορικών δεδομένων, σώμα κειμένων ακαδημαϊκού λόγου κ.ά.).

British National Corpus (BNC)

- «Εμβληματικό» γενικό σώμα κειμένων αναφοράς για τα βρετανικά αγγλικά.
- Οι αρχές συγκρότησης του BNC, κυρίως σε ό,τι αφορά τις μεθόδους και τα κριτήρια δειγματοληψίας του κειμενικού υλικού, αποτέλεσαν πρότυπο αναφοράς για την κατάρτιση εθνικών σωμάτων κειμένων σε πολλές γλώσσες παγκοσμίως.
- Περιλαμβάνει περίπου 100 εκατομμύρια λέξεις από μεικτά δεδομένα, εκ των οποίων το 90% είναι γραπτά κείμενα και το 10% μεταγραμμένα προφορικά δεδομένα.
- Το BNC είναι ένα στατικό (static corpus), δηλαδή σταθερό ως προς το μέγεθος σώμα κειμένων (περιλαμβάνει κειμενικά δείγματα μιας συγκεκριμένης χρονικής περιόδου (1980-1993) χωρίς έκτοτε ανανέωση του κειμενικού του υλικού.

Η ελληνική πραγματικότητα

Εξειδικευμένα Σώματα Κειμένων

- Τα εξειδικευμένα σώματα κειμένων συγκροτούνται με στόχο τη μελέτη της γλώσσας:
- σε ειδικές κειμενικές ποικιλίες ή είδη λόγου (π.χ. επιστημονικού, ακαδημαϊκού κ.ά.), σε ένα μέσο (π.χ. ηλεκτρονική επικοινωνία), σε κοινωνιογλωσσικές ποικιλίες, όπως π.χ. μια διαλεκτική ποικιλία, μια κοινωνιόλεκτος με βάση την ηλικία, το φύλο, το μορφωτικό επίπεδο κ.λπ. των ομιλητών κ.ο.κ.
- Δεν αντιπροσωπεύουν τη χρήση της γενικής γλώσσας, αλλά παρουσιάζουν σε υψηλή συχνότητα γλωσσικά φαινόμενα που απαντούν σπανιότερα ή απουσιάζουν από τα γενικά σώματα κειμένων (π.χ. ονοματοποίηση σε επιστημονικό λόγο, αργκοτισμοί σε νεανική επικοινωνία).
- Τα σώματα αυτά είναι συνήθως μικρότερα σε μέγεθος από τα γενικά.
- Η ομοιογένεια των κειμένων που απαρτίζουν τη συλλογή είναι αυτή που εξασφαλίζει την αντιπροσωπευτικότητά της, καθώς και ποσοτικές παράμετροι, όπως ο αριθμός των λέξεων (π.χ. περίπου 1.000 λέξεις) ανά κειμενικό δείγμα και ο αριθμός των αρχείων (π.χ. περίπου 10-15 αρχεία) ανά κειμενικό είδος.

Εξειδικευμένα Σώματα Κειμένων

- COLT (Bergen Corpus of London Teenage Language) περιέχει προφορικές συνομιλίες από εφήβους του Λονδίνου, ηλικίας 13-17 ετών.
- Διδακτικά Βιβλία (Γυμνασίου και Λυκείου) του Παιδαγωγικού Ινστιτούτου της Πύλης για την Ελληνική Γλώσσα, περίπου 2 εκατ. Λέξεων.

Εξειδικευμένα Σώματα Κειμένων

Σώματα κειμένων ακαδημαϊκού λόγου (academic corpora):

- Michigan Corpus of Academic Spoken English (MICASE, 1.8 εκατομμύρια λέξεις): Συλλογή από μεταγραμμένα προφορικά δεδομένα ακαδημαϊκών ομιλιών και συνομιλιών στα αμερικανικά αγγλικά σε ποικίλα γνωστικά πεδία, από το Πανεπιστήμιο του Michigan, ελεύθερα διαθέσιμη μέσω διαδικτυακής διεπαφής.
- British Academic Spoken English Corpus (BASE, 1.6 εκατομμύρια λέξεις): Σώμα προφορικών κειμένων ακαδημαϊκού λόγου στα βρετανικά αγγλικά, που συγκροτήθηκε κατά το πρότυπο του MICASE. Ελεύθερα διαθέσιμο μέσω της διεπαφής του Sketch Engine.
- British Academic Written English Corpus (BAWE, 7 εκατομμύρια λέξεις): Σώμα γραπτών κειμένων ακαδημαϊκού λόγου στα βρετανικά αγγλικά, γλωσσικά επισημειωμένο. Ελεύθερα διαθέσιμο μέσω της διεπαφής του Sketch Engine.

Αναπτυξιακά σώματα κειμένων (developmental language corpora):

- CHILDES database: Συνομιλίες παιδιών σε διάφορες ηλικίες και προσπέλαση μέσω διαδικτυακής διεπαφής.
- LUCY: Γραπτές εργασίες εφήβων και παιδιών στα βρετανικά αγγλικά, για συγκριτικές έρευνες με γραπτά μη φυσικών ομιλητών. Διατίθεται σε καταφορτώσιμη μορφή χωρίς ενσωματωμένο εργαλείο αναζήτησης.
- Louvain Corpus of Native English Essays (LOCNESS, 324 χιλιάδες λέξεις): Γραπτά δοκίμια Βρετανών και Αμερικανών φοιτητών, για συγκριτικές μελέτες με αντίστοιχα σώματα κειμένων μη φυσικών ομιλητών, όπως το International Corpus of Learner English.

