



Π2.2 Αναφορά Τεκμηρίωσης

DATE: 07/10/2019

ΔΡΑΣΗ ΕΘΝΙΚΗΣ ΕΜΒΕΛΕΙΑΣ «ΕΡΕΥΝΩ-ΔΗΜΙΟΥΡΓΩ- ΚΑΙΝΟΤΟΜΩ»

T1EΔK-05094 - Λεξιπαίγνιο

Ανάπτυξη προηγμένου, καινοτόμου και state-of-the-art υπολογιστικού περιβάλλοντος, για τη δημιουργία ψηφιακών εκπαιδευτικών παιχνιδιών που απευθύνονται σε μαθητές με στόχο: α) τη βελτίωση του γλωσσικού και γενικότερα μαθησιακού τους επιπέδου και β) την ανάπτυξη ποικίλων δεξιοτήτων τους σχετικά με το λεξιλόγιο και τη γλώσσα γενικότερα, αλλά και το περιεχόμενο ειδικών γνωστικών αντικειμένων (π.χ. γεωλογία-γεωγραφία, βιολογία κ.ά.).



ΗΜΕΡΟΜΗΝΙΑ ΕΝΑΡΞΗΣ ΕΡΓΟΥ	09/07/2018
ΔΙΑΡΚΕΙΑ ΕΡΓΟΥ	34 Μήνες



ΕΠΑνεΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ
ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ
ΚΑΙΝΟΤΟΜΙΑ



Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Π2.2 Αναφορά Τεκμηρίωσης	
ΑΡΧΙΚΗ ΗΜΕΡΟΜΗΝΙΑ ΠΑΡΑΔΟΣΗΣ:	08/2019
ΠΡΑΓΜΑΤΙΚΗ ΗΜΕΡΟΜΗΝΙΑ ΠΑΡΑΔΟΣΗΣ:	10/2019
ΕΙΔΟΣ ΠΑΡΑΔΟΤΕΟΥ:	Έκθεση
ΣΥΝΤΑΚΤΗΣ:	Α. Βαγγελάτος
ΕΚΔΟΣΗ ΕΓΓΡΑΦΟΥ:	ν. 1.4

ΣΥΝΤΕΛΕΣΤΕΣ ΠΑΡΑΔΟΤΕΟΥ	
Όνομα	Φορέας
Άννα Ιορδανίδου	ITYE
Μόνικα Γαβριηλίδου	ITYE
Μαρία Φουντάνα	ITYE
Αριστείδης Βαγγελάτος	ITYE
Νίκος Αλεξόπουλος	ITYE
Βαρβάρα Γιαννούλη	ITYE
Χρήστος Τσαλίδης	NEUROCOM
Βασίλης Σπιταδάκης	NEUROCOM
Ιωάννης Σταματοπούλος	NEUROCOM
Γιώργος Παναγιωτάκης	NEUROCOM
Μαρία Φωστιέρη	NEUROCOM
Γιάννης Καφούσης	NEUROCOM
Αναστάσιος Καλαενζτής	NEUROCOM
Κωνσταντίνος Αλεξάκης	NEUROCOM
Παναγιώτης Γάκης	NEUROCOM

«Το παραδοτέο υλοποιήθηκε στο πλαίσιο της Δράσης ΕΡΕΥΝΩ – ΔΗΜΙΟΥΡΓΩ - ΚΑΙΝΟΤΟΜΩ και συγχρηματοδοτήθηκε από την Ευρωπαϊκή Ένωση και εθνικούς πόρους μέσω του Ε.Π. Ανταγωνιστικότητα, Επιχειρηματικότητα & Καινοτομία (ΕΠΑνΕΚ) (κωδικός έργου: Τ1ΕΔΚ-05194)»

Πίνακας Περιεχομένων

Εισαγωγή	4
1. Αναγνώριση όρων και απαραίτητες υποδομές	5
2. Αρχιτεκτονική του συστήματος.....	5
2.1 Σύνοψη Αρχιτεκτονικής.....	5
2.2 Επεξεργασία φυσικής γλώσσας - Μηχανική μάθηση	6
2.3 Εργαλεία υποστήριξης χρηστών.....	7
3. Γλωσσικό Υπόβαθρο	9
3.1 Μεθοδολογία ανάλυσης κειμένων για εντοπισμό όρων	9
3.2 Επιμέλεια Όρων	10
Παράρτημα Ι: Περιγραφή σώματος εκπαιδευτικών κειμένων	12
Βιβλίο Βιολογίας Α΄ Γυμνασίου	14
Βιβλίο Οικιακής Οικονομίας Α΄ Γυμνασίου.....	15
Φάκελος Μαθήματος - Προσωρινό Έντυπο Υλικό στα Θρησκευτικά Α΄ Γυμνασίου	16
Βιβλίο Μουσικής Α΄ Γυμνασίου.....	16
Παράρτημα ΙΙ – Υλικό από άλλες πηγές («Ιφιγένεια», «Φωτόδεντρο»)	17
Φωτόδεντρο – Εκπαιδευτικές δραστηριότητες που παρουσιάζουν ενδιαφέρον:.....	17
Ιφιγένεια: Εκπαιδευτικές δραστηριότητες που παρουσιάζουν ενδιαφέρον:.....	18
Ιφιγένεια: Κειμενική Πληροφορία από Εκπαιδευτικές δραστηριότητες που είναι αξιοποιήσιμη στο corpus της Γεωγραφίας:.....	18
Παράρτημα ΙΙΙ: Θεωρητική αναφορά στα Σώματα Κειμένων	20
Ορισμός του όρου «σώμα κειμένων»	20
Είδη σωμάτων κειμένων	20
Τα σώματα κειμένων στη γλωσσολογική έρευνα.....	21
Βασικά χαρακτηριστικά της γλωσσολογικής ανάλυσης που βασίζεται σε σώματα κειμένων	22
Η σύγχρονη εποχή: τομείς εφαρμογής.....	23
Τα σώματα κειμένων στην κατασκευή οντολογίας	25
Βιβλιογραφία.....	26

Παράρτημα IV – Μονολεκτικοί και Πολυλεκτικοί όροι.....	29
Παράρτημα V – Στατιστικά Στοιχεία Όρων.....	29

Εισαγωγή

Το έργο ΛΕΞΠΑΙΓΝΙΟ, είναι ένα ερευνητικό έργο, ενταγμένο στη Δράση ΕΡΕΥΝΩ – ΔΗΜΙΟΥΡΓΩ - ΚΑΙΝΟΤΟΜΩ που συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση και εθνικούς πόρους μέσω του Ε.Π. «Ανταγωνιστικότητα, Επιχειρηματικότητα & Καινοτομία» (ΕΠΑνΕΚ) (κωδικός έργου:Τ1ΕΔK-05194). Για την υλοποίησή του συνεργάζονται δύο φορείς: το ερευνητικό Ινστιτούτο Τεχνολογίας Υπολογιστών και Εκδόσεων (ITYE) και η εταιρεία Neurocom A.E.

Στόχος του έργου είναι η Ανάπτυξη προηγμένου, καινοτόμου και state-of-the-art υπολογιστικού περιβάλλοντος, για τη δημιουργία ψηφιακών εκπαιδευτικών παιχνιδιών που απευθύνονται σε μαθητές με σκοπό: α) τη βελτίωση του γλωσσικού και γενικότερα μαθησιακού τους επιπέδου και β) την ανάπτυξη ποικίλων δεξιοτήτων τους σχετικά με το λεξιλόγιο και τη γλώσσα γενικότερα, αλλά και το περιεχόμενο ειδικών γνωστικών αντικειμένων (π.χ. γεωλογία-γεωγραφία, βιολογία, κ.α.).

Το παρόν παραδοτέο, Π2.2: *Αναφορά Τεκμηρίωσης* τεκμηριώνει όλες τις εργασίες που έλαβαν χώρα στην ΕΕ2, από την συγκρότηση του corpus, έως την αναγνώριση των όρων (μονολεκτικών και πολυλεκτικών). Είναι συμπληρωματικό του παραδοτέου Π2.1: *Περιγραφή συντακτικών σχημάτων για αναγνώριση όρων με το φορμαλισμό «Κανών»*) παραθέτοντας την απαραίτητη περιγραφή και τεκμηρίωση διαδικασιών που υλοποιήθηκαν και υποδομών που χρησιμοποιήθηκαν.

Στα παρακάτω κεφάλαια, σύμφωνα με τα προβλεπόμενα στην περιγραφή των παραδοτέων του έργου, παρουσιάζονται: μια εισαγωγή στην αναγνώριση των όρων, η αρχιτεκτονική του συστήματος και το γλωσσικό υπόβαθρο που απαιτήθηκε.

Στα Παραρτήματα, περιλαμβάνονται:

Παράρτημα Ι: Περιγραφή του Σώματος Κειμένων (Σ.Κ.)

Παράρτημα ΙΙ: Υλικό από άλλες πηγές και

Παράρτημα ΙΙΙ: Θεωρητική τεκμηρίωση των Σ.Κ.

Παράρτημα ΙV: Πολυλεκτικοί και Μονολεκτικοί όροι.

Παράρτημα V: Στατιστικά στοιχεία όρων

1. Αναγνώριση όρων και απαραίτητες υποδομές

Η ανάγκη για αναγνώριση των όρων στο πλαίσιο του έργου ΛΕΞΙΠΑΙΓΝΙΟ και συγκεκριμένα για το μάθημα της «Γεωγραφίας – Γεωλογίας» απαίτησε την ύπαρξη μιας σειράς υποδομών που είτε διατέθηκαν από τους φορείς που συνεργάζονται στο πλαίσιο του έργου, είτε αναπτύχθηκαν στο πλαίσιο του ίδιου του έργου. Στη συνέχεια περιγράφονται τόσο οι υποδομές όσο και η διαδικασία που υλοποιήθηκε.

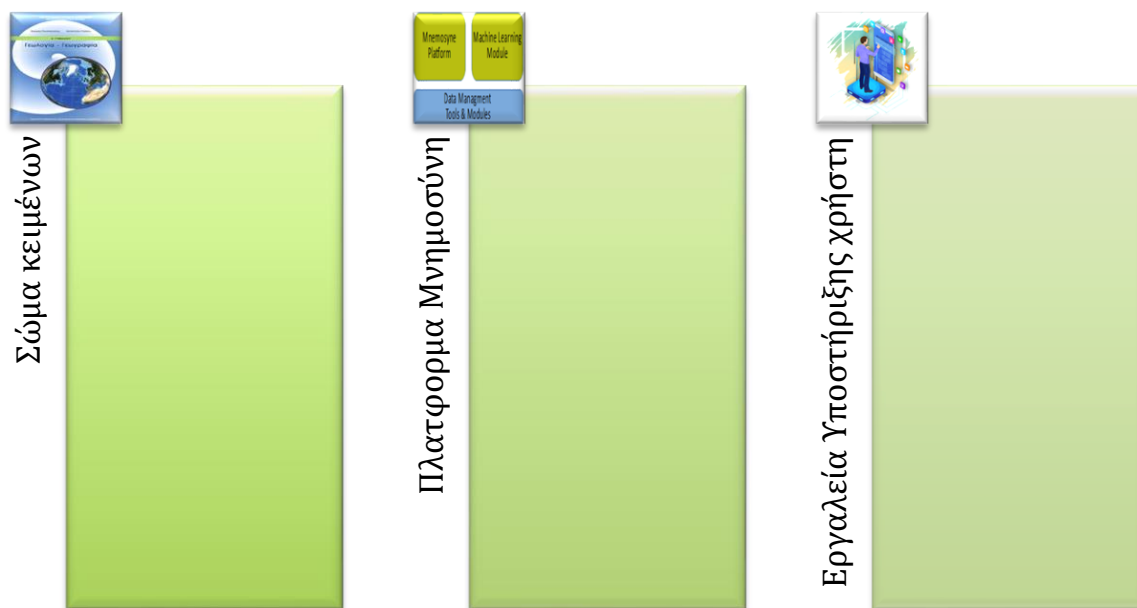
Τα βασικά συστατικά μέρη της υποδομής περιλαμβάνουν:

- Το σύστημα επεξεργασίας φυσικής γλώσσας,
- Τους αλγορίθμους μηχανικής μάθησης που ενισχύουν την ανάλυση,
- Το γλωσσικό υπόβαθρο,
- Τις εφαρμογές χρήσης από διαφορετικές κατηγορίες χρηστών,

Η περιγραφή των προαναφερθέντων συστατικών επικουρείται και από τις μεταξύ τους συνδέσεις, με τρόπο ώστε να συγκροτούν ένα ενιαίο σύστημα. Ιδιαίτερη έμφαση δίνεται στην περιγραφή του γλωσσικού υποβάθρου, καθώς αυτό καθορίζει σε μεγάλο βαθμό την ποιότητα των τελικών αποτελεσμάτων ως προς την ακρίβεια εντοπισμού ειδικά των όρων (στο πλαίσιο της ΕΕ2). Η μεθοδολογία της γλωσσικής ανάλυσης και τα αποτελέσματα (ταξινομία, κανόνες, λεξικά) περιγράφονται και με παραδείγματα και οθόνες (screenshots) αποτύπωσης της διαδικασίας.

2. Αρχιτεκτονική του συστήματος

2.1 Σύνοψη Αρχιτεκτονικής



Εικόνα 1: Αρχιτεκτονική συστήματος Επεξεργασίας CORPUS (Λεξιπαίγνιο)

Στην παραπάνω εικόνα παρουσιάζεται συνοπτικά η αρχιτεκτονική του συστήματος ανάλυσης του σώματος κειμένων για το Λεξιπαίγνιο. Αυτή απαρτίζεται από ξεχωριστά συστατικά μέρη (modules) που διεπιδρούν, συνεργάζονται και ανταλλάσσουν δεδομένα μεταξύ τους. Καθένα από αυτά εκτελεί μια καθορισμένη λειτουργία με είσοδο και έξοδο συγκεκριμένων τύπων δεδομένων.

Τα συστατικά μέρη της πλατφόρμας διακρίνονται στις ακόλουθες τρεις (3) γενικές ενότητες:

- **Σώμα Κειμένων** – Το σώμα κειμένων που συγκροτήθηκε περιγράφεται στο παραδοτέο Π2.1 ενώ στο Παράρτημα I και II του τρέχοντος Παραδοτέου, αναφέρονται τα περιεχόμενά του ενώ στο Παράρτημα III μια θεωρητική αναφορά στα Σώματα Κειμένων γενικότερα
- **Mnemosyne** – αποτελεί την πλατφόρμα / υποδομές επεξεργασίας των δεδομένων
- **Εργαλεία** – εφαρμογές υποστήριξης του χρήστη.