Εξειδικευμένα Σώματα Κειμένων

- **Σώματα κειμένων εκμάθησης ξένης γλώσσας (learner corpora):**
- Ανάλυση λαθών (error analysis) και χρήσης συγκεκριμένου λεξιλογίου ή γραμματικών και συντακτικών δομών από τους μαθητές μιας ξένης γλώσσας.
- Συνηθισμένες είναι οι συγκριτικές έρευνες με σώματα κειμένων εκμάθησης και σώματα κειμένων φυσικών ομιλητών.

Εξειδικευμένα Σώματα Κειμένων

- International Corpus of Learner English (ICLE, 3.7 εκατομμύρια λέξεις): γραπτά δοκίμια μαθητών της αγγλικής, μέσου και προχωρημένου επιπέδου, διαθέσιμο σε έντυπη μορφή και CD επί πληρωμή.
- Learner Corpora Made Easy!

http://koreanlearnercorpusblog.blogspot.gr/

- Learner corpora around the world-CECL

https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

Δυνατότητα ανανέωσης του υλικού: στατικά (static) vs δυναμικά (dynamic)

- Ανάλογα με τον αν παραμένουν σταθερά ως προς το μέγεθός τους από τη στιγμή της συγκρότησή τους ή αν ανανεώνονται συστηματικά.
- **Ποιο είναι το βασικό πλεονέκτημα των δυναμικών σωμάτων κειμένων;**
- Συνεχής ροή γλωσσικού υλικού, που επιτρέπει την παρακολούθηση πρόσφατων γλωσσικών αλλαγών, γι' αυτό και συχνά αναφέρονται ως σώματα ελέγχου ή σώματα παρακολούθησης (monitor corpora) (π.χ. Bank of English).

COCA

- The Corpus of Contemporary American English (COCA): περιέχει 400 εκατομμύρια λέξεις από το 1990-2012 και είναι το πρώτο μεγάλο σε μέγεθος και ισορροπημένο ως προς τις κειμενικές ποικιλίες σώμα κειμένων της αμερικανικής αγγλικής.
- Ο σχεδιασμός του επιτρέπει την ανανέωσή του σχεδόν ετησίως κατά 20 εκατομμύρια λέξεις, ισάριθμα κατανεμημένες στα διάφορα κειμενικά είδη→ αξιόπιστο μοντέλο καταγραφής των γλωσσικών αλλαγών.
- Είναι προσπελάσιμο μέσω της διαδικτυακής διεπαφής Corpus BYU.

https://www.wordfrequency.info/comparison.asp

Word frequency data

introduction samples compare non-English related sites get data

Our data is based on two different corpora: the 14 billion word iWeb corpus, and the 560 million word Corpus of Contemporary American English (COCA; 450 million words when the wordlists were created). COCA is the only large corpus of English that is both up-to-date (the latest texts are from late 2017) and which is based on a wide range of genres (e.g. spoken, fiction, newspapers, magazines, and academic writing). Most of the following refers to the COCA word lists.

Why worry about what corpus is used? After all, there are many English word lists and frequency lists out on the Web (see in particular the British National Corpus and the American National Corpus). Some are good, and others are very poor in quality. Not all frequency lists are created equal.

One should be very, very suspicious of word lists that are taken from messy web data, outdated texts, or corpora that are too small to effectively model what is happening in the real world. Or worse, word lists that don't give you *any* idea what they are based on. As the saying goes: "*garbage in* (*bad texts*), *garbage out* (*frequency lists*)".

Here's some questions you might ask yourself as you consider downloading or purchasing a word list:

Depth and accuracy. Why do so many wordlists on the web contain just the top 1000-3000 words of English? Why not the top 20,000 or 60,000? It's because even a bad corpus (the collection of texts that the word lists are based on) can produce a moderately accurate list for the very most frequent words. But because the corpus is neither deep nor balanced enough, you start getting messy data for medium and lower frequency words. Ask to see samples of the top 20,000 or 60,000 words (e.g. every 7th or 10th word). If they don't have it, then you should be very, very suspicious of that word list.

Word freq

introduction samples compare

non-English
Spanish
Portuguese

related sites
English-Corpora.org
Full-text data
Collocates
N-grams
WordAndPhrase
Academic vocabulary

get data

Our data is based on two different corpora: the 14 billion word iWeb corpus, and the 560 million word English (COCA; 450 million words when the wordlists were created). COCA is the only large corpus of Engli texts are from late 2017) and which is based on a wide range of genres (e.g. spoken, fiction, newspapers Most of the following refers to the COCA word lists.

Why worry about what corpus is used? After all, there are many English word lists and frequency lists ou British National Corpus and the American National Corpus). Some are good, and others are very poor i created equal.

One should be very, very suspicious of word lists that are taken from messy web data, outdated text effectively model what is happening in the real world. Or worse, word lists that don't give you *any* idea wh goes: "*garbage in* (*bad texts*), *garbage out* (*frequency lists*)".

Here's some questions you might ask yourself as you consider downloading or purchasing a word list:

Depth and accuracy. Why do so many wordlists on the web contain just the top 1000-3000 words of English? Why not the top 20,000 or 60,000? It's because even a bad corpus (the collection of texts that the word lists are based on) can produce a moderately accurate list for the very most frequent words. But because the corpus is neither deep nor balanced enough, you start getting messy data for medium and lower frequency words. Ask to see samples of the top 20,000 or 60,000 words (e.g. every 7th or 10th word). If they don't have it, then you should be very, very suspicious of that word list.

Χρονική κάλυψη: συγχρονικά και διαχρονικά

- Συγχρονικά: περιλαμβάνουν κείμενα από μία συγκεκριμένη χρονική περίοδο.
- Ένα διαχρονικό σώμα κειμένων περιλαμβάνει κείμενα που καλύπτουν μία ή περισσότερες ιστορικές περιόδους ή ολόκληρη την ιστορική εξέλιξη της υπό μελέτη γλώσσας.

- Διαχρονικό σώμα ελληνικών κειμένων του 20ού αιώνα

  http://greekcorpus20.phil.uoa.gr/

- The Helsinki Corpus of English: Περιέχει 1,5 εκατομμύριο λέξεις από ποικιλία κειμενικών ειδών της περιόδου 750 έως 1700 (παλαιά, μέση και πρώιμη σύγχρονη περίοδο της βρετανικής αγγλικής). Ελεύθερα προσπελάσιμο.
- Corpus of Historical Corpus of English (COHA): το μεγαλύτερο διαχρονικό σώμα κειμένων με περισσότερες από 400 εκατομμύρια λέξεις για την αμερικανική αγγλική. Περιλαμβάνει κείμενα της περιόδου 1810 έως 2009 (λογοτεχνία, περιοδικά εφημερίδες και ακαδημαϊκά κείμενα). Είναι ελεύθερα προσβάσιμο από το Corpus BYU.

Επιλογή γλώσσας: μονόγλωσσα vs πολύγλωσσα

- Ανάλογα με τον αριθμό των γλωσσών που περιλαμβάνουν, τα σώματα κειμένων διακρίνονται σε μονόγλωσσα (monolingual corpora) και πολύγλωσσα (multilingual corpora) (McEnery & Hardie 2012· Baker et al. 2006).
- Τα μονόγλωσσα σώματα κειμένων καλύπτουν μόνο μία γλώσσα ή μία γλωσσική ποικιλία.
- Τα πολύγλωσσα σώματα κειμένων περιέχουν κείμενα από διαφορετικές γλώσσες ή γλωσσικές ποικιλίες και διακρίνονται περαιτέρω σε παράλληλα (parallel) και συγκρίσιμα (comparable).

Επιλογή γλώσσας: Παράλληλα Σώματα Κειμένων

- Αποτελούνται από τα ίδια κείμενα σε διαφορετικές γλώσσες (π.χ. κοινό πρωτότυπο και μτφρ. σε άλλες γλώσσες, διασκευές και αποδόσεις του ίδιου έργου, διάλογοι σε τηλεοπτική εκπομπή και υπότιτλοι που τους συνοδεύουν).
- Περιλαμβάνουν κυρίως επίσημα έγγραφα (π.χ. κυβερνητικά ενημερωτικά φυλλάδια, πρακτικά κοινοβουλίων, εκθέσεις διεθνών οργανισμών όπως η Ευρωπαϊκή Ένωση και ο Ο.Η.Ε.) και τεχνικά εγχειρίδια σε πολύγλωσσες μεταφράσεις.
- Βασικό γνώρισμα των παράλληλων σωμάτων κειμένων είναι η *ευθυγράμμιση* (alignment) ανάμεσα σε παραγράφους, προτάσεις ή λέξεις, δηλαδή η συσχέτιση ένα προς ένα των κειμένων στα επίπεδα αυτά.
- Χρησιμοποιούνται κατά κόρον για περιγραφικές και εμπειρικές μελέτες στο πλαίσιο των μεταφραστικών σπουδών, ως εργαλείο δημιουργίας ασκήσεων και διαγωνισμάτων στη διδασκαλία της μετάφρασης και της ξένης γλώσσας, ως πηγή άντλησης δεδομένων και εμπλουτισμού δίγλωσσων λεξικών, αλλά και για την εκπαίδευση και βελτίωση συστημάτων στατιστικής μηχανικής μετάφρασης.

Επιλογή γλώσσας: Συγκρίσιμα Σώματα Κειμένων

- Ένα συγκρίσιμο σώμα κειμένων περιλαμβάνει κείμενα από δύο ή περισσότερες γλώσσες ή γλωσσικές ποικιλίες τα οποία είναι διαφορετικά μεταξύ τους, δηλαδή δεν είναι τα πρωτότυπα κείμενα και οι μεταφράσεις τους, και των οποίων η συγκρισιμότητα έγκειται στο κοινό πλαίσιο δειγματοληψίας που ακολουθείται κατά τη συγκρότησή τους (π.χ. ίδια περίοδος δειγματοληψίας) και τις κοινές προδιαγραφές (π.χ. ίδια αναλογία από τα ίδια κειμενικά είδη στην ίδια περίοδο δειγματοληψίας) (McEnery & Hardie 2012: 19-20).
- Π.χ. σύνολο ειδησεογραφικών κειμένων για το προσφυγικό από ελληνικές και βρετανικές εφημερίδες, άρθρα της Wikipedia για την ίδια θεματική περιοχή σε άλλες γλώσσες.