Η «Πλατφόρμα Μνημοσύνη» (**Mnemosyne Platform**) αποτελεί το συστατικό μέρος στο οποίο βασίζεται η κυρίως επεξεργασία της πληροφορίας (σώμα κειμένων). Η ανάπτυξή της στηρίχθηκε στο σύστημα Mnemosyne™, που έχει αναπτύξει η Neurocom, μια αρθρωτή πλατφόρμα Επεξεργασίας Φυσικής Γλώσσας και ανάλυσης κειμένων (text analytics) με δυνατότητα ευέλικτης προσαρμογής σε οποιοδήποτε περιβάλλον επεξεργασίας πληροφορίας, χάρη στα ενσωματωμένα υποσυστήματα σημασιολογικής ανάλυσης και εξαγωγής πληροφορίας (semantic information extraction), επεξεργασίας και ανάλυσης γνώμης (sentiment analysis), αποδελτίωσης και οργάνωσης πληροφορίας. Η πλατφόρμα αυτή, τροφοδοτήθηκε με τους απαραίτητους κανόνες (Π2.1) με βάση τον ειδικό φορμαλισμό «Κανών». Ο φορμαλισμός «Κανών», εκτός από τη δυνατότητα ορισμού μορφοσυντακτικών σχημάτων, ευαίσθητων περιεχομένου (context sensitive constructs) που λαμβάνουν υπόψη το συμφραστικό τους περιβάλλον, έχει τη δυνατότητα να αξιοποιεί επιμέρους χαρακτηριστικά των λέξεων (μορφοσυντακτικά, υφολογικά κτλ.).

Τα συστατικά μέρη και τα εργαλεία υποστήριξης των χρηστών αξιοποιούνται για την διευκόλυνση των χρηστών στις διαδικασίες ελέγχου που υλοποιούν. Ένα από τα εργαλεία είναι η αναζήτηση όρων και η παρουσίαση των συμφραζομένων τους (concordancer). Σημαντικές υποδομές που υποστηρίζουν αυτά είναι: το KNIME Analytics Platform (<https://www.knime.com/>), κυρίως, για ανάγκες επεξεργασίας τύπου ETL (Extract, Transform, Load) και το Apache Jena (<https://jena.apache.org/>) για τη διαχείριση της ταξινόμιας.

2.2 Επεξεργασία φυσικής γλώσσας - Μηχανική μάθηση

Η υλοποίηση του έργου στηρίζεται στη χρήση και προσαρμογή του υπάρχοντος συστήματος επεξεργασίας φυσικής γλώσσας (ΕΦΓ) Mnemosyne™, πλαισιούμενο από καινοτόμους αλγόριθμους σημασιολογικής ανάλυσης και βαθιάς μάθησης, οι οποίοι εμπλουτίστηκαν περαιτέρω από τους υπάρχοντες γραμματικούς και σημασιολογικούς κανόνες που ετοιμάστηκαν για το μάθημα της Γεωγραφίας - Γεωλογίας και τα εκφραστικά σχήματα που διερευνήθηκαν στην συλλογή κειμένων (corpus) που συγκροτήθηκε.

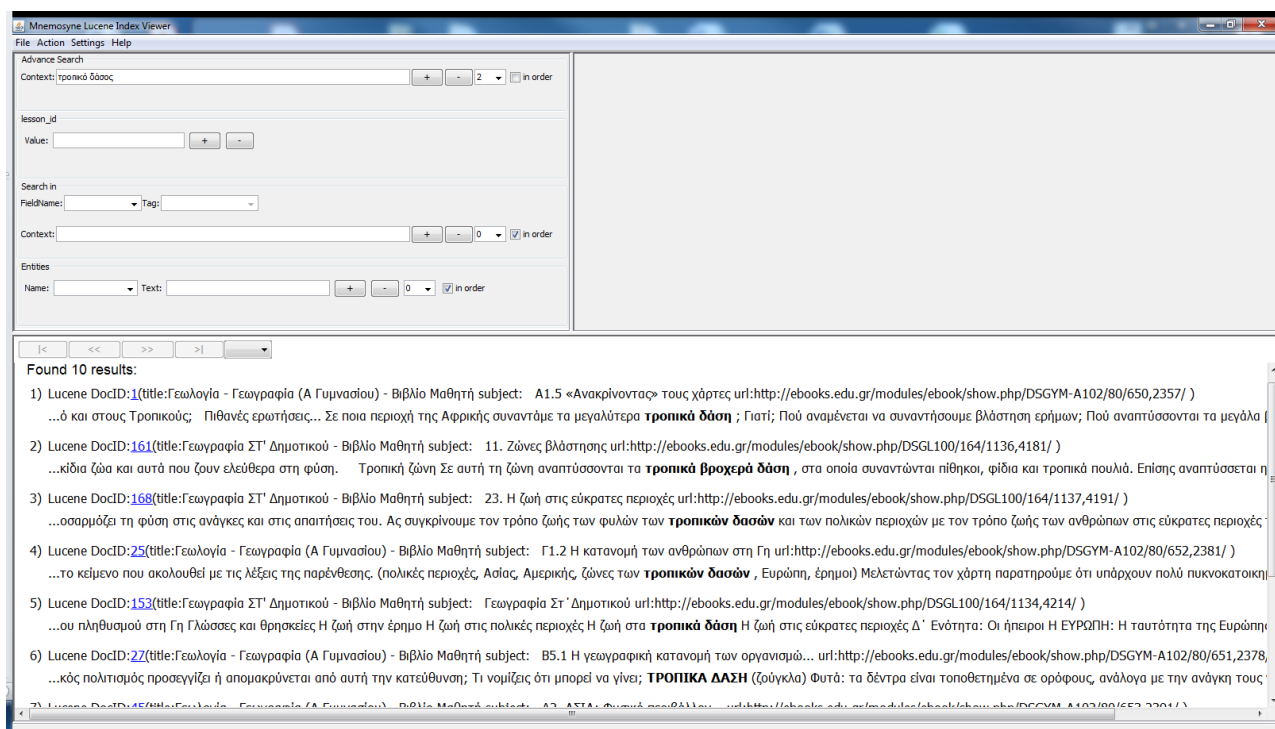
Ο φορμαλισμός «ΚΑΝΩΝ» ανήκει στην κατηγορία των «Ενοποιητικών Γραμματικών» (Unification Grammars). Σε συνδυασμό με τη χρήση ενός πρότυπου «Αλγόριθμου Ανάλυσης» (Resolution Algorithm), επιτρέπει την αποτύπωση σύνθετων συντακτικών γλωσσικών δομών βάσει των οποίων γίνεται η σημασιολογική αποσαφήνιση και εξαγωγή πληροφορίας από κείμενα (Nissim Francez, Shuly Wintner, «Unification Grammars», Cambridge University Press, 2012).

Η επεξεργασία φυσικής γλώσσας υποστηρίζεται από λειτουργίες και μοντέλα μηχανικής μάθησης. Στο βασικό επίπεδο εφαρμογής των παραδοσιακών αλγορίθμων μηχανικής μάθησης σε συνδυασμό με τους κανόνες που έχουν καθοριστεί έγινε η αναγνώριση των όρων (μονολεκτικών και πολυλεκτικών) που

εμφανίζονται στο σώμα κειμένων (βλέπε αναλυτικά παραδοτέο Π2.1). Οι όροι είναι σύνολα λέξεων που αντιστοιχούν σε έννοιες και πραγματώνονται γλωσσικά με συγκεκριμένες μορφοσυντακτικές δομές που αναγνωρίζει το σύστημα. Στην περίπτωση υποψήφιων όρων οι αλγόριθμοι μηχανικής μάθησης καθορίζουν αν είναι όροι και πώς θα χρησιμοποιηθούν. Στο τέλος, ακολουθεί ένας χειρονακτικός έλεγχος (από τους φιλόλογους/γλωσσολόγους).

2.3 Εργαλεία υποστήριξης χρηστών

Για την υποστήριξη των χρηστών, ειδικά στη διαδικασία αξιολόγησης / επιμέλειας των όρων, προσαρμόστηκε στις ανάγκες του έργου μια εφαρμογή εμφάνισης συμφραζομένων (concordance) των όρων μέσα στο Σώμα Κειμένων (ΣΚ). Πιο συγκεκριμένα η εφαρμογή δίνει στο χρήστη τη δυνατότητα να αναζητήσει έναν όρο (μονολεκτικό ή πολυλεκτικό) και επιστρέφει τις εμφανίσεις του όρου με τα συμφραζομένα τους καθώς και στοιχεία για το πού βρέθηκε (αλλά και link στο βιβλίο όπου βρέθηκε) – βλ. παρακάτω σχήμα.



Στο παραπάνω παράδειγμα ο χρήστης έχει αναζητήσει τον όρο «τροπικό δάσος» και η εφαρμογή έχει επιστρέψει δέκα (10) αποτελέσματα με τα συμφραζομένα τους. Υπάρχει ακόμα ο σύνδεσμός που αν τον ακολουθήσει ο χρήστης, ανοίγει το ψηφιακό βιβλίο που αντιστοιχεί στην εμφάνιση, που είναι αναρτημένο στο «Φωτόδεντρο» (βλ. επόμενη εικόνα). Έτσι ο χρήστης μπορεί να εκτιμήσει την χρήση του όρου, τόσο στο πλαίσιο μιας φράσης, όσο και στο πλαίσιο του ίδιου του βιβλίου – κεφαλαίου – παραγράφου που ανευρέθηκε.

Γεωγραφία (ΣΤ Δημοτικού) - Βιβλίο Μαθητή (Εμπλουτισμένο)

11 Ζώνες βλάστησης

Εικόνα 11.2: Τούνδρα



Εύκρατη ζώνη

Στην εύκρατη ζώνη αναπτύσσονται οι εξής φυτικές διατάξεις: δάση από κωνοφόρα δέντρα (πεύκα, έλατα) που αποτελούν την **τάιγκα**. Την τάιγκα συναντάμε αμέσως μετά την τούνδρα, καθώς και σε περιοχές με μεγάλο υψόμετρο. Επίσης αναπτύσσονται δάση από φυλλοβόλα δέντρα (οξές, βελανιδιές), η μεσογειακή βλάστηση (ελιές, αμπέλια) και λιβάδια με ξηρό κλίμα που ονομάζονται **στέπες**. Στη ζώνη αυτή συναντάμε μεγάλη ποικιλία ζωικών οργανισμών. Είναι τα κατοικίδια ζώα και αυτά που ζουν ελεύθερα στη φύση.

Εικόνα 11.3: Βλάστηση στην εύκρατη ζώνη



Εικόνα 11.4: Σαβάννα

Τροπική ζώνη

Σε αυτή τη ζώνη αναπτύσσονται τα τροπικά βροχόρα δάση, στα οποία συναντώνται πίθηκοι, φίδια και τροπικά πουλιά. Επίσης αναπτύσσεται η **σαβάννα**, που αποτελείται από φοίνικες, κάκτους και φυλλοβόλα δέντρα με μικρά φύλλα και αγκάθια. Εδώ ζουν ελέφαντες, ζέβρες, τίγρεις, λιοντάρια κ.λπ. Ακόμη αναπτύσσεται η **στέπα** (όπως και στην εύκρατη ζώνη), η οποία αποτελείται από μικρούς θάμνους και μικρά δέντρα. Στις στέπες ζουν χορτοφάγα ζώα (άλογο, καμήλα), τρωκτικά (χάμστερ, κουνέλια), σαύρες και φίδια.

Στην ίδια εφαρμογή, αν ο χρήστης επιλέξει τον σύνδεσμο που υπάρχει στο αριθμητικό που χαρακτηρίζει τον όρο, ανοίγει ένα παράθυρο με το κείμενο στο οποίο περιλαμβάνεται ο όρος.

Lucene DocId:161(title:Γεωγραφία ΣΤ' Δημοτικού - Βιβλίο Μαθητή subject: 11. Ζώνες βλάστησης url:http://ebooks.edu...

Εύκρατη ζώνη

Στην **εύκρατη ζώνη** αναπτύσσονται οι εξής **φυτικές διατάξεις**: δάση από **κωνοφόρα δέντρα** (πεύκα, έλατα) που αποτελούν την **τάιγκα**. Την τάιγκα συναντάμε αμέσως μετά την τούνδρα, καθώς και σε περιοχές με **μεγάλο υψόμετρο**. Επίσης αναπτύσσονται δάση από **φυλλοβόλα δέντρα** (οξές, βελανιδιές), η **μεσογειακή βλάστηση** (ελιές, αμπέλια) και λιβάδια με **ξηρό κλίμα** που ονομάζονται **στέπες**. Στη ζώνη αυτή συναντάμε **μεγάλη ποικιλία ζωικών οργανισμών**. Είναι τα **κατοικίδια ζώα** και αυτά που ζουν ελεύθερα στη φύση. Σε αυτή τη ζώνη αναπτύσσονται τα **τροπικά βροχόρα δάση**, στα οποία συναντώνται πίθηκοι, φίδια και **τροπικά πουλιά**. Επίσης αναπτύσσεται η **σαβάννα**, που αποτελείται από φοίνικες, κάκτους και **φυλλοβόλα δέντρα** με **μικρά φύλλα** και αγκάθια. Εδώ ζουν ελέφαντες, ζέβρες, τίγρεις, λιοντάρια κ.λπ. Ακόμη αναπτύσσεται η **στέπα** (όπως και στην **εύκρατη ζώνη**), η οποία αποτελείται από **μικρούς θάμνους** και **μικρά δέντρα**. Στις στέπες ζουν **χορτοφάγα ζώα** (άλογο, καμήλα), τρωκτικά (χάμστερ, κουνέλια), σαύρες και φίδια. Ο άνθρωπος καλλιεργεί **πολλά φυτά**, για να καλύψει τις διατροφικές του ανάγκες, χρησιμοποιώντας λιπάσματα και φυτοφάρμακα. Συζητήστε τα θετικά και αρνητικά αποτελέσματα αυτής της χρήσης. **Γεωγραφικό γλωσσάριο**

Βλάστηση: ο **τρόπος ανάπτυξης** των φυτών μιας περιοχής. Βρύα: πολύ **μικρά φυτά** (μερικά τα παρατηρούμε μόνο με **μεγεθυντικό φακό**), που σχηματίζουν τάπητες και καλύπτουν βράγια, δέντρα, κεραμίδια κ.λπ. Λειχήνες: **φυτικοί οργανισμοί** που ζουν σε όλες τις περιοχές του κόσμου, αλλά κυριαρχούν σε περιοχές, όπου δεν μπορούν να επιβιώσουν άλλα φυτά, όπως στις **πολικές περιοχές** και στα πολύ **μεγάλα υψόμετρα**. Πανίδα: το σύνολο των ζώων που ζουν σε μια περιοχή μια **συγκεκριμένη χρονική στιγμή**. Φυτική διάπλαση: είδος **βλάστησης** χαρακτηριστικό μιας περιοχής. Χλωρίδα: το σύνολο των **αυτοφώνων φυτών** που υπάρχουν σε μια περιοχή. Ομαδική δραστηριότητα (προαιρετική)

Χωριζόμαστε σε τρεις ομάδες, επιλέγουμε έναν μεσημβρινό και καταγράφουμε τα **είδη βλάστησης** και τα ζώα, που συναντάμε κατά το φανταστικό μας ταξίδι

Entities: Clear All Tags

{ΠΟΛΥΛΕΚΤΙΚΟ}

Στο κείμενο αυτό, ο χρήστης μπορεί να επιλέξει να εμφανιστούν μορφοσυντακτικές πληροφορίες όπως: υποψήφιοι πολυλεκτικοί όροι, τα λήμματα και άλλες πληροφορίες (μέρος του λόγου, πτώση, κτλ).