Επιλογή γλώσσας: Συγκρίσιμα Σώματα Κειμένων

- Στην περίπτωση των συγκρίσιμων σωμάτων κειμένων δεν υπάρχει αντιστοίχιση, αλλά συγκρίνονται μεταξύ τους δύο ή περισσότερες γλώσσες με βάση τις λεξιλογικές επιλογές ή τις συντακτικές δομές, χωρίς να εμπλέκεται ο παράγοντας της μετάφρασης.
- ICE (International Corpus of English), που περιλαμβάνει περίπου 1 εκατ. λέξεις από πολλές γεωγραφικές ποικιλίες της αγγλικής (π.χ. του Καναδά, της Μεγ. Βρετανίας, του Χονγκ Κονγκ, της Νιγηρίας κ.λπ.) με βάση την ίδια μεθοδολογία δειγματοληψίας δεδομένων (ίδια αναλογία κειμένων από τα ίδια κειμενικά είδη της ίδιας χρονικής περιόδου).

Εμπλουτισμός ΗΣΚ με τρία είδη πληροφορίας

- μεταδεδομένα (metadata),
- κειμενικοί χαρακτηρισμοί (textual annotations)
- γλωσσικοί χαρακτηρισμοί ή γλωσσικές επισημειώσεις (linguistic annotations)

(McEnery and Hardie 2012: 14, 30)

Εμπλουτισμός ΗΣΚ με τρία είδη πληροφορίας

- Η διαδικασία εμπλουτισμού των κειμένων με μεταδεδομένα και κειμενικούς χαρακτηρισμούς ονομάζεται **κωδικοποίηση κειμένων (text encoding)** και αποτελεί μαζί με την απόδοση **γλωσσικών χαρακτηρισμών** το τελευταίο στάδιο κατά τη συγκρότηση σωμάτων κειμένων.

Μεταδεδομένα (metadata)

- Πληροφορίες που αφορούν το ίδιο το κείμενο (π.χ. για γραπτό κείμενο: ποιος το έγραψε, πότε, σε ποια γλώσσα, σε ποιο κειμενικό είδος ανήκει, σε ποιο μέσο δημοσιεύθηκε κ.ά.
- Στην περίπτωση του προφορικού υλικού, τα μεταδεδομένα μπορεί να περιλαμβάνουν λεπτομέρειες σχετικά με το πότε και πού έγινε η ηχογράφηση, αν μια συνομιλία είναι αυθόρμητος καθημερινός λόγος ή επίσημη ομιλία, πόσα άτομα συνομιλούν, ποια είναι η ηλικία και το φύλο τους, ποιες οι συνομιλιακές τους σχέσεις κ.ο.κ.).
- Τα μεταδεδομένα κωδικοποιούνται είτε μέσα στο κείμενο, στην ενότητα πληροφοριών που προηγείται του κειμένου και ονομάζεται κεφαλίδα (header), είτε διατηρούνται σε χωριστό έγγραφο ή βάση δεδομένων.

Κειμενικός χαρακτηρισμός (textual annotation)

- Επισήμανση στοιχείων της δομής ή/και της μορφοτύπησης (format) του ΗΣΚ: π.χ. σημεία στίξης, λέξεις, προτάσεις, παράγραφοι, τίτλοι, υπότιτλοι, πλάγιοι και έντονοι χαρακτήρες, ή (σε προφ. ΗΣΚ) η αλλαγή ομιλητών σε μια συνομιλία.

Κειμενικός χαρακτηρισμός (textual annotation) Παράδειγμα σε XML (Extensible Markup Language) (σε Γούτσος & Φραγκάκη 2015:35)

(1) <motto> Στον Γ. Μ. Κότσιανο… </motto> <title> Πρόλογος </title> Η παρούσα μεταπτυχιακή εργασία εκπονήθηκε στο ΒΕΜΜΟ της ιατρικής σχολής του πανεπιστημίου Κρήτης. Για την επιτυχή ολοκλήρωση της με βοήθησαν αρκετοί άνθρωποι τους οποίους θα ήθελα να ευχαριστήσω.   Το θέμα της εργασίας υποδείχθηκε από τον Καθηγητή του πανεπιστημίου Κρήτης Μ. Τσιλιμπάρη ο οποίος ήταν και ο επιβλέπων καθηγητής και θα ήθελα να τον ευχαριστήσω για την ευκαιρία που μου έδωσε να συνεργαστώ μαζί του αλλά και για την βοήθεια που μου προσέφερε προκειμένου να καταστεί δυνατή η υλοποίηση της εργασίας.   Ευχαριστώ όλα τα μέλη της εξεταστικής επιτροπής (Λ. Ναουμίδη και Χ. Γκίνη) που με τις επισημάνσεις του μου δόθηκε η δυνατότητα να διορθώσω και να βελτιώσω το παρόν κείμενο.   Ιδιαίτερα θα ήθελα να…