3. Γλωσσικό Υπόβαθρο

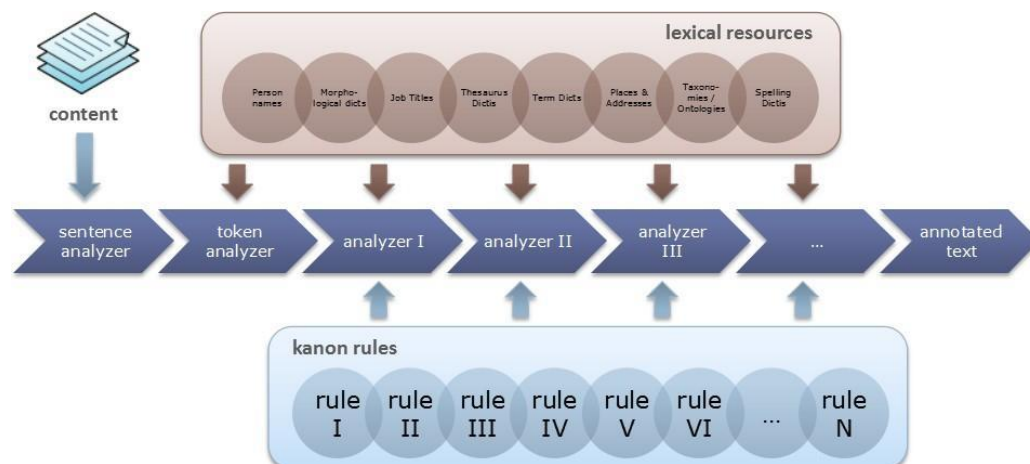
Η ανάλυση κειμένων για εντοπισμό και εξαγωγή γεγονότων απαιτεί γλωσσικούς πόρους και τεχνικές γλωσσικής ανάλυσης που επιτρέπουν τον εντοπισμό σύνθετων συντακτικών και σημασιολογικών γλωσσικών δομών βάσει των οποίων επιτυγχάνεται η αναγνώριση των όρων. Για την εξαγωγή των όρων απαιτήθηκε η αξιοποίηση των παρακάτω γλωσσικών πόρων:

1. **Λεξικά:** α) μορφολογικό λεξικό με 90.000 λήμματα και 1.200.000 παραγόμενους κλιτικούς τύπους και β) θησαυρό συνωνύμων και αντιθέτων 22.000 λημμάτων, που είχε αναπτύξει η Neurocom.
2. **Κανόνες:** Διαφορετικές κατηγορίες κανόνων και σχεδιάτυπων που αναπτύχθηκαν στο πλαίσιο του παρόντος έργου και εξασφαλίζουν τη μορφοσυντακτική αναγνώριση λέξεων ή ευρύτερων δομών, ώστε να είναι δυνατή η αναγνώριση όρων.

3.1 Μεθοδολογία ανάλυσης κειμένων για εντοπισμό όρων

Η μεθοδολογία ανάλυσης κειμένων για τον εντοπισμό όρων που ακολουθήθηκε στο έργο συνίσταται σε μία σειρά βημάτων που αξιοποιούν την πλατφόρμα και τις υποδομές που αναλύθηκαν παραπάνω.

Οι γενικές φάσεις επεξεργασίας του παρουσιάζονται στην Εικόνα 3:



Εικόνα 3: Φάσεις επεξεργασίας και ανίχνευσης όρων

Αναλυτικότερα, τα βήματα που ακολουθήθηκαν για την ανάλυση και εξαγωγή των γλωσσικών πληροφοριών από τα κείμενα είναι τα ακόλουθα:

1. **Συγκρότηση σώματος κειμένων** (βλέπε Π2.1 και Παράρτημα 1). Το αρχικό σώμα κειμένων αποτελείται κυρίως από τα σχολικά βιβλία Γεωγραφίας - Γεωλογίας.
2. **Κατάτμηση των κειμένων σε περιόδους, αναγνώριση σημείων στίξης και μικρότερων ενοτήτων στο πλαίσιο των περιόδων (tokenisation)**. Η διαδικασία αυτή επιτρέπει την αναγνώριση των ενοτήτων μέσα στις οποίες εντοπίζονται τα γεγονότα σχολιασμού (π.χ. περίοδος, πρόταση).

3. Αυτόματη μορφοσυντακτική αναγνώριση των λέξεων του κειμένου και λημματοποίηση.

- Η **λημματοποίηση** αφορούσε την απόδοση σε κάθε λέξη του ληματικού τύπου που την αντιπροσωπεύει (π.χ. «Βρείτε στην υδρόγειο σφαίρα τον Ισημερινό» → Ληματοποίηση: ΒΡΙΣΚΩ ΣΤΗΝ ΥΔΡΟΓΕΙΟΣ ΣΦΑΙΡΑ Ο ΗΣΗΜΕΡΙΝΟΣ).
- Κατά τη **μορφοσυντακτική** αναγνώριση αποδόθηκαν σε κάθε λέξη επισημειώσεις (tags) που αφορούν το μέρος του λόγου (π.χ. ουσιαστικό, επίθετο, ρήμα).

4. Εξαγωγή μονολεκτικών και πολυλεκτικών όρων: α) μέσω κανόνων που αφορούν την αναγνώριση ονοματικών φράσεων (ΟΦ) και ρηματικών φράσεων (ΡΦ) και β) με στατιστική ανάλυση που υπολογίζει λέξεις που απαντούν συχνά στα ίδια συμφραζόμενα. Για την στατιστική ανάλυση χρησιμοποιήθηκαν δύο ελεύθερες εφαρμογές ανοικτού κώδικα: το **KNIME** (<https://www.knime.com/>), μια πλατφόρμα για ανάλυση κειμένων, και το **KH-Coder** (<https://kncoder.net/en/>), μια εφαρμογή για ποσοτική (quantitative) ανάλυση περιεχομένου και εξαγωγή πληροφορίας (text mining).

5. Δημιουργία λεξικών. Πρόκειται για δύο (2) λεξικά:

- **Λεξικό πολυλεκτικών όρων**, που περιλαμβάνει πολυλεκτικά λήμματα και φράσεις, το οποίο αποτελεί το αποτέλεσμα της τελικής επιμέλειας των όρων που προέκυψαν από το βήμα 4, από τους γλωσσολόγους/φιλόλογους.
- **Λεξικό μονολεκτικών όρων**, που περιλαμβάνει μονολεκτικά λήμματα, το οποίο αποτελεί το αποτέλεσμα της τελικής επιμέλειας των όρων που προέκυψαν από το βήμα 4, από τους γλωσσολόγους/φιλόλογους.
- **Δημιουργία ταξινομίας**. Η ταξινόμια του έργου περιγράφεται στο παραδοτέο Π2.3.

Δημιουργία και εφαρμογή κανόνων. Για τη δημιουργία των κανόνων χρησιμοποιήθηκε ο φορμαλισμός «KANON». Οι κανόνες περιγράφονται στο παραδοτέο Π2.1.

3.2 Επιμέλεια Όρων

Σε σχέση με την αναγνώριση των όρων η διαδικασία ακολούθησε τα παρακάτω στάδια.

Α) Αυτόματη αναγνώριση υποψήφιων όρων από τα συστήματα, με την αξιοποίηση τόσο των γλωσσικών κανόνων (βλέπε προηγούμενη παράγραφο) όσο και στατιστικών αποτελεσμάτων. Η διαδικασία αυτή, όπως περιγράφηκε στα προηγούμενα περιλαμβάνει ανάλυση των κειμένων, κατάτμηση σε tokens, αφαίρεση των λεγόμενων «stop words», αναγνώριση του μέρους του λόγου των λημμάτων, και ανάδειξη των «υποψήφιων όρων». Τα αποτελέσματα της φάσης αυτής ελήχθησαν σε ένα λογιστικό φύλο (excel) στο οποίο στη συνέχεια προστέθηκε πληροφορία που μπορούσε να βοηθήσει τους γλωσσολόγους/φιλόλογους στην αξιολόγησή τους (μαζί με την εφαρμογή συμφραζομένων που περιγράφηκε στα προηγούμενα). Η πληροφορία έχει να κάνει με: α) τον συνολικό αριθμό των εμφανίσεων του όρου, β) Τις μοναδικές εμφανίσεις στο συγκεκριμένο κεφάλαιο, γ) Το βιβλίο στο οποίο βρέθηκε, δ) στατιστικά χαρακτηριστικά (π.χ. τον δείκτη: Mean(IDF) που είναι ο αριθμός των εμφανίσεων στο κεφάλαιο προς τον συνολικό αριθμό εμφανίσεων στο σώμα, που φανερώνει τη «σημαντικότητα» του όρου για το συγκεκριμένο κεφάλαιο). Επί πλέον, έγινε μια αναζήτηση στην Βικιπαιδεία και όταν υπήρχε ο όρος εκεί, εντάσσονταν και πληροφορίες από εκεί.

Στο παρακάτω βλέπουμε τη μορφή του λογιστικού φύλλου, το οποίο επεξεργάστηκαν / επιμελήθηκαν οι γλωσσολόγοι, οι οποίοι χαρακτήρισαν (με «ΝΑΙ») τους υποψήφιους όρους που θεώρησαν ότι είναι όροι Γεωγραφίας – Γεωλογίας.

ΑΡΙΟΚΑΤΣΙΚΟ	ΝΑΙ	Β' Γυμνασίου		1	1	Αίγαρος	YES	8	Σωλογία,Θηλαστικά	Είδη εκτεθειμένα στον κίνδυνο αφανισμού.Βασί: Ο αίγαρος ή κοινό αγριοκάτσικο (Capra aegagrus, Aeg. ο αίγαρος) είναι θηλαστικό και μονοκαστό.	Γεωλογία - Γεωγραφία (Β' Γυμνασίου) - Βιβλίο Μαθητή	2,263301229	0,002220001		
ΑΡΙΟΚΟΝΙΑΡΙΑ	ΝΑΙ	Ε' Δημοτικού		2	2						Γεωγραφία Ε' Δημοτικού - Βιβλίο Μαθητή	1,984527313	0,003802085		
ΑΡΙΟΚΡΑΝΕΣ	ΝΑΙ	Ε' Δημοτικού		1	1						Γεωγραφία Ε' Δημοτικού - Βιβλίο Μαθητή	2,263301229	0,00295858		
ΑΡΙΟΚΟΝΙΛΙΩΔ	ΝΑΙ	Ε' Δημοτικού		5	5						Γεωγραφία Ε' Δημοτικού - Βιβλίο Μαθητή	1,593266667	0,002385711		
ΑΡΙΟΠΡΟΒΑΤΟΥ	ΝΑΙ	Ε' Δημοτικού		1	1	Αγριού	YES	8	Σωλογία,Θηλαστικά	Είδη εκτεθειμένα στον κίνδυνο αφανισμού.Βασί: Το αγριού (γυναικά και ως μουφλόν) είναι είδος άγριου προβάτου και ανήκει στην οικογένεια Αιγίδια.	Γεωγραφία Ε' Δημοτικού - Βιβλίο Μαθητή	2,263301229	0,00310559		
ΑΡΙΟΣ		Ε' Δημοτικού		8	7						Γεωγραφία Ε' Δημοτικού - Βιβλίο Μαθητή	1,45156715	0,00393717	1,142657343	
ΑΡΙΟΚΤΗΜΑ	ΝΑΙ	Β' Γυμνασίου		5	4	Αγριοκίτρινος Συναδέρ, Λόγγ	YES	3		Γεωγραφία της Κόστα Ρίκα	Το Finca Sanaador σφύδα ξεκίνησε το 1979 με σκοπό να παρέχει σπίτια στους πρόσφυγες της Νικαράγας.	Γεωλογία - Γεωγραφία (Β' Γυμνασίου) - Βιβλίο Μαθητή	1,68797462	0,001545675	1,25
ΑΡΙΟΣ	ΝΑΙ	Β' Γυμνασίου		1	1						Γεωλογία - Γεωγραφία (Β' Γυμνασίου) - Βιβλίο Μαθητή	2,263301229	0,001344086		

Το σύνολο των όρων, είτε πρόκειται για πολυλεκτικούς είτε για μονολεκτικούς, αναρτήθηκε στην ιστοσελίδα του έργου.

Ως έμμεσο επακόλουθο της παραπάνω διαδικασία, είναι ότι δημιουργήθηκε και ένα «Ευρετήριο» (index) για κάθε βιβλίο που εντάχθηκε στο σώμα κειμένων. Η δημιουργία ευρετηρίου σε βιβλία, δεν είναι μια απλή διαδικασία, αλλά απαιτεί προσπάθεια, χρόνο, επιμονή και εξαιρετική επιμέλεια. Η αυτοματοποίηση που εισάγεται με την παραπάνω διαδικασία, βοηθάει και προς αυτήν την κατεύθυνση.

Παράρτημα Ι: Περιγραφή σώματος εκπαιδευτικών κειμένων

Βιβλίο Γεωλογίας – Γεωγραφίας Α' Γυμνασίου

Συγγραφείς: Παυλόπουλος, Κ. & Γαλάνη, Α.