TEI (Text Encoding Initiative) (Burnard 2002)

```
<teiHeader>
<fileDesc>
   <distributor> BASE and Oxford Text Archive </distributor>
   <idno> ahlct001 </idno>
   <recording n="10364" dur="01:07:50">
   <date> 01/12/1998 </date>
   <equipment> <p> video </p> </equipment> </fileDesc>
<profileDesc>
   <langUsage> <language id="en"> English </language>
   <particDesc> <person role="main speaker" n="n" id="nm0001" sex="m">
      <p> nm0001, main speaker, non-student, male </p> </person>
      <person role="participant" n="s" id="sm0003" sex="m"> <p> sm0003,
      participant, student, male </p> </person> <person role="observer"
      n="o" id="om0004" sex="m"> <p>om0004, observer, observer, male</p>
      </person> </particDesc>
   <textClass> <item n="speechevent"> Lecture </item> <item
   n="partlevel">PG/staff</item> </teiHeader>
```

Γλωσσικός χαρακτηρισμός (linguistic annotation)

- Επισημείωση (annotation) ονομάζεται η *«προσθήκη ερμηνευτικών γλωσσικών πληροφοριών σε ένα σώμα κειμένων»* (Leech 2005).
- Διαδικασία κατά την οποία οι λέξεις που έχουν αναγνωριστεί κατά το στάδιο του κειμενικού χαρακτηρισμού (corpus markup) επισημειώνονται με ποικίλες γλωσσικές πληροφορίες:

α) με το **λήμμα**, δηλαδή την ουδέτερη μορφή στην οποία ανάγονται όλοι οι κλιτικοί τύποι μιας λέξης (π.χ. οι κλιτικοί τύποι *αντιδρούσε, αντιδρά, αντιδράσει* ανάγονται στο λήμμα «αντιδράω/ώ»),

β) τη **γραμματική κατηγορία** (*επισημείωση για μέρη του λόγου/part-of-speech ή POS tagging*: π.χ. αν είναι ουσιαστικά, ρήματα, επιρρήματα, προθέσεις, σύνδεσμοι κτλ.),

γ) τις **συντακτικές δομές** (*συντακτική επισημείωση/parsing):* χαρακτηρισμός της συντακτικής δομής σε επίπεδο πρότασης π.χ. αναγνώριση ΟΦ, ΡΦ, ΠΦ),

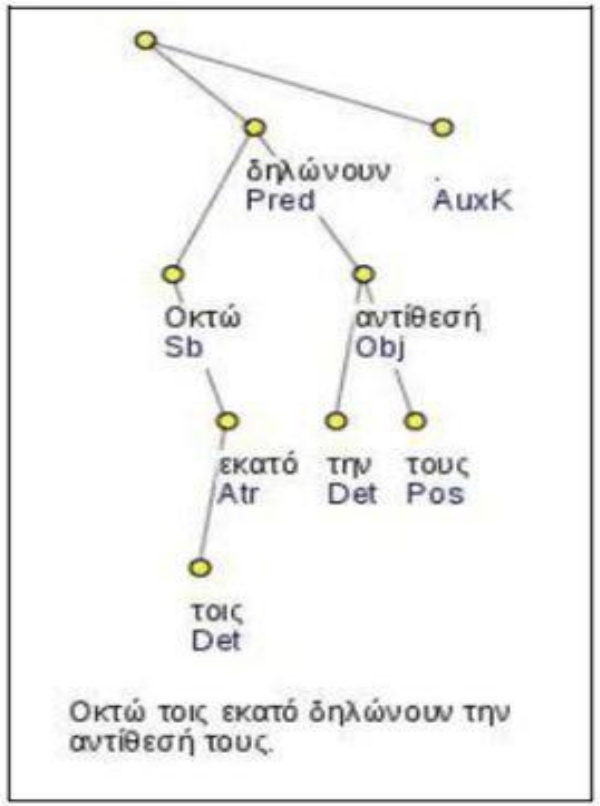
δ) **σημασιολογική επισημείωση** (semantic annotation): χαρακτηρισμός λέξεων ή προτάσεων *του* κειμένου για πτυχές της σημασίας,

ε) **κειμενική επισημείωση** (discourse annotation): χαρακτηρισμός τμημάτων του κειμένου *σε* διαπροτασιακό επίπεδο.

Γλωσσικός χαρακτηρισμός (linguistic annotation) με βάση κωδικοποίηση Penn treebank POS tagset
(σε Γούτσο & Φραγκάκη 2015: 36)

(3) Κύριοι_NNS συνάδελφοι_NNS ,_, τη_DT συνεδρίασή_NN μας_PRP$ παρακολουθούν_VBP από_IN τα_DT άνω_RB δυτικά_JJ θεωρεία_NNS,_, αφού_IN ξεναγήθηκαν_VB στην_IN έκθεση_NN της_DT αίθουσας_NN "_"ΕΛΕΥΘΕΡΙΟΣ_NNP ΒΕΝΙΖΕΛΟΣ_NNP"_" για_IN τα_DT ογδόντα_CD χρόνια_NNS ενσωμάτωσης_NN της_DT Θράκης_NNP στην_IN Ελλάδα_NNP,_, εβδομήντα_CD οκτώ_CD μαθητές_NNS και_CC πέντε_JJ συνοδοί_NNS καθηγητές_NNS τους_PRP$ από_IN το_DT 3_CD ο_JJ Γυμνάσιο_NN Ιεράπετρας_NNP Λασιθίου_NNP Κρήτης_NNP,_, καθώς_RB και_CC τριάντα_CD ένας_CD μαθητές_NNS και_CC τέσσερις_CD συνοδοί_NNS -_- δάσκαλοί_NNS τους_PRP$ από_IN το_DT Δημοτικό_JJ Σχολείο_NN Λιμένα_NNP Χερσονήσου_NNP του_DT Ηρακλείου_NNP Κρήτης_NNP._.