Μπορεί να βρεθεί στο <http://ebooks.edu.gr/modules/ebook/show.php/DSGYM-A102/148/1056,3798/>

Στο βιβλίο της Γεωλογίας – Γεωγραφίας Α' Γυμνασίου περιλαμβάνονται:

- ο Ορολογία και ορισμοί (π.χ. παράλληλοι, μεσημβρινοί, Ισημερινός, πρώτος μεσημβρινός, Συντεταγμένες <http://ebooks.edu.gr/modules/ebook/show.php/DSGYM-A102/148/1057,3800/>, GIS <http://ebooks.edu.gr/modules/ebook/show.php/DSGYM-A102/148/1057,3802/>, χάρτες γενικής χρήσης, θεματικοί χάρτες <http://ebooks.edu.gr/modules/ebook/show.php/DSGYM-A102/148/1057,3803/> κ.α.)
- ο Χαρτογραφικά σύμβολα (π.χ. <http://ebooks.edu.gr/modules/ebook/show.php/DSGYM-A102/148/1057,3802/>, <http://ebooks.edu.gr/modules/ebook/show.php/DSGYM-A102/148/1057,3803/>)
- ο Ασκήσεις εντοπισμού τοποθεσίας πάνω στον χάρτη (π.χ. ορισμός τοποθεσίας με τη βοήθεια συντεταγμένων: <http://ebooks.edu.gr/modules/ebook/show.php/DSGYM-A102/148/1057,3800/> <http://ebooks.edu.gr/modules/ebook/show.php/DSGYM-A102/148/1057,3801/>)
- ο Πληροφορίες / οδηγίες για τη δημιουργία θεματικού χάρτη (π.χ. <http://ebooks.edu.gr/modules/ebook/show.php/DSGYM-A102/148/1057,3803/>)
- ο Τα ακόλουθα αποσπάσματα είτε λογοτεχνικά είτε επιστημονικά με γεωγραφική πληροφορία: <http://ebooks.edu.gr/modules/ebook/show.php/DSGYM-A102/148/1057,3799/>, <http://ebooks.edu.gr/modules/ebook/show.php/DSGYM-A102/148/1058,3805/>, <http://ebooks.edu.gr/modules/ebook/show.php/DSGYM-A102/148/1059,3826/>, <http://ebooks.edu.gr/modules/ebook/show.php/DSGYM-A102/148/1058,3805/>)

Σχόλια και παρατηρήσεις για την Δομή

- ο Δεν περιλαμβάνονται indexes με εικόνες ή λεξιλόγιο ή ορισμούς.
- ο Δεν υπάρχει μεγάλη συστηματικότητα στη δομή π.χ. άλλοτε οι έννοιες – λέξεις κλειδιά δίνονται με μικρά άλλοτε με κεφαλαία, άλλοτε χρωματισμένα με διαφορετικά χρώματα.
- ο Στην πρώτη σελίδα κάθε μίας από τις 3 στις 4 Ενότητες του βιβλίου περιλαμβάνονται: εικόνες, κείμενα, αποφθέγματα, φωτογραφίες (συνήθως χωρίς λεζάντες).
- ο Το διαδραστικό βιβλίο, ωστόσο, διασυνδέει τα κείμενα με εικόνες και υλικό π.χ. γίνονται συνδέσεις με το Φωτόδεντρο
 - σε χάρτες (πολλούς)
 - σε κάποιες δραστηριότητες (λίγες)
 - σε συλλογές φωτογραφιών (λίγες)
 - επίσης, γίνονται συνδέσεις με Wikipedia
 - σύνδεσμοι σε σχετικές ιστοσελίδες.

- Το βιβλίο του εκπαιδευτικού βρίσκεται σε μη-επεξεργάσιμη μορφή (από σάρωση σε μορφή εικόνας, βλ. σχόλιο πιο πάνω) και περιλαμβάνει οδηγίες διδασκαλίας και ενδεικτικές απαντήσεις στα ερωτήματα του βιβλίου και του τετραδίου εργασιών που απευθύνονται στους μαθητές.
- Το βιβλίο εργασιών περιλαμβάνει **ΕΠΙΠΛΕΟΝ κείμενα, αποφθέγματα, εικόνες (συνήθως χωρίς λεζάντες), συνδέσμους** και δραστηριότητες.

Μοτίβα – Patterns

Από την ανάλυση του κειμένου των βιβλίων (με έμφαση σε αυτό του μαθητή), προέκυψαν οι παρακάτω παρατηρήσεις σε σχέση με μοτίβα (patterns) που μπορεί να βοηθήσουν στο έργο:

- Ορισμοί:
 - Λέξεις που ξεκινούν με κεφαλαίο, συνεχίζουν με μικρά και ακολουθούνται από άνω και κάτω τελεία κι έχουν χρώμα κόκκινο ή μωβ ή κόκκινο και μωβ στην ίδια λέξη (π.χ. σελ. 11 / A1.1 γεωγραφικές συντεταγμένες).
 - Λέξεις με κεφαλαία που καταλήγουν σε άνω και κάτω τελεία κι έχουν χρώμα πράσινο, bold (π.χ. σελ. 36 / B1.2 Χωρίζοντας το φυσικό περιβάλλον σε ενότητες, σελ. 40 / B2.1, σελ. 72 / B4.2).
 - Λέξεις που βρίσκονται σε πράσινο πλαίσιο, με κίτρινο φόντο, ξεκινούν με κεφαλαίο και συνεχίζουν με μικρά (η πρώτη λέξη είναι γραμμένη με bold γράμματα), καταλήγουν σε άνω και κάτω τελεία κι έχουν χρώμα μαύρο (π.χ. σελ. 48 / B3.1).
 - Αντίστοιχη περίπτωση με το παραπάνω: λέξεις που βρίσκονται σε πράσινο πλαίσιο, με άσπρο φόντο, είναι γραμμένες με bold γράμματα, εντάσσονται σε κείμενο με μαύρα γράμματα. (σελ. 44, 53). Μικρή τροποποίηση: στη σελ. 56 υπάρχουν πράσινα πλαίσια, με άσπρο φόντο που μιλούν για την «Ιαπωνία», αλλά η λέξη δεν είναι γραμμένη με bold, ενώ στα πλαίσια περιλαμβάνονται και bullets, καθώς και κενά προς συμπλήρωση.
- Ονομασία ποταμών, πεδιάδων και ορέων:
 - Ονομασία ποταμών: λέξη με κεφαλαία, και bold, μαύρο χρώμα, μέσα σε πορτοκαλί πλαίσιο το οποίο βρίσκεται πάνω από κίτρινο πλαίσιο. (π.χ. σελ. 60 -61 / B3.5, σελ. 63 / B3.6, σελ. 66 / B3.7, σελ. 78-80 / B4.4)
- Πληροφορίες για όρη, πεδιάδες και ποτάμια:
 - Περιγραφή ποταμών: κίτρινο πλαίσιο κάτω από το πορτοκαλί πλαίσιο, όπως περιγράφεται παραπάνω (π.χ. σελ. 60 -61 / B3.5, σελ. 63 / B3.6, σελ. 66 / B3.7, σελ. 78-80 / B4.4) Μικρή τροποποίηση: Στις σελίδες 84-85 λέξη με κεφαλαία, και bold, μαύρα ή άσπρα γράμματα, μέσα σε πλαίσια διαφόρων χρωμάτων (πράσινο, κίτρινο, λαδί, πορτοκαλί, κόκκινο κλπ) που βρίσκονται πάνω από κίτρινο πλαίσιο, στο οποίο περιλαμβάνεται και εικόνα.

Υπομνήματα κάτω ή δίπλα από χάρτες (άλλοτε σε πλαίσιο κι άλλοτε εκτός)

- Χάρτης θερμικών ζωνών της γης – μέση θερμοκρασία σε βαθμούς κελσίου (σελ. 13 / A1.2)
- Υπόμνημα με χαρτογραφικά σύμβολα (σελ. 17 / A1.3)
- Υπόμνημα σε χάρτη ανάγλυφου της Ιαπωνίας (σελ. 18 / A1.3)
- Υπόμνημα σε χάρτη (σελ. 21 / A1.4), Στις σελίδες 26, 27 τα υπομνήματα του βιβλίου δεν φαίνονται καθαρά.

Βιβλίο Βιολογίας Α' Γυμνασίου

Συγγραφείς: Μαυρικάκη Ε., Γκούβρα Μ., Καμπούρη Α.

Μπορεί να βρεθεί στο <http://ebooks.edu.gr/modules/ebook/show.php/DSGYM-A103/517/3362,13558/>

Η καλά οργανωμένη δομή του βιβλίου της Βιολογίας Α' Γυμνασίου επιτρέπει την αυτόματη εξαγωγή θεματικού λεξιλογίου και άλλων δομημένων πληροφοριών για τη δημιουργία παιχνιδιών. Ειδικότερα, σε κάθε μια από τις 7 ενότητες περιλαμβάνονται:

- ο «Προηγούμενες γνώσεις που θα χρειαστώ»: Πρόκειται για εικόνες/σχεδιαγράμματα με λεζάντες/επεξηγήσεις.
- ο «Καινούριες γνώσεις που θα αποκτήσω»: Δίνονται συνοπτικά σε bullets.
- ο Τίτλος ενότητας - Εισαγωγικό κείμενο - αριθμημένες υπενότητες με τον τίτλο τους.
- ο «Ερωτήσεις Προβλήματα Δραστηριότητες»: Πρόκειται για σύντομες δοκιμασίες με τη μορφή κειμένων με κενά, ερωτήσεις πολλαπλών επιλογών, ασκήσεις σχολιασμού εικόνων, ανοιχτές ερωτήσεις κλπ. Οι Ερωτήσεις Προβλήματα Δραστηριότητες εμφανίζονται στο βιβλίο όπου κρίνεται σκόπιμο, συνήθως ύστερα από κάποιο αριθμό υποενοτήτων.
- ο «Στα ασπόνδυλα», «Στα σπονδυλωτά»: Κάθε ενότητα περιλαμβάνει ένα δισέλιδο με εικόνες/σχεδιαγράμματα και αντίστοιχο σχολιασμό/λεζάντες με στόχο τη μελέτη και τη σύγκριση διαφορετικών περιπτώσεων αναφορικά με τη λειτουργία που πραγματεύεται η ενότητα α) για τα ασπόνδυλα και β) για τα σπονδυλωτά ζώα.
- ο «Ας σκεφτούμε»: Σύντομο κείμενο επέκτασης του περιεχομένου της ενότητας με ερώτηση-προτροπή για περισσότερη σκέψη και κριτική ματιά από τον μαθητή.
- ο «Η Βιολογία και ...οι άλλες»: Σύντομο κείμενο σύνδεσης της Βιολογίας με άλλες επιστήμες, τέχνες ή την καθημερινή ζωή, έχοντας ως σημείο εκκίνησης μια συγκεκριμένη ενότητα του βιβλίου.
- ο «Μικρές έρευνες και εργασίες»: Ανάθεση εργασιών στους μαθητές.
- ο «Περίληψη»: Σύντομη περίληψη στο τέλος της ενότητας.
- ο «Λέξεις-Κλειδιά»: Στο τέλος της ενότητας.

Επίσης, στο τέλος του βιβλίου δίνονται:

- ο Χρονολόγιο της Βιολογίας
- ο Λεξιλόγιο όρων
- ο Ευρετήριο όρων

Βιβλίο Οικιακής Οικονομίας Α' Γυμνασίου

Συγγραφείς: Αποστολόπουλος Κ., Γεωργιτσογιάννη Ε., Κανέλλου Α., Σαϊτη Α., Σδράλη Δ., Τριάδη Δ.

Μπορεί να βρεθεί στο <http://ebooks.edu.gr/modules/ebook/show.php/DSGL103/57/436,2196/>

Καλά οργανωμένη δομή του βιβλίου. Σε κάθε μια από τις 7 ενότητες περιλαμβάνονται:

- Στόχοι της ενότητας/υποενότητας και Λέξεις-Κλειδιά (λαδί πλαίσιο).
- Οι σημαντικότεροι όροι της ενότητας/υποενότητας δίνονται είναι γραμμένοι με bold γράμματα και δίνονται σε ξεχωριστό γαλάζιο πλαίσιο.
- Οι σημαντικές πληροφορίες του κειμένου δίνονται με συνοπτικό τρόπο σε μωβ ή πράσινα πλαίσια.
- Στο τέλος κάθε ενότητας υπάρχει η «Ανακεφαλαίωση» (σε κίτρινο πλαίσιο).
- Οι εικόνες του βιβλίου είναι αριθμημένες και έχουν υπότιτλους.

Επίσης, στο τέλος του βιβλίου δίνονται:

- Γλωσσάριο
- Ευρετήριο όρων και ονομάτων

Φάκελος Μαθήματος - Προσωρινό Έντυπο Υλικό στα Θρησκευτικά Α' Γυμνασίου

Συγγραφείς: Γριζοπούλου Ο., Υφαντής Π., Γώγου Β., Μπάρλος Α.

Μπορεί να βρεθεί στο <http://ebooks.edu.gr/modules/ebook/show.php/DSGYM-A121/777/5080,23140/>

Ο χαρακτηρισμός «Φάκελος» δηλώνει τη «χαλαρή» δομή του βιβλίου. Στο βιβλίο υπάρχουν τρεις μεγάλες ενότητες. Σε διάφορα σημεία κάθε ενότητας υπάρχει γαλάζιο πλαίσιο με τίτλο «Προτάσεις για εργασίες και δραστηριότητες στην τάξη». Το βιβλίο περιλαμβάνει πλήθος λογοτεχνικών αποσπασμάτων και φωτογραφιών, οι περισσότερες από τις οποίες έχουν λεζάντες.

Στο τέλος του βιβλίου δίνεται κατάλογος Πηγών εικονιστικού και λοιπού βοηθητικού υλικού.

Βιβλίο Μουσικής Α' Γυμνασίου

Συγγραφείς: Σταυρίσης Μ., Σκορδή Μ., Χατζηγεωργίου-Λυμπουρή Α.

Μπορεί να βρεθεί στο <http://ebooks.edu.gr/modules/ebook/show.php/DSGYM-A118/357/2400,23156/>

Το βιβλίο περιλαμβάνει 8 ενότητες. Σε κάθε ενότητα υπάρχουν:

- «Λέξεις Κλειδιά» που πρόκειται να μελετηθούν στην ενότητα.
- «Ακρόαση» ενός ή περισσότερων μουσικών έργων.
- «Τραγούδι»: Απόσπασμα ενός γνωστού τραγουδιού. Δίνονται οι νότες και τα λόγια του τραγουδιού.