Συντακτική επισημείωση (σε Γούτσο & Φραγκάκη 2015: 37)



Εικόνα 2.2 *Συντακτική επισημείωση πρότασης με βάση το* Greek Dependency Treebank.

Χρησιμότητα...

Ο χρήστης μπορεί:

- να περιορίσει τις αναζητήσεις του σε ένα μόνο κειμενικό είδος (π.χ. κείμενα γνώμης) και μόνο στη δομική ενότητα «τίτλος» των κειμένων,
- να αναζητήσει τις πιο συχνές λεξικές συνάψεις ή συμφράσεις της (collocations), δηλαδή τις στατιστικά σημαντικές και όχι τις τυχαίες συνεμφανίσεις της με άλλες λέξεις (π.χ. *στρώμα θαλάσσης*), ή τις φράσεις στις οποίες χρησιμοποιείται μια λέξη (*διά θαλάσσης, πυρ, γυνή και θάλασσα*),
- να συγκρίνει τις εμφανίσεις μιας συντακτικής δομής (π.χ. ουσιαστικό + γενική) σε κείμενα επιστημονικά και σε κείμενα λογοτεχνικά κ.ο.κ.
- να διερευνήσει το λεξιλόγιο που χρησιμοποιεί ένας συγγραφέας, ή το λεξιλόγιο που χρησιμοποιείται συχνότερα σε ένα γνωστικό πεδίο.

...αλλά και προβλήματα

- Εικόνα ακαταστασίας στο κείμενο, που υπονομεύει την ακεραιότητά του (Sinclair 2004: 191).
- Το ΗΣΚ καθίσταται λιγότερο εύκολα προσβάσιμο και μακρόσυρτο (McEnery et al. 2006).
- Η επισημείωση επιβάλλει μια ορισμένη γλωσσολογική ανάλυση και εξαρτάται από το θεωρητικό σχήμα που υιοθετείται κάθε φορά (π.χ. *στην:* πρόθεση ή συνδυασμός πρόθεσης και άρθρού; / λημματοποίηση: π.χ. οι τύποι *τριάντα πέντε* και *καθηγητής-σύμβουλος* ανήκουν σε ένα ή σε δύο λήμματα;).

Εν κατακλείδι...

- Αντίθετα, ο Leech (2005) πιστεύει ότι η επισημείωση προσθέτει αξία στο σώμα κειμένων.
- Οι McEnery και Wilson (1996: 32) σημειώνουν ότι με την επισημείωση οι γλωσσικές πληροφορίες που περιέχει ένα σώμα κειμένων γίνονται ρητές και η ανάκτηση και ανάλυσή τους πολύ πιο γρήγορες και εύκολες.

Εν κατακλείδι...

- Ο Leech (1997) διατυπώνει επτά αρχές που θα πρέπει να τηρούνται κατά την επισημείωση:

1. Θα πρέπει να είναι δυνατό να αφαιρείται η επισημείωση από ένα επισημειωμένο σώμα κειμένων ώστε να προκύπτει η απλή εκδοχή του.
2. Θα πρέπει να είναι δυνατό να εξάγονται οι ίδιες οι επισημειώσεις από το κείμενο.
3. Το σχήμα επισημείωσης θα πρέπει να βασίζεται σε κατευθυντήριες οδηγίες που να είναι διαθέσιμες στον χρήστη του σώματος κειμένων.
4. Θα πρέπει να είναι σαφές ποιος και πώς πραγματοποίησε την επισημείωση.
5. Θα πρέπει να είναι σαφές στον χρήστη του σώματος κειμένων ότι η επισημείωση δεν είναι αλάθητη, αλλά αποτελεί απλώς ένα χρήσιμο εργαλείο.
6. Τα σχήματα επισημείωσης θα πρέπει να βασίζονται κατά το δυνατόν σε ευρέως αποδεκτές και θεωρητικά ουδέτερες αρχές.
7. Κανένα σχήμα επισημείωσης δεν θα πρέπει να θεωρείται εκ των προτέρων ως πρότυπο. Τα πρότυπα θα πρέπει να προσαρμόζονται στα νέα δεδομένα που προκύπτουν από την ερευνητική πράξη και να μην προσκολλώνται σε έτοιμα θεωρητικά μοντέλα.

Διαθεσιμότητα

- Τα γενικά σώματα κειμένων είναι ελεύθερα διαθέσιμα, π.χ. μέσω διεπαφής στο διαδίκτυο.
- Η ελεύθερη και αυτούσια διάθεση του υλικού των σωμάτων κειμένων προσκρούει στα πνευματικά δικαιώματα των κειμένων που περιλαμβάνονται σε αυτά.
- Για όσα κείμενα προέρχονται από ιδιωτικές πηγές είναι απαραίτητο να υπάρχει έγγραφη άδεια των δημιουργών ή των εκδοτών.
- Σε περιπτώσεις μεταγραφής προφορικών δεδομένων υπάρχει η δεοντολογία τήρησης προσωπικών δεδομένων, οπότε ζητείται η συγκατάθεση των συμμετεχόντων και απαλείφονται, μέσω μιας διαδικασίας ανωνυμοποίησης, όσα στοιχεία ανήκουν στα προσωπικά τους δεδομένα.
- Οι περιορισμοί που συνδέονται με τα πνευματικά δικαιώματα εξαρτώνται από τη νομοθεσία κάθε χώρας και σε γενικές γραμμές μπορούν να παρακαμφθούν αν δεν διατεθεί αυτούσιο το γλωσσικό υλικό (κείμενα σε συνεχόμενη μορφή) στην ευρύτερη κοινότητα.

ΗΣΚ από την έρευνα στο μαθητή

- Τα ΗΣΚ γλωσσικής εκμάθησης είναι συλλογές κειμένων, γραπτών ή προφορικών, που έχουν παραχθεί από τους ίδιους τους μαθητές της ξένης γλώσσας

Είναι:

α. ηλεκτρονικά (και άρα εύκολα στην ανάλυση)

β. μεγάλα (και άρα αξιόπιστα)

Αυθεντικά;

- Τα ΗΣΚ Γλωσσικής Εκμάθησης αποτελούν πηγή αυθεντικού λόγου, μοναδική απόδειξη της **διαγλώσσας** των μαθητών κατά τη διαδικασία της εκμάθησης.

Διαγλώσσα;

- Η γλώσσα του μαθητή της Ξένης Γλώσσας διαφέρει από αυτήν του φυσικού ομιλητή. Τόσο **ποσοτικά** όσο και **ποιοτικά** έχει διαφορετική συχνότητα λέξεων, φράσεων και δομών, άλλα στοιχεία δεν χρησιμοποιούνται και άλλα χρησιμοποιούνται υπερβολικά.
- Ένα ΗΣΚ γλωσσικής εκμάθησης χαρακτηρίζεται από **λάθη** (αποκλίσεις) που βοηθούν στην κατανόηση της διαδικασίας της μάθησης

Εφαρμογές των ΗΣΚ Γλωσσική έρευνα Λεξικογραφία Γραμματική

- Γλωσσική έρευνα
- Λεξικογραφία
- Μετάφραση (παράλληλα ΗΣΚ)
- Ορολογία
- Ιστορία της γλώσσας
- Διδακτική

Διδακτική και ΗΣΚ Εφαρμογή στη διδασκαλία

- **Σε χαρακτηρισμένο ΗΣΚ**, όταν ο διδάσκων παίρνει παραδείγματα από το ΗΣΚ και τα επεξεργάζεται ως άσκηση με τους μαθητές.
- **Σε μη-χαρακτηρισμένο ΗΣΚ**, όταν οι μαθητές ψάχνουν να βρουν τη χρήση μιας λέξης μόνοι τους. (Δεδομενοκεντρικής εκμάθησης/DDL: Data Driven Learning).
- **Σε ΗΣΚ Γλωσσικής εκμάθησης** (learner corpus) ο ένας μαθητής προσπαθεί να εξηγήσει στον άλλον τα λάθη του («αμοιβαία μάθηση»).

Διδακτική και ΗΣΚ Εφαρμογή στην αξιολόγηση

- Ως αρχείο γραπτών προερχόμενων από εξετάσεις
- Ως εργαλείο για να αναπτυχθεί υλικό αξιολόγησης
- Για να βελτιωθούν οι τεχνικές αξιολόγησης
- Για να βελτιωθεί η ποιότητα της βαθμολόγησης
- Για να σταθεροποιηθεί η μορφή των τεστ.

Τι δεν είναι τα ΗΣΚ για την διδακτική:

- Μέθοδος
- Αντικαταστάτης του εκπαιδευτικού

ΑΛΛΑ:

- Είναι εργαλείο, διευκολυντής, βοηθός…

Ζητήματα παιδαγωγικής μεθόδευσης της γλωσσικής διδασκαλίας

- Στη δυναμική εκδοχή της αξιοποίησης των ΗΣΚ, οι ίδιοι οι μαθητές/ σπουδαστές χρησιμοποιούν τον υπολογιστή, για να έχουν άμεση πρόσβαση στο υλικό τους και να πραγματοποιούν αναζητήσεις.
- Σε μια πιο συντηρητική εκδοχή, άμεση πρόσβαση έχει μόνο ο δάσκαλος, ο οποίος τα αξιοποιεί, για να προετοιμάσει το υλικό που θα φέρει στην τάξη
- η αξιοποίησή τους για την υλοποίηση δραστηριοτήτων εξερευνητικής μάθησης (discovery ή exploratory learning) που δίνουν στον διδασκόμενο την ευκαιρία να παίξει το **ρόλο του ερευνητή** ή ακόμη και του **εξερευνητή της γλώσσας** (linguistic explorer, Bernardini, 2000, 2002, 2004)
- μαθητής ως ερευνητής (ή εξερευνητής) της γλώσσας είτε εμπλέκεται σε μια **επαγωγική διαδικασία**, δηλαδή εξετάζει πλήθος παραδειγμάτων που αντλεί από ένα corpus, για να ανακαλύψει τις κανονικότητες της γλωσσικής χρήσης και τους κανόνες του γλωσσικού συστήματος, είτε σε **παραγωγική** (deductive) διαδικασία στο πλαίσιο της οποίας διαμορφώνει υποθέσεις και στη συνέχεια ελέγχει την ισχύ τους, αξιοποιώντας τα δεδομένα του σώματος κειμένων