Στο τέλος του βιβλίου δίνεται Κατάλογος Όρων/Συμβόλων/Εννοιών, καθώς και Ευρετήριο.

Παράρτημα II – Υλικό από άλλες πηγές («Ιφιγένεια», «Φωτόδεντρο»)

Φωτόδεντρο – Εκπαιδευτικές δραστηριότητες που παρουσιάζουν ενδιαφέρον:

- ΚΑΤΑΣΚΕΥΗ ΧΑΡΤΩΝ, <http://photodentro.edu.gr/aggregator/lo/photodentro-lor-8521-7248>
- ΠΡΟΣΑΝΑΤΟΛΙΣΜΟΣ - ΕΞΑΣΚΗΣΗ ΣΤΟΝ ΧΩΡΟ ΚΑΙ ΤΟΝ ΧΑΡΤΗ, <http://photodentro.edu.gr/aggregator/lo/photodentro-lor-8521-6814>
- ΣΧΕΤΙΚΗ ΚΑΙ ΑΠΟΛΥΤΗ ΘΕΣΗ - ΨΑΧΝΟΝΤΑΣ ΣΤΗΝ ΑΛΕΞΑΝΔΡΟΥΠΟΛΗ <http://photodentro.edu.gr/aggregator/lo/photodentro-lor-8521-3302>
- ΥΠΟΛΟΓΙΖΟΝΤΑΣ ΑΠΟΣΤΑΣΕΙΣ ΥΠΟ ΚΛΙΜΑΚΑ <http://photodentro.edu.gr/aggregator/lo/photodentro-lor-8521-3278>
- ΦΤΙΑΧΝΩ ΤΟ ΥΠΟΜΝΗΜΑ ΕΝΟΣ ΧΑΡΤΗ (ΑΝΤΙΣΤΟΙΧΙΣΗ) <http://photodentro.edu.gr/aggregator/lo/photodentro-lor-8521-3216> (σε εμένα δεν λειτουργεί αλλά φαίνεται ενδιαφέρον)
- ΑΝΑΚΑΛΥΠΤΩ ΤΟΝ ΚΟΣΜΟ ΜΕΣΑ ΑΠΟ ΤΟΝ ΥΠΟΛΟΓΙΣΤΗ, <http://photodentro.edu.gr/edusoft/r/8531/229>, Λογισμικό E-Slate, Έργο των Πλειάδων που παρακολουθούσα. Επισυνάπτω το βιβλίο μαθητή. Δείτε κυρίως τις ενότητες 5 και 6 (χρειάζεται έλεγχος λειτουργίας του E-Slate)
- Γεωγραφικοί Όροι: <http://photodentro.edu.gr/aggregator/lo/photodentro-lor-8521-3650>
- 216 ΔΙΣΔΙΑΣΤΑΤΟΙ ΚΑΙ ΤΡΙΣΔΙΑΣΤΑΤΟΙ ΧΑΡΤΕΣ ΚΟΣΜΟΥ, ΗΠΕΙΡΩΝ, ΕΛΛΑΔΑΣ ΚΑΙ ΚΥΠΡΟΥ, <http://photodentro.edu.gr/aggregator/lo/photodentro-lor-8521-5118>
- ΑΝΑΓΛΥΦΟ, ΔΙΑΙΡΕΣΗ ΤΗΣ ΕΛΛΑΔΑΣ, <http://photodentro.edu.gr/aggregator/lo/photodentro-lor-8521-7249>
- ΑΠΟ ΤΑ ΓΕΩΓΡΑΦΙΚΑ ΔΕΔΟΜΕΝΑ ΣΤΟ ΤΟΠΩΝΥΜΙΟ, <http://photodentro.edu.gr/aggregator/lo/photodentro-lor-8521-3483> και αντίστροφα <http://photodentro.edu.gr/aggregator/lo/photodentro-lor-8521-3474> (σε εμένα φαίνεται να μην λειτουργούν τα τοπωνύμια)
- ΙΑΣΩΝ: ΑΠΟΣΤΑΣΕΙΣ ΣΤΗΝ ΕΛΛΑΔΑ: <http://photodentro.edu.gr/aggregator/lo/photodentro-lor-8521-9148> και ΜΕΤΡΗΣΗ ΑΠΟΣΤΑΣΗΣ <http://photodentro.edu.gr/aggregator/lo/photodentro-lor-8521-3481>
- ΠΑΓΚΟΣΜΙΟΣ ΧΑΡΤΗΣ ΚΑΙ ΥΔΡΟΓΕΙΟΣ ΣΦΑΙΡΑ <http://photodentro.edu.gr/aggregator/lo/photodentro-lor-8521-3477>

Ιφιγένεια: Εκπαιδευτικές δραστηριότητες που παρουσιάζουν ενδιαφέρον:

Εκπαιδευτικά σενάρια διαθέσιμα στη βιβλιοθήκη εκπαιδευτικών δραστηριοτήτων “Ιφιγένεια” περιέχουν ενδιαφέρουσες δραστηριότητες π.χ. σχετικές με τη Γεωγραφία, από τις οποίες μπορούν να αντληθούν ιδέες και πληροφορίες. Ενδεικτικά παραθέτουμε την ακόλουθη, μίνι συλλογή για την αξιοποίηση χαρτών:

- 5 Σενάρια για Γεωγραφικές συντεταγμένες και χρήση χαρτών: http://ifigeneia.cti.gr/repository/index.php?option=com_docman&task=cat_view&gid=267&Itemid=65 :
 - ο Ειδικά η «Χρήση χαρτών στην καθημερινή ζωή» περιλαμβάνει ένα ppt με χρήσιμες πληροφορίες για την ανάγνωση ενός χάρτη και τη δημιουργία υπομνήματος – θα μπορούσε να συνδυαστεί με αντίστοιχες δραστηριότητες στο Φωτόδεντρο (δείτε παρακάτω). Ομοίως και οι Γεωγραφικές Συντεταγμένες του κ. Ματσούκα «μικροδιδασκαλία που θέλει να δείξει τη σχέση της γεωγραφίας με τον πολιτισμό».

Ιφιγένεια: Κειμενική Πληροφορία από Εκπαιδευτικές δραστηριότητες που είναι αξιοποιήσιμη στο corpus της Γεωγραφίας:

Στο συνοδευτικό υλικό των εκπαιδευτικών σεναρίων που είναι διαθέσιμα στη βιβλιοθήκη εκπαιδευτικών δραστηριοτήτων “Ιφιγένεια”, περιλαμβάνονται παρουσιάσεις (αρχεία σε μορφή .ppt /.pptx) που περιέχουν ενδεχομένως αξιοποιήσιμη πληροφορία. Χρήσιμα μπορεί να φανούν και τα συνοδευτικά φύλλα εργασίας ως προς το περιεχόμενό τους, καθώς, μέσα από τα ερωτήματα που περιλαμβάνονται σε αυτά, γίνεται εμφανής η εστίαση σε «έννοιες που προβληματίζουν τους μαθητές». Ακολουθούν ενδεικτικά παραδείγματα:

- ο Δραστηριότητα με τίτλο: «Η Χρήση των Χαρτών στην Καθημερινή Ζωή – Λιάπη Α.» (http://ifigeneia.cti.gr/repository/index.php?option=com_docman&task=doc_download&gid=3033&Itemid=65) - pptx με πληροφορίες.
- ο Δραστηριότητα με τίτλο: «Γεωγραφικές Συντεταγμένες» (http://ifigeneia.cti.gr/repository/index.php?option=com_docman&task=doc_download&gid=1826&Itemid=65) - pptx με πληροφορίες.
- ο Δραστηριότητα με τίτλο: «planitis_gea» (http://ifigeneia.cti.gr/repository/index.php?option=com_docman&task=doc_download&gid=641&Itemid=65) γίνεται μέσα στο σενάριο αναφορά στις έννοιες «εναλλαγή μέρας – νύχτας» και «εναλλαγή των εποχών». Επίσης, μέσα στις δραστηριότητες, γίνεται αναφορά στο φαινόμενο της «περιοδικότητας» με τη μορφή ερωτήματος. Επίσης, στα φύλλα εργασίας δίνεται το στίγμα συγκεκριμένων περιοχών (π.χ. Ισημερινός (γεωγραφικό πλάτος =0) Ινδικός Ωκεανός (γεωγραφικό μήκος 75 μοίρες δυτικά).
- ο Δραστηριότητα με τίτλο: «Φαινόμενο Θερμοκηπίου» (http://ifigeneia.cti.gr/repository/index.php?option=com_docman&task=cat_view&gid=270&Itemid=65) - pptx με πληροφορίες.

- ο Δραστηριότητα με τίτλο: «Ιδιόσφαιρα» (http://ifigeneia.cti.gr/repository/index.php?option=com_docman&task=doc_download&gid=643&Itemid=65), στο σενάριο, υπάρχει επεξήγηση παρανοήσεων σε σχέση με τις αντιλήψεις των μαθητών. Επίσης, υπάρχει αναφορά στην υπάρχουσα γνώση για τη γη την εποχή του Κολόμβου.

Δραστηριότητα με τίτλο: «Ενδογενείς – Εξωγενείς Δυνάμεις_ Αμαλία Αναγνωστοπούλου» (http://ifigeneia.cti.gr/repository/index.php?option=com_docman&task=doc_download&gid=1007&Itemid=65) - pptx με πληροφορίες.

Παράρτημα III: Θεωρητική αναφορά στα Σώματα Κειμένων

Ορισμός του όρου «σώμα κειμένων»

Πολλοί είναι οι ορισμοί που έχουν δοθεί κατά καιρούς στον όρο σώμα κειμένων (corpus). Για το παρόν έργο, θα δεχθούμε ως ορισμό εργασίας αυτόν που δόθηκε στο πλαίσιο του προγράμματος EAGLES από το Sinclair (1996): σώμα κειμένων είναι 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. Ένα σώμα κειμένων όχι μόνον πρέπει να είναι αντιπροσωπευτικό (representative) και ισορροπημένο (balanced), όπως τονίζει ο Rundell (1995), αλλά πρέπει και να έχει επιλεγεί με βάση τους στόχους που έχουν θέσει οι άνθρωποι που το έχουν σχεδιάσει.

Αντίστοιχα, ηλεκτρονικό σώμα κειμένων (ΗΣΚ) είναι ένα σώμα κειμένων φυσικής γλώσσας που αποθηκεύεται σε μορφή αναγνώσιμη και επεξεργάσιμη από τον ηλεκτρονικό υπολογιστή (Renouf 1987, Leech & Fligelstone, 1992). Αξίζει να επισημάνουμε ότι, ενώ παλαιότερα ένα σώμα κειμένων μπορεί να ήταν απλώς ένα σύνολο κειμένων, βιβλίων, αποσπασμάτων από βιβλία ή απομαγνητοφωνημένων ομιλιών, αντίθετα στη σημερινή εποχή ένα σώμα κειμένων εμφανίζεται ως επί το πλείστον σε ηλεκτρονική μορφή. Είναι γεγονός ότι σήμερα είναι σημαντικός πλέον ο αριθμός των κειμένων που είναι είτε άμεσα διαθέσιμος είτε εύκολα μετατρέψιμος σε ηλεκτρονική μορφή. Έτσι, όπως επισημαίνει και ο Aarts (1990), στις μέρες μας εξυπακούεται ότι όταν μιλάμε για σώμα κειμένων εννοούμε σώμα μηχαναγνώσιμων (machine-readable) κειμένων. Επιπλέον, τα ΗΣΚ είναι ειδικά κωδικοποιημένα για τυποποιημένες και ομοιογενείς εργασίες ανάκτησης γλωσσικής πληροφορίας, ενώ τα γλωσσικά τμήματα που περιέχουν περιλαμβάνουν πληροφορίες για την προέλευση και την εξωγλωσσική λειτουργία τους.

Είναι φανερό ότι υπάρχουν συλλογές ηλεκτρονικών κειμένων που δεν μπορούν να χαρακτηριστούν ως ΗΣΚ. Έτσι, σύμφωνα με τα παραπάνω, δεν αποτελούν ΗΣΚ, για παράδειγμα, ο παγκόσμιος ιστός, εφόσον δεν έχει συγκεκριμένο και σταθερό μέγεθος ούτε έχει σχεδιαστεί για γλωσσολογικούς σκοπούς,

Είδη σωμάτων κειμένων

Τα ΗΣΚ διαφοροποιούνται ανάλογα με τη χρήση για την οποία προορίζονται και τις προδιαγραφές σχεδιασμού τους. Οι τυπολογίες που έχουν προταθεί είναι πολλές (βλ. Aston & Burnard 1998, McEnery & Wilson 1996) και διακρίνουν τα ΗΣΚ ως προς πολλά και διαφορετικά μεταξύ τους επίπεδα. Ενδεικτικά αναφέρουμε μερικές διακρίσεις ως προς:

- το περιεχόμενο των κειμένων (γενικό – ειδικευμένο ΗΣΚ)
- τον τρόπο γλωσσικής παραγωγής (γραπτού/προφορικού λόγου – μεικτό ΗΣΚ)
- τη χρονική κάλυψη των κειμένων (συγχρονικό – διαχρονικό ΗΣΚ)
- το γλωσσολογικό σχολιασμό των κειμένων (σχολιασμένο – μη σχολιασμένο ΗΣΚ)
- τον αριθμό των γλωσσών (μονόγλωσσο – πολύγλωσσο ΗΣΚ)
- τη δυνατότητα ανανέωσης του σώματος κειμένων (κλειστό – ανοιχτό ΗΣΚ)
- το πεπερασμένο μέγεθος (στατικό – δυναμικό ΗΣΚ ή ΗΣΚ αναφοράς – ελέγχου)

- κτλ.

Πολλοί ερευνητές έχουν χρησιμοποιήσει και άλλες παραμέτρους για την κατηγοριοποίηση των σωμάτων κειμένων. Έτσι, για παράδειγμα, σύμφωνα με τον Biber (1994) υπάρχουν κατηγοριοποιήσεις που αφορούν το αν ένα σώμα κειμένων αποτελείται από πλήρη κείμενα ή από αποσπάσματα κειμένων (sampled corpus), αν τα ποσοστά κειμένων που ανήκουν σε ένα τομέα είναι τυχαία ή προκαθορισμένα (balanced corpus) κ.ο.κ.

Στο παρόν έργο, στόχος είναι να σχεδιαστεί και να υλοποιηθεί ένα ηλεκτρονικό σώμα εκπαιδευτικών κειμένων, ανά θεματική περιοχή σε σχέση με τα μαθήματα που διδάσκονται οι μαθητές. Πρόκειται για ένα ειδικευμένο ΗΣΚ (specialized corpus) που περιλαμβάνει κείμενα που προέρχονται από θεματικές περιοχές της σχολικής εκπαίδευσης με στόχο την απεικόνιση της συγκεκριμένης υπογλώσσας (sublanguage).

Σε αντίθεση με τα γενικά ΗΣΚ ή ΗΣΚ αναφοράς (reference corpora) που σχεδιάζονται με στόχο να αναπαραστήσουν όσο το δυνατόν πιο αντιπροσωπευτικά το σύνολο του λεξιλογίου μιας συγκεκριμένης γλώσσας και να προσφέρουν αναλυτικές πληροφορίες για τη γλώσσα αυτή¹, τα ειδικευμένα ΗΣΚ ή ΗΣΚ ειδικής χρήσης (specialized/special corpora)² αποτελούνται αποκλειστικά από τα κείμενα στα οποία στοχεύει η ανάλυση και χαρακτηρίζεται από εύρος «σημαδεμένων» γλωσσικών χαρακτηριστικών. Τα ειδικευμένα ΗΣΚ δεν αντιπροσωπεύουν τη χρήση της γενικής γλώσσας, αλλά παρουσιάζουν σε υψηλή συχνότητα γλωσσικά φαινόμενα που βρίσκουμε πολύ σπάνια ή και καθόλου σε ένα γενικό ΗΣΚ. Ειδικά σε επίπεδο λεξιλογίου, τα ειδικευμένα ΗΣΚ περιλαμβάνουν κείμενα από συγκεκριμένα θεματικά πεδία και υψηλές συχνότητες εμφάνισης των όρων που σχετίζονται με αυτά.

Ειδική κατηγορία ΗΣΚ συνιστά το υποσώμα κειμένων (subcorpus) που είναι υποσύνολο ενός ΗΣΚ. Αποτελεί είτε στατικό στοιχείο ενός σύνθετου σώματος κειμένων, είτε δυναμικά προσδιοριζόμενη επιλογή από ένα ΗΣΚ για τις ανάγκες μίας συγκεκριμένης ανάλυσης. Αν, για παράδειγμα, θέλουμε να κάνουμε μια συγκριτική μελέτη μεταξύ των διαφόρων κειμενικών ειδών (π.χ. για να εξετάσουμε το εύρος της χρήσης μιας λέξης ή για να συγκρίνουμε τις υφολογικές παραλλαγές της μεταξύ των διαφόρων τομέων κτλ.) θα είναι πιο χρήσιμο να χωρίσουμε το ΗΣΚ μας σε υπο-ΗΣΚ. Η συγκρότηση ενός υπο-ΗΣΚ μπορεί να γίνεται με διάφορα κριτήρια, όπως είναι: η γλώσσα των κειμένων, τη γεωγραφική ποικιλία (π.χ. κοινή γλώσσα ή διάλεκτο), το κειμενικό είδος (δημοσιογραφικό, λογοτεχνικό, επιστημονικό, τεχνικό κτλ.), το ειδικό γνωστικό πεδίο, το επικοινωνιακό επίπεδο, κτλ. Συνήθως ένα ΗΣΚ αποτελείται από πολλά υπο-ΗΣΚ, τα οποία οργανώνονται σε ιεραρχική δομή και μπορούν να λειτουργήσουν και αυτοτελώς.

Τα σώματα κειμένων στη γλωσσολογική έρευνα

Η δυνατότητα πρόσβασης που έχουμε σήμερα σε μεγάλους όγκους κειμενικών δεδομένων σε ηλεκτρονική μορφή έπαιξε αποφασιστικό ρόλο στην ανάπτυξη ενός νέου ρεύματος στη γλωσσολογία: τη γλωσσολογία των σωμάτων κειμένων (corpus linguistics). Η γλωσσολογία των σωμάτων κειμένων εμφανίστηκε δυναμικά στις αρχές της δεκαετίας του 1980 ως απάντηση στη γενετική γραμματική και τις νεοπλατωνικές

¹ Το γενικό ΗΣΚ είναι σαφώς και αυστηρώς προσδιορισμένο και ισορροπημένο. Πρέπει να είναι αρκετά μεγάλο για να αναπαραστήσει ικανοποιητικά το πιο χαρακτηριστικό λεξιλόγιο και τις σημαντικότερες ποικιλίες, έτσι ώστε να μπορεί να χρησιμοποιηθεί ως βάση ανάπτυξης για αξιόπιστα λεξικά, γραμματικές, θησαυρούς και άλλα έργα γλωσσικής αναφοράς. Το μεγαλύτερο, αυτή τη στιγμή, ΗΣΚ αναφοράς για τη γραπτή μορφή της νέας ελληνικής, που είναι προσβάσιμο και μέσω διαδικτύου (<http://corpus.ilsp.gr>) είναι ο Εθνικός Θησαυρός για την Ελληνική Γλώσσα (ΕΘΕΓ) του Ινστιτούτου Επεξεργασίας του Λόγου (Hatzigeorgiou *et al.* 2000) που περιλαμβάνει σήμερα 46 εκατομμύρια λέξεις. Σε φάση υλοποίησης βρίσκεται επίσης το Σώμα Ελληνικών Κειμένων (ΣΕΚ) που εκπονείται με τη συνεργασία των Πανεπιστημίων Αθηνών, Κύπρου, King's College (London) και Birmingham (<http://www.ucy.ac.cy/sek>) το οποίο αναμένεται να περιλαμβάνει 30 εκατομμύρια λέξεις, προερχόμενες από το γραπτό και τον προφορικό λόγο (Γούτσος 2003).

² Ειδικευμένο ΗΣΚ της νέας ελληνικής, εξειδικευμένο για την ορολογική έρευνα, έχει αναπτυχθεί από το Ινστιτούτο Επεξεργασίας του Λόγου και ονομάζεται ΟΡΟΣΗΜΟ (Μάντζαρη *et al.* 1999). Περιλαμβάνει περίπου 3.000.000 λέξεις, σε κείμενα από 19 γνωστικές περιοχές, με 25.000 όρους επιλεγμένους και σχολιασμένους από τους ίδιους τους συγγραφείς των κειμένων.

θεωρίες που πρόσβευε, αφού σήμανε την επιστροφή της γλωσσολογικής θεωρίας στον εμπειρισμό. Στόχος της είναι να μελετήσει τη γλώσσα στο φυσικό της περιβάλλον, δίνοντας έτσι έμφαση στη γλωσσική πραγμάτωση (linguistic performance) έναντι της γλωσσικής ικανότητας (linguistic competence) η οποία βρισκόταν στο επίκεντρο του ενδιαφέροντος για τη γενετική γραμματική.

Πρέπει να διευκρινίσουμε ότι η γλωσσολογία των σωμάτων κειμένων δε συνιστά αυτοτελή κλάδο της γλωσσολογίας, με τον τρόπο που συνιστά, για παράδειγμα, η σημασιολογία, η μορφολογία, η φωνολογία, η κοινωνιογλωσσολογία κτλ., υπό την έννοια ότι κάθε γλωσσολογικός κλάδος έχει στόχο την περιγραφή και ανάλυση μιας συγκεκριμένης όψης της γλώσσας. Αντίθετα, η γλωσσολογία των σωμάτων κειμένων αποτελεί μια μεθοδολογική προσέγγιση που μπορεί να εφαρμοστεί σε διάφορους κλάδους της γλωσσολογίας. Έτσι, για παράδειγμα, κάνουμε λόγο για σύνταξη ή σημασιολογία που βασίζεται σε σώματα κειμένων (σε αντιδιαστολή προς τη σύνταξη ή τη σημασιολογία που δε βασίζεται σε σώματα κειμένων).

Βασικά χαρακτηριστικά της γλωσσολογικής ανάλυσης που βασίζεται σε σώματα κειμένων

Τα βασικά χαρακτηριστικά της γλωσσολογικής ανάλυσης που χρησιμοποιεί σώματα κειμένων, όπως τα συνοψίζουν οι Mc Enery & Wilson (1996), είναι τα εξής:

- η εμπειρική προσέγγιση,
- η χρήση ΗΣΚ ως βάση της ανάλυσης,
- η χρήση H/Y για τη συγκέντρωση και ανάλυση των δεδομένων,
- οι ποσοτικές και ποιοτικές τεχνικές ανάλυσης.

Η **εμπειρική προσέγγιση** αποτελεί τη βάση για τη γλωσσολογία των σωμάτων κειμένων (Leech 1992): ο γλωσσολόγος πρώτα συλλέγει και παρατηρεί την πληροφορία που εντοπίζει μέσα στα κείμενα, και στη συνέχεια αναπτύσσει ένα θεωρητικό μοντέλο που να συμπεριλαμβάνει το σύνολο των πληροφοριών που έχει συγκεντρώσει. Αυτό το μοντέλο μπορεί στη συνέχεια να επαληθευτεί ή να αντιπαραβληθεί με άλλα, ανάλογα μοντέλα. Με άλλα λόγια, η θεωρία στηρίζεται στην παρατήρηση και όχι το αντίστροφο. Πιο αναλυτικά:

1. Η συγκεκριμένη προσέγγιση βασίζεται σε πραγματικά κείμενα, άρα σε **εμπειρικά δεδομένα**, και όχι στην ενδοσκόπηση του γλωσσολόγου. Το σώμα κειμένων εμφανίζεται εδώ ως ο αντίποδας των κατασκευασμένων παραδειγμάτων που επινοούνται επί τούτου από το γλωσσολόγο για τις ανάγκες μιας ορισμένης θεωρίας ή μιας μελέτης.
2. Η γλωσσολογία των σωμάτων κειμένων παρέχει στους γλωσσολόγους ερευνητές τη δυνατότητα να διατυπώνουν αντικειμενικές και αξιόπιστες κρίσεις σχετικά με το υλικό που μελετούν, αφού τα σώματα κειμένων φέρνουν στο φως γλωσσικά φαινόμενα που είναι επαληθεύσιμα, γεγονός που αποτελεί σημαντικό κριτήριο για κάθε έρευνα που θέλει να λέγεται **επιστημονική**.
3. Το σώμα κειμένων αποτελεί γενικά το εργαλείο μιας **περιγραφικής ανάλυσης** της γλώσσας, που το παρατηρεί προκειμένου να ανασυστήσει εκ των υστέρων τις κανονικότητες που τυχόν το διέπουν. Αντίθετα, δεν μπορεί να χρησιμοποιηθεί για μια ρυθμιστική προσέγγιση, αφού το «καθαρό» σώμα κειμένων σε καμία περίπτωση δεν μπορεί να υπακούσει σε εκ των προτέρων καθορισμένους κανόνες.
4. Η μελέτη των σωμάτων κειμένων αποτελεί τη βάση για την περιγραφή μιας γλώσσας, χωρίς ωστόσο να αποτελεί εμπόδιο στην ανάπτυξη θεωριών, αφού η παρατήρηση των κειμένων και η γλωσσική περιγραφή είναι απαραίτητες προϋποθέσεις για την **ανάπτυξη μιας θεωρητικής προσέγγισης**.

5. Τα εμπειρικά δεδομένα αναγκάζουν το γλωσσολόγο να λάβει υπόψη του όλα τα γλωσσικά φαινόμενα, αφού όλα ανεξαιρέτως τα φαινόμενα που καταγράφονται θα πρέπει και να ερμηνευτούν. Ως εκ τούτου, ένα μοντέλο που βασίζεται στην ανάλυση των σωμάτων κειμένων μπορεί να είναι ταυτόχρονα **ποσοτικό** (συχνότητες, αναλογίες κτλ.) και **ποιοτικό** (ερμηνεία και ανάλυση των φαινομένων).

Καλό είναι πάντως, όπως υποδεικνύει ο Svartvik (1992), η εμπειρική προσέγγιση να συνδυάζεται με την ενδοσκόπηση και την κριτική ματιά του γλωσσολόγου, αν θέλουμε να σχεδιάσουμε γλωσσικά μοντέλα που να είναι πραγματικά πλήρη, αξιόπιστα και αντιπροσωπευτικά.

Η σύγχρονη εποχή: τομείς εφαρμογής

Στις μέρες μας έχει πλέον αποδειχθεί ότι η μελέτη των σωμάτων κειμένων παρουσιάζει πολλά πλεονεκτήματα για τη γλωσσολογική έρευνα, κυρίως ως πηγή συστηματικής άντλησης δεδομένων και ως βάση για τη διατύπωση θεωρητικών υποθέσεων (Svartvik 1992, Leech 1992). Τα τελευταία χρόνια η μελέτη της γλώσσας στο φυσικό της περιβάλλον κέρδισε σημαντικό έδαφος, αφού εκτιμάται ότι οι σχετικές μελέτες διπλασιάζονται περίπου κάθε 5 χρόνια (Altenberg 1991). Η γλωσσολογία των σωμάτων κειμένων έχει κάνει έντονα αισθητή την παρουσία της σε πολλούς τομείς της γλωσσολογικής έρευνας καθώς παρέχει νέες δυνατότητες στην περιγραφή και ανάλυση της γλώσσας. Τα σώματα κειμένων είναι μία σημαντική πηγή πληροφορίας για τη γλωσσολογία, προπαντός για κάποιους κλάδους της εφαρμοσμένης γλωσσολογίας, όπως η λεξικογραφία, η αυτόματη μετάφραση, η αναγνώριση της ομιλίας, η παραγωγή ορολογικών έργων αναφοράς όπως γλωσσάρια υπογλωσσών, ειδικευμένα λεξικά, ορολογικές βάσεις δεδομένων κτλ., καθώς και η ανάπτυξη θεματικών θησαυρών και οντολογιών.

Τα σώματα κειμένων στην ορολογική έρευνα

Τα τελευταία χρόνια η χρήση των ΗΣΚ στην ορολογική έρευνα έχει γενικευτεί. Παραδοσιακά εξειδικευμένα λεξικά, αναγνωρισμένη ορολογία, ονοματολογίες, θησαυροί, κατηγοριοποιήσεις και βάσεις δεδομένων δεν επαρκούν για να προσφέρουν μια ολοκληρωμένη λύση στο πρόβλημα της συστηματικής καταγραφής, πρόσβασης και ανάκτησης της εξειδικευμένης γνώσης και των όρων (μονολεκτικών και πολυλεκτικών) που πραγματώνουν γλωσσικά τις έννοιες ενός επιστημονικού τομέα. Τα ΗΣΚ έχουν καθιερωθεί πλέον ως εργαλεία αναφοράς και συμβάλλουν σημαντικά τόσο στην ανάπτυξη υπολογιστικών εργαλείων που σχετίζονται με την επεξεργασία ορολογίας όσο και στην αξιολόγηση και τεκμηρίωση των αποτελεσμάτων τους.