Ζητήματα παιδαγωγικής μεθόδευσης της γλωσσικής διδασκαλίας

- Η χρήση των ΗΣΚ εντάσσεται στον ευρύτερο προβληματισμό σχετικά με την υποστηριζόμενη από την τεχνολογία διδασκαλία και εκμάθηση γλωσσών (Computer Assisted Language Learning, CALL).
- Οι παιδαγωγικού χαρακτήρα αλλαγές που συνεπάγεται η αξιοποίηση των ΗΣΚ στη διδασκαλία συνοψίζονται στα εξής:

1. αναθεώρηση του ρόλου και του τρόπου δράσης των μαθητών/σπουδαστών που αναλαμβάνουν να εξερευνήσουν πτυχές της γλώσσας,
2. αναπροσαρμογή του ρόλου του δασκάλου που μετατρέπεται σε υποστηρικτή της έρευνας,
3. αλλαγή του περιεχομένου της διδασκαλίας, η οποία δεν είναι πια μετάδοση της γνώσης, αλλά διαμεσολαβητική διαδικασία που υποστηρίζει ανακαλυπτικές δραστηριότητες των μαθητών

Μειονεκτήματα

- Δεν μπορούν να χρησιμοποιηθούν για όλα τα επίπεδα και όλες τις ανάγκες
- Τα ΗΣΚ που είναι βασισμένα σε ενημερωτικά κείμενα δεν είναι κατάλληλα για μικρά επίπεδα
- Το υλικό μπορεί να φανεί δυσνόητο ή πολύπλοκο στο μαθητή και η έκτασή του να τον τρομάζει
- Το ΗΣΚ δεν δίνει πληροφορίες για τον επιτονισμό, χειρονομίες κ.λ.π. των ομιλητών και για το επικοινωνιακό περιβάλλον.
- Ο δάσκαλος έχει μειωμένο έλεγχο στη μαθησιακή διαδικασία.
- Μπορεί να δημιουργηθούν από το μαθητή περισσότερες ερωτήσεις και απορίες, γιατί εκτίθεται σε έναν πολύ μεγαλύτερο αριθμό γλωσσικών παραδειγμάτων
- Η επιμόρφωση των εκπαιδευτικών δεν θεωρείται πάντα δεδομένη, όπως και ο ανάλογος εξοπλισμός στο σχολικό περιβάλλον

Βιβλιογραφικές αναφορές

- Atkins, B. T. S. and Rundell, M. (2008) The Oxford Guide to Practical Lexicography. Oxford: Oxford University Press.
- Baker, P., Hardie , A. & McEnery, T. (2006). A Glossary of Corpus Linguistics. Edinburgh University Press.
- Barnbrook, G. (1996). Language and Computers. Edinburgh: Edinburgh University Press.
- Burnard, L. (2002). Validation Manual for Written Language Resources. Available at: http://projects.oucs.ox.ac.uk/elra/D1.xml?ID=body.1_div.1 [Accessed 5 March 2015].
- Γούτσος, Δ. & Φραγκάκη, Γ. (2015). Εισαγωγή στη Γλωσσολογία Σωμάτων Κειμένων. Ελληνικά ακαδημαϊκά ηλεκτρονικά συγγράμματα και βοηθήματα.
- Kučera, K. (2002). The Czech National Corpus: Principles, design and results. Literary and Linguistic Computing 17 (2): 245-257.
- Leech, G. (1991) The State of the Art in Corpus Linguistics. In K. Aijmer and B. Altenberg (eds), English Corpus Linguistics. London: Longman, 8–30.
- Leech, G. (1997). Introducing corpus annotation. In R. Garside, G. Leech & A. McEnery (eds) Corpus Annotation. London: Longman, 1-18.

Βιβλιογραφικές αναφορές

- Leech, G. (2005). Adding linguistic annotation. In M. Wynne (ed.) Developing Linguistic Corpora: A Guide to Good Practice. Oxford: Oxbow Books, 17-29.
- McEnery, T. & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice. Cambridge: Cambridge University Press.*
- McEnery, T. & Wilson, A. (1996). *Corpus Linguistics: An Introduction. Edinburgh: Edinburgh University* Press.
- McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus Based Language Studies. An Advanced Resource Book.* London: Routledge.
- Sinclair, J. (1991). *Corpus, Concordance, Collocation. Oxford: Oxford University Press.*
- Sinclair, J. (2004). *Trust the Text: Language, Corpus and Discourse. London: Routledge.*
- Sinclair, J. (2005). Corpus and Text. Basic Principles. In M. Wynne (ed.) *Developing Linguistic Corpora: A Guide to Good Practice. Oxford: Oxbow Books,* 1-16.
- Sinclair, J. (2008). Borrowed Ideas. In A. Gerbig & O. Mason (eds) *Language, People, Numbers. Corpus Linguistics and Society. Amsterdam: Rodopi,* 21-42.
- Stubbs, M. (2001) *Words and Phrases: Corpus Studies of Lexical Semantics*. New York: Blackwell.


THANK YOU
FOR YOUR ATTENTION