Η μεγαλύτερη πρακτική απήχηση από τη χρήση των ΗΣΚ στο χώρο της ορολογίας είναι η δυνατότητα αυτόματης αναγνώρισης όρων³. Οι ερευνητικές προσεγγίσεις που εφαρμόζονται για την αντιμετώπιση των ιδιαίτερων προβλημάτων που παρουσιάζει η αναγνώριση όρων βασίζονται σε τεχνικές και θεωρίες που προέρχονται από διάφορους ερευνητικούς τομείς. Έτσι, ανάλογα με τον ερευνητικό χώρο από τον οποίο εμπνέονται, διακρίνουμε (Ανανιάδου & Ζερβάνου 2004):

- τις μεθόδους που προέρχονται από το χώρο της επιστήμης της πληροφορίας, και ειδικότερα τον τομέα της ανάκτησης πληροφοριών: τέτοιες είναι οι στατιστικές μέθοδοι και άλλες τεχνικές δεικτοδότησης, κατηγοριοποίησης και ομαδοποίησης κειμένων (ποσοτική προσέγγιση, βλ. Bourigault 1992)

³ Ας σημειωθεί ότι οι Jacquemin & Bourigault (2003) χρησιμοποιούν τον όρο *αυτόματος εντοπισμός όρων* (term detection) ο οποίος καλύπτει τόσο την *αναγνώριση όρων* (term recognition) όσο και την *πρόσκτηση όρων* (term acquisition).

- τις μεθόδους που βασίζονται σε γλωσσολογικές θεωρίες, κυρίως στους τομείς της λεξιλογικής, μορφολογικής και μορφοσυντακτικής έρευνας (ποιοτική προσέγγιση, βλ. Calzolari & Bindi 1990)
- τις υβριδικές μεθόδους, που συνήθως συνδυάζουν στατιστικές τεχνικές με γλωσσολογικές θεωρίες και μεθόδους (βλ. ενδεικτικά Justeson & Katz 1995⁴, Daille *et al.* 1994⁵, Frantzi & Ananiadou 1999⁶, Maynard & Ananiadou 2000⁷, Georgantopoulos & Piperidis 2000)
- και τις μεθόδους που προέρχονται από το χώρο της τεχνητής νοημοσύνης και εφαρμόζουν διαφόρους αλγορίθμους μηχανικής μάθησης (Nobata *et al.* 1999, Collier *et al.* 2000⁸).

Η χρήση των σωμάτων κειμένων επιτρέπει την εξέταση των όρων μέσα στο συμφραστικό τους περιβάλλον, γεγονός που βελτιώνει σημαντικά τις τεχνικές αυτόματης αναγνώριση και εξαγωγή τους και συμβάλλει στην ανάπτυξη ειδικών υπολογιστικών συστημάτων (Frantzi & Ananiadou 1997).

Τα σώματα κειμένων είναι απαραίτητα επίσης όχι μόνο στην ανάπτυξη εργαλείων για την ορολογική έρευνα, αλλά και στην αξιολόγηση των αποτελεσμάτων που δίνουν τα εργαλεία αυτά. Έτσι, όπως έχει διαπιστωθεί, οι στατιστικές προσεγγίσεις είναι πιο αποδοτικές στον εντοπισμό όρων με υψηλή συχνότητα εμφάνισης, αλλά μάλλον ασταθείς στον εντοπισμό πιο σπάνιων όρων, με αποτέλεσμα να παράγουν πολλή «σιωπή» (ανεπαρκή εξαγωγή πληροφορίας). Αντίθετα, οι γλωσσολογικές προσεγγίσεις αποδίδουν καλύτερα αποτελέσματα όσον αφορά όρους που εμφανίζονται με χαμηλή συχνότητα, αλλά παράγουν περισσότερο «θόρυβο» (λανθασμένο εντοπισμό πληροφορίας). Προκειμένου να συνδυαστούν τα πλεονεκτήματα των δύο αυτών προσεγγίσεων και να βελτιωθεί η αποτελεσματικότητά τους, έχουν αναπτυχθεί υβριδικές προσεγγίσεις που συνδυάζουν τη στατιστική με τη γλωσσολογική πληροφορία.

Η χρήση σωμάτων κειμένων επιτρέπει στον ερευνητή να εντοπίσει τους όρους (ήδη υπάρχοντες, νεοσύστατους ή ακόμη και εν τη γενέσει τους) και να βγάλει ασφαλή συμπεράσματα σχετικά με τη χρήση τους και τη συχνότητα εμφάνισής τους σε μία δεδομένη υπογλώσσα. Παρέχει επίσης τη δυνατότητα να εξεταστούν οι συμφραστικές τους σχέσεις, γεγονός που οδηγεί στον εντοπισμό των γλωσσικών τους χαρακτηριστικών σε μορφολογικό, συντακτικό και σημασιολογικό επίπεδο (Bowker 1996, 2000). Συγκεκριμένα, η εξέταση των συμφραζομένων επιτρέπει τον εντοπισμό των βασικών μορφολογικών του παραγώγων, την εξέταση των συντακτικών περιορισμών που επιβάλλονται στις συμφράσεις ενός όρου, καθώς και της σημασιολογικής του προσωδίας (semantic prosody, βλ. Sinclair 1987).

Τέλος, η χρήση σωμάτων κειμένων μπορεί να συμβάλει στον εντοπισμό ασυνέπειας στη χρήση ταυτόσημων όρων. Η ασυνέπεια μπορεί να παρατηρηθεί σε ορθογραφικό, μορφολογικό ή σημασιολογικό

⁴ TERMS: σύστημα με στόχο την αναγνώριση πολυλεκτικών όρων σε κείμενα οποιουδήποτε αντικείμενου και τύπου. Κατά την αξιολόγηση αναφέρεται ότι το σύστημα επιτυγχάνει ποσοστά ανάκλησης της τάξης του 71% και ακρίβειας 67-96% σε σύγκριση με δεδομένα αναγνώρισης όρων στο ίδιο κειμενικό σώμα από ειδικούς επιστήμονες. Παρά το γεγονός ότι το σύστημα επιτυγχάνει ικανοποιητικά αποτελέσματα, υποστηρίζεται ότι η χρήση μορφοσυντακτικού αναλυτή θα μπορούσε να βελτιώσει την ανάλυση, παρά το ενδεχόμενο ποσοστό σφάλματος (Jacquemin 2001).

⁵ ACABIT: υβριδικός αναλυτής που αποτελείται από επιφανειακό μορφοσυντακτικό αναλυτή και στατιστικά φίλτρα για την αναγνώριση όρων στη γαλλική γλώσσα.

⁶ Τιμή C/CN: μέθοδος που συνδυάζει στατιστικές μετρήσεις με γλωσσικές πληροφορίες και μπορεί να εφαρμοστεί σε οποιοδήποτε γνωστικό αντικείμενο και είδος κειμένου για αναγνώριση πολυλεκτικών και ένθετων όρων.

⁷ TRUCKS: μέθοδος που αποτελεί επέκταση της τιμής C/CN και εστιάζει στον εντοπισμό πληροφοριών για το γλωσσικό περιβάλλον ενός όρου μέσω διαφόρων πηγών με στόχο τη βελτιστοποίηση των στατιστικών μετρήσεων για αναγνώριση όρων. Διακρίνει τρία είδη συγκεκριμένης πληροφορίας (μορφοσυντακτική, σημασιολογική και ορολογική). Μεταξύ άλλων κάνει χρήση ορολογικών θησαυρών, όπως ο θησαυρός ιατρικής ορολογίας UMLS.

⁸ Μέθοδος εποπτευόμενης εκμάθησης που βασίζεται σε κρυμμένα μαρκοβιανά μοντέλα (Hidden Markov Models) για τον εντοπισμό όρων μοριακής βιολογίας σε κείμενα από το MEDLINE (Medical Literature Analysis and Retrieval System On Line).

επίπεδο, και η αδυναμία εντοπισμού και διόρθωσής της ενδέχεται να αποτελέσει σημαντικό πρόβλημα για τη μετέπειτα επεξεργασία και αξιοποίηση των όρων. Χάρη στους συμφραστικούς πίνακες που προκύπτουν με αναζήτηση πιθανών μορφών με τις οποίες μπορεί να συναντάται ένας όρος, είναι δυνατή η γρήγορη και εύκολη ομοιογενοποίησή του σώματος κειμένων, με αποτέλεσμα την αύξηση του ρυθμού επαναληψιμότητας⁹. Εννοείται ότι όσο μεγαλύτερη είναι η συνέπεια στη χρήση των όρων, τόσο μεγαλύτερη είναι η επαναληψιμότητα.

Τα σώματα κειμένων στην κατασκευή οντολογίας

Εκτός από την αναγνώριση των όρων ενός συγκεκριμένου γνωστικού πεδίου, τα σώματα κειμένων παρέχουν επίσης σημαντική βοήθεια στον εννοιολογικό προσδιορισμό αυτών των όρων. Με άλλα λόγια, βοηθούν στον εντοπισμό των σημασιολογικών σχέσεων που συνδέουν τους όρους μέσα στο εννοιολογικό σύστημα που αντικατοπτρίζει τη γνώση ενός συγκεκριμένου πεδίου. Αυτό επιτρέπει τη δημιουργία οντολογιών προσδιορισμένων εμπειρικά από κείμενα μιας συγκεκριμένης υπογλώσσας.

Για τη διερεύνηση των σημασιολογικών σχέσεων ενός όρου X, η Bowker (1996) προτείνει την αναζήτηση με χρήση τυποποιημένων εκφράσεων, π.χ. «το X ανήκει στο Y» ή «το X είναι μέρος του Y». Τέτοιες εκφράσεις αποκαλύπτουν σχέσεις σημασιολογικής *υπωνυμίας* (δηλ. μέρους-όλου) και μπορούν να χρησιμοποιηθούν ως συνθήκες αναζήτησης σε ένα σώμα κειμένων ειδικής χρήσης στο οποίο θα εντοπίσουν προτάσεις όπου το X ως προς το Y θα λειτουργεί ως *υπώνυμο*.

Προκειμένου να διερευνήσει κανείς το σύνολο των σημασιολογικών σχέσεων ενός όρου, θα πρέπει να καθορίσει τη θέση αυτού του όρου σε ένα ειδικό γνωστικό πεδίο αίροντας τυχόν ασάφεια που τον χαρακτηρίζει. Η ασάφεια είναι ένα γενικό φαινόμενο των φυσικών γλωσσών, ειδικότερα όμως στην κατασκευή μιας οντολογίας, η ασάφεια των λεξιλογικών μονάδων μπορεί να είναι μορφοσυντακτική, συντακτική, σημασιολογική και ασάφεια ταξινόμησης¹⁰.

Η σημασιολογική ασάφεια οφείλεται κυρίως σε φαινόμενα πολυσημίας και ομωνυμίας. Έτσι, ενώ, θεωρητικά, οι όροι πρέπει να είναι *μονοσήμαντοι* (δηλαδή ένας και μόνο όρος πρέπει να αντιστοιχεί σε μία και μόνον έννοια), εντούτοις πολύ συχνά στην πράξη εμφανίζονται φαινόμενα *αμφισημίας* και *πολυσημίας* (όταν ένας όρος αντιστοιχεί σε πολλές έννοιες), καθώς και *συνωνυμίας* (όταν πολλοί όροι εκφράζουν την ίδια έννοια). *Πολύσημοι* είναι οι όροι που έχουν την ίδια φωνολογική μορφή αλλά αναφέρονται σε διαφορετικές συγγενείς έννοιες, είτε σε διαφορετικούς τομείς (π.χ. ο όρος *φακός* απαντά στον τομέα της φωτογραφίας, της φυσικής και της ανατομίας), είτε και στο ίδιο γνωστικό αντικείμενο (π.χ. ο όρος *γονίδιο* στη βιολογία έχει δύο έννοιες: α) το καταγεγραμμένο τμήμα του DNA που μεταφράζεται σε πρωτεΐνη, και β) η περιοχή του DNA που κατονομάζεται και φέρει ένα γενετικό φαινότυπο). Οι *ομώνυμοι* όροι έχουν την ίδια φωνολογική μορφή, αλλά αναφέρονται σε εντελώς διαφορετικές έννοιες (π.χ. *διαβήτης*: ασθένεια / γεωμετρικό όργανο). Η σημασιολογική αποσαφήνιση των όρων μπορεί να επιτευχθεί με στατιστικές μετρήσεις ομοιότητας μεταξύ των εξαχθέντων όρων και του γλωσσικού τους περιβάλλοντος.

Η ασάφεια ταξινόμησης είναι ένα άλλο είδος ασάφειας που δημιουργείται όταν μία έννοια, και συνεπώς ο όρος που την εκφράζει, μπορεί να κατηγοριοποιηθεί με πολλούς τρόπους. Αυτού του είδους η ασάφεια αφορά τους λεγόμενους *πολυδιάστατους όρους* (Bowker 1995). Ως πολυδιαστατικότητα ενός όρου ορίζεται ένα ειδικό είδος αμφισημίας που αφορά τη δυνατότητα ενός όρου να ανήκει σε δύο ή και περισσότερα ιεραρχικά δίκτυα εννοιών ταυτόχρονα και εμφανίζεται ως πολλαπλή κληρονομικότητα σε μία ιεραρχία όρων (όταν ένας κόμβος σε μία δένδρική ιεραρχία εμφανίζεται να έχει περισσότερους από ένα γονικούς

⁹ Εννοούμε το δείκτη που προσδιορίζει το πλήθος των όρων που επαναλαμβάνονται μέσα σε ένα κείμενο.

¹⁰ Ένα άλλο πρόβλημα είναι ότι μια λεξιλογική μονάδα μπορεί να είναι όρος σε κάποιο ειδικό γνωστικό πεδίο, αλλά όχι σε κάποιο άλλο. Επίσης, όροι μπορεί να είναι και λέξεις προερχόμενες από τη γενική γλώσσα, οι οποίες όμως σε ένα συγκεκριμένο γνωστικό πεδίο χρησιμοποιούνται με ειδική έννοια.

κόμβους). Για παράδειγμα, ο όρος *πνευμονική φυματίωση* μπορεί να ταξινομηθεί και ως ασθένεια του αναπνευστικού συστήματος και ως μεταδοτική ασθένεια. Η άρση της ασάφειας είναι σημαντική προκειμένου να μπορέσουμε να κατηγοριοποιήσουμε έναν όρο ανάλογα με τη σημασία του. Ο βαθμός λεπτομέρειας που επιθυμούμε να έχουμε στην ταξινόμηση και διάκριση μεταξύ των εννοιών εξαρτάται από την εφαρμογή, που μπορεί π.χ. να σχετίζεται με εξαγωγή πληροφορίας ή με ανάπτυξη οντολογιών.

Δεδομένου ότι η εννοιολογική λειτουργία ενός όρου (είτε μονοσήμαντου είτε πολυδιάστατου) σχετίζεται άμεσα με τα συμφραζόμενα¹¹, ο ειδικός μελετητής μπορεί να αξιοποιήσει τα ποιοτικά και ποσοτικά στοιχεία που προσφέρουν οι συμφραστικοί πίνακες ενός ειδικευμένου ΗΣΚ, προκειμένου να μελετήσει τη συχνότητα συνεμφάνισης των λεξικών στοιχείων, για να εντοπίσει τις εννοιολογικές λειτουργίες που προσδιορίζουν το συγκεκριμένο όρο στην υπογλώσσα που μελετά.

Βιβλιογραφία

- Aarts, J. (1990): « Corpus Linguistics: An Appraisal », Computers in Literary and Linguistic Research - Proceedings of the Fifteenth International Conference, Y. Choueka (Éd.), VILLE, Champion-Slatkine, σ. 13-27.
- Altenberg, B. (1991): “A bibliography of publications relating to English computer corpora”, Johansson, S. & Stenstrom, A. B. Berlin (eds.), English computer corpora, Mouton de Gruyter, σ. 355-396.
- Aston, G. & Burnard, L. (1998): The BNC handbook. Exploring the British National Corpus with SARA, Edinburgh, Edinburgh University Press.
- Atkins, Sue, Jeremy Clear, and Nicholas Ostler. (1992) "Corpus Design Criteria". Literary and Linguistic Computing 7, 1-16.
- Biber, D. (1994): “An analytical framework for register studies”, στο Biber, D. & Finegan, E. (eds.), Sociolinguistic Perspectives on Register, Studies in Sociolinguistics, Oxford, Oxford University Press.
- Bourigault, D. (1992): “Surface grammatical analysis for the extraction of terminological noun phrases”, Proceedings of Coling '92, Nantes, σ. 977-981.
- Bowker, L. (1995): A Multidimensional Approach to Classification in Terminology: Working Within a Computational Framework, PhD thesis, University of Manchester, Institute of Science and Technology.
- Bowker, L. (1996): “Towards a corpus-based approach to terminography”, Terminology 3, σ. 27-52.
- Bowker, L. (2000): “Methodology for exploiting target language corpora”, International Journal Corpus Linguistics 5, σ. 17-52.
- Burnard, L. 1999. Using SGML for linguistic analysis: the case of the BNC. In Markup languages theory and practice, 31-51. Cambridge, Mass: MIT Press
- Calzolari, N. & Bindi, R. (1990): “Acquisition of lexical information from a large textual Italian corpus”, Proceedings of Coling '90, Helsinki, σ. 54-59.

¹¹ Οι Maynard & Ananiadou (1998) περιγράφουν μια μεθοδολογία που στηρίζεται σε ειδικευμένα ΗΣΚ και αξιοποιεί τη συμφραστική πληροφορία του όρου για να αποσαφηνίσει πολυδιάστατους όρους.

- Daille, B., Gaussier, E. & Lange, J.-M. (1994): “Towards Automatic Extraction of Monolingual
- Georgantopoulos, B. & Piperidis, S. (2000): “A hybrid technique for automatic term extraction”, Proceedings of International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications – ACIDCA 2000, Monastir, Tunisia, σ. 124-128.
- Hatzigeorgiou, N., Gavriilidou, M., Piperidis, S., Carayannis, G., Papakostopoulou, A., Spiliotopoulou, A., Vacalopoulou, A., Labropoulou, P., Mantzari, E., Papageorgiou, H. & Demiros, I. (2000): “Design and implementation of the online ILSP Greek Corpus”, Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000), Athens, ELRA, σ. 1737-1742.
- Leech, G. & Fligelstone, S. (1992): “Computers and Corpus Analysis”, στο Computers and Written Texts, C. Butler (ed.), Oxford, Blackwell.
- Leech, G. (1991): “The state of the art in corpus linguistics”, στο Aijmer and Altenberg (eds.), σ. 8-29.
- Leech, G. (1992): “Corpora and theories of linguistic performance”, στο Svartvik 1992, σ. 105-122.
- Maynard, D. & Ananiadou, S. (2000): “TRUCKS: A Model for Automatic Multi-Word Term Recognition”, Journal of Natural Language Processing 8/1, σ. 101-125.
- McEnery, T. & Wilson, A. (1996): Corpus Linguistics, Edinburgh, Edinburgh University Press.
- Sinclair, J. & Ball, J. (1996): EAGLES Text typology.
- Sinclair, J. 1982. Reflections on computer corpora in English language research. In Computer corpora in English language research, ed. Stig Johansson: 1-6. Bergen.
- Sinclair, J. 1987: Looking up: an account of the COBUILD project in lexical computing, London, Collins.
- Sinclair, J. 1989. Corpus creation. In Language, learning and community, eds. C Candlin and T McNamara, 25-33: NCELTR Macquarie University.
- Sinclair, J. 1991. Corpus, concordance, collocation. Oxford: Oxford University Press.
- Sinclair, J. 1995. From theory to practice. In Spoken English on Computer, eds. G. Leech, G. Myers and J. Thomas, 99-112. Harlow: Longman.
- Sinclair, J. 2001. Preface. In Small corpus studies and ELT, eds. Mohsen Ghadessy, Alex Henry and Robert L. Roseberry, vii-xv. Amsterdam/Philadelphia: John Benjamins.
- Sinclair, J. 2003. Corpora for lexicography. In A practical guide to lexicography, ed. P Van Sterkenberg. Amsterdam: John Benjamins.
- Sinclair, J. 2004. Intuition and annotation - the discussion continues. In Advances in corpus linguistics. Papers from the 23rd International Conference on English Language Research on Computerized corpora (ICAME 23). Göteborg 22-26 May 2002., eds. Karin Aijmer and Bengt Altenberg, 39-59. Amsterdam/New York: Rodopi.

- Sinclair, J. 2005. "Corpus and Text - Basic Principles", στο Developing Linguistic Corpora: a Guide to Good Practice, ed. M. Wynne. Oxford, Oxbow Books, σ. 1-16 Available from: <http://ahds.ac.uk/linguistic-corpora/>
- Svartvik, J. (1992): Directions in Corpus Linguistics, Berlin: Mouton de Gruyter.
- Ανανιάδου, Σ. & Ζερβάνου, Κ. (2004): «Αναγνώριση όρων σε υπολογιστικά συστήματα: Προβλήματα και μέθοδοι», στο Ελληνική ορολογία: έρευνα και εφαρμογές, Μ. Κατσογιάννου & Ε. Ευθυμίου (επιμ.), Αθήνα, Εκδόσεις Καστανιώτη.
- Γούτσος, Δ. (2003): «Σώμα Ελληνικών Κειμένων: σχεδιασμός και υλοποίηση», Πρακτικά του 6ου Διεθνούς Συνεδρίου Ελληνικής Γλωσσολογίας, Ρέθυμνο, Πανεπιστήμιο Κρήτης.
- Μάντζαρη, Έ., Γαβριηλίδου, Μ., Λαμπροπούλου, Π. & Καραγιάννης, Γ. (1999): «Συλλογή ηλεκτρονικών ορολογικών πόρων: μεθοδολογία συγκέντρωσης και αποτελέσματα», στο Ελληνική γλώσσα και ορολογία – Ανακοινώσεις 2ου Συνεδρίου, Αθήνα, ΕΛΕΤΟ, σ. 133-142.

Παράρτημα IV – Μονολεκτικοί και Πολυλεκτικοί όροι

Στο σημείο αυτό, και ως ξεχωριστά αρχεία excel, επισυνάπτονται οι μονολεκτικοί (P2.2-ebooks_lemma_terms_with_wikipedia_info_v6) και οι Πολυλεκτικοί (P2.2-ebooks_multiwords_with_wikipedia_info_v6) όροι που προέκυψαν από την ΕΕ2.

Παράρτημα V – Στατιστικά Στοιχεία Όρων

Στο σημείο αυτό, και ως ξεχωριστά αρχεία (συμπίεσμένο zip αρχείο: P2.2-ctermns.zip), επισυνάπτονται οι Πολυλεκτικοί όροι που προέκυψαν από την εφαρμογή των κανόνων στο σώμα κειμένων με στατιστικά στοιχεία εμφάνισης του πρώτου, του δεύτερου αλλά και των δύο όρων μαζί. Πιο συγκεκριμένα, μέσα στο συμπίεσμένο αρχείο υπάρχουν 24 αρχεία κειμένου:

Όνομα	Τύπος	Συμπίεσμένο μέγεθος	Πρωστό
 terms.gr	Αρχείο GR	12 KB	Όχι
 TERMS_1_1.txt	Έγγραφο κειμένου	186 KB	Όχι
 TERMS_1_2.txt	Έγγραφο κειμένου	50 KB	Όχι
 TERMS_1_3.txt	Έγγραφο κειμένου	44 KB	Όχι
 TERMS_1_4.txt	Έγγραφο κειμένου	11 KB	Όχι
 TERMS_1_5.txt	Έγγραφο κειμένου	26 KB	Όχι
 TERMS_1_6.txt	Έγγραφο κειμένου	21 KB	Όχι
 TERMS_1_7.txt	Έγγραφο κειμένου	4 KB	Όχι
 TERMS_1_8.txt	Έγγραφο κειμένου	58 KB	Όχι
 TERMS_1_9.txt	Έγγραφο κειμένου	5 KB	Όχι
 TERMS_1_10.txt	Έγγραφο κειμένου	5 KB	Όχι
 TERMS_1_11.txt	Έγγραφο κειμένου	2 KB	Όχι
 TERMS_1_12.txt	Έγγραφο κειμένου	5 KB	Όχι
 TERMS_1_13.txt	Έγγραφο κειμένου	10 KB	Όχι
 TERMS_1_14.txt	Έγγραφο κειμένου	6 KB	Όχι
 TERMS_1_15.txt	Έγγραφο κειμένου	7 KB	Όχι
 TERMS_1_16.txt	Έγγραφο κειμένου	1 KB	Όχι
 TERMS_1_17.txt	Έγγραφο κειμένου	2 KB	Όχι
 TERMS_1_18.txt	Έγγραφο κειμένου	1 KB	Όχι
 TERMS_1_19.txt	Έγγραφο κειμένου	1 KB	Όχι
 TERMS_1_20.txt	Έγγραφο κειμένου	0 KB	Όχι
 TERMS_1_21.txt	Έγγραφο κειμένου	0 KB	Όχι
 TERMS_1_22.txt	Έγγραφο κειμένου	18 KB	Όχι
 TERMS_1_23.txt	Έγγραφο κειμένου	0 KB	Όχι

Το αρχείο «terms.gr» περιέχει τους κανόνες (rules), από τους οποίους προήλθαν οι όροι.

Τα υπόλοιπα αρχεία, περιέχει το καθένα, τα αποτελέσματα (όρους) του αντίστοιχου κανόνα, με μια συγκεκριμένη σειρά και με κάποιους αριθμούς που δείχνουν: τον αριθμό εμφανίσεων του πρώτου όρου, τον αριθμό εμφανίσεων του δεύτερου όρου και τον αριθμό των εμφανίσεων και των δύο μαζί (πολυλεκτικός όρος). Π.χ. *Γεωγραφικό/10 μήκος/2* : ο όρος «Γεωγραφικό» έχει εμφανιστεί 10 φορές, ο όρος «μήκος» έχει εμφανιστεί 2 φορές και ο πολυλεκτικός όρος «Γεωγραφικό μήκος» έχει εμφανιστεί 2 φορές.



TERMS_1_1.txt - Σημειωματάριο

Αρχείο	Επεξεργασία	Μορφή	Προβολή	Βοήθεια
Ngrams Flexible Model. #NumOfNGrams: 8272				
κοινωνικό/3	σύνολο/2	2		
κοινωνικό/3	επίπεδο/1	1		
υγρή/1	περίοδος/1	1		
υγρά/2	απόβλητα/1	1		
υγρά/2	καλοκαίρια/1	1		
Μεγάλων/3	Λιμνών/3	3		
τεχνητής/1	λίμνης/1	1		
ανθυγιεινά/1	έλη/1	1		
πατρογονικών/1	εστιών/1	1		
Γεωγραφικό/10	διαμέρισμα/2	2		
Γεωγραφικό/10	μήκος/2	2		
Γεωγραφικό/10	γλώσσάριο/4	4		
Γεωγραφικό/10	πλάτος/2	2		
Καλές/1	Τέχνες/1	1		
μεγαλύτερο/74	μέγεθος/1	1		
μεγαλύτερο/74	προμηθευτή/3	3		
μεγαλύτερο/74	πληθυσμό/1	1		
μεγαλύτερο/74	ποταμό/1	1		
μεγαλύτερο/74	πάχος/1	1		
μεγαλύτερο/74	βάθος/3	3		
μεγαλύτερο/74	νησί/1	1		
μεγαλύτερο/74	ποσοστό/7	7		
μεγαλύτερο/74	υψόμετρο/1	1		
μεγαλύτερο/74	αεροπλάνο/1	1		
μεγαλύτερο/74	τμήμα/7	7		
μεγαλύτερο/74	μέρος/38	38		
μεγαλύτερο/74	θηλαστικό/1	1		
μεγαλύτερο/74	πρόβλημα/1	1		
μεγαλύτερο/74	τούνελ/1	1		
μεγαλύτερο/74	δάσος/1	1		
μεγαλύτερο/74	κράτος/2	2		
μεγαλύτερο/74	βαθμό/1	1		
μεγαλύτερο/74	μήκος/1	1		
μεγαλύτερο/74	μισθό/1	1		
απέραντες/4	πεδιάδες/4	4		
πολλές/52	συρράξεις/1	1		
πολλές/52	λεπτομέρειες/1	1		
πολλές/52	βροχές/6	6		
πολλές/52	συνήθειες/1	1		
πολλές/52	πόλεις/1	1		
πολλές/52	πληροφορίες/2	2		