ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΗΣ ΣΧΕΔΙΑΣΗΣ ΚΑΙ ΠΑΡΑΓΩΓΗΣ

Διπλωματική Εργασία

Ανάπτυξη μορφολογικού ηλεκτρονικού λεξικού της Νέας Ελληνικής γλώσσας για εφαρμογές Επεξεργασίας Φυσικής Γλώσσας

ΟΝΟΜΑΤΕΠΩΝΥΜΟ ΦΟΙΤΗΤΗ:

ΙΑΚΩΒΟΣ ΧΑΡΔΑΛΟΥΠΑΣ

Α.Μ.: 71445820

Επιβλέπων Καθηγητής

Ευάγγελος Παπακίτσος

Ε.ΔΙ.Π. Α΄ Βαθμίδας

ΑΙΓΑΛΕΩ 2023

UNIVERSITY OF WEST ATTICA

Department of
Industrial Design & Production Engineering

Diploma Thesis

Development of a morphological lexicon of the Modern Greek language for Natural Language Processing applications.

Student name and surname

IAKOVOS CHARDALOUPAS

Registration Number

71445820

Supervisor name and surname

Evangelos Papakitsos

Egaleo, 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΗΣ ΣΧΕΔΙΑΣΗΣ ΚΑΙ ΠΑΡΑΓΩΓΗΣ

Διπλωματική Εργασία

Ανάπτυξη μορφολογικού ηλεκτρονικού λεξικού της Νέας Ελληνικής γλώσσας για εφαρμογές Επεξεργασίας Φυσικής Γλώσσας

ΟΝΟΜΑΤΕΠΩΝΥΜΟ ΦΟΙΤΗΤΗ: ΙΑΚΩΒΟΣ ΧΑΡΔΑΛΟΥΠΑΣ

Α.Μ.: 71445820

Μέλη Εξεταστικής Επιτροπής συμπεριλαμβανομένου και του Εισηγητή

Η παρούσα διπλωματική εργασία εγκρίθηκε ομόφωνα από την τριμελή εξεταστική επιτροπή, η οποία ορίστηκε από την Γ.Σ. του Τμήματος Μηχανικών Βιομηχανικής Σχεδίασης και Παραγωγής του Πανεπιστημίου Δυτικής Αττικής, σύμφωνα με το νόμο και τον εγκεκριμένο Οδηγό Σπουδών του τμήματος.

Επιβλέπων: Ε. ΠΑΠΑΚΙΤΣΟΣ

Επιτροπή Αξιολόγησης:

Ε. Παπακίτσος Ε.ΔΙ.Π. Α΄ Βαθμίδας	Evangelos Papakitsos Digitally signed by Evangelos Papakitsos Date: 2023.07.18 11:57:45 +03'00'
Ν. Λάσκαρης Επίκουρος Καθηγητής	Nikolaos Laskaris Digitally signed by Nikolaos Laskaris Date: 2023.07.21 09:58:33 +03'00'
Χ. Δρόσος Ε.ΔΙ.Π. Α΄ Βαθμίδας	Christos Drosos Digitally signed by Christos Drosos Date: 2023.07.18 13:02:27 +03'00'

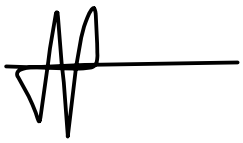
ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΠΤΥΧΙΑΚΗΣ/ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο κάτωθι υπογεγραμμένος Χαρδαλούπας Ιάκωβος του Ιωάννη, με αριθμό μητρώου 71445820 φοιτητής του Πανεπιστημίου Δυτικής Αττικής της Σχολής Μηχανικών του Τμήματος Βιομηχανικής Σχεδίασης και Παραγωγής, δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της πτυχιακής/διπλωματικής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Ο Δηλών



ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα με την παρούσα αναφορά να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή Ευάγγελο Παπακίτσο για την υποστήριξη, την καθοδήγηση αλλά και την υπομονή του κατά την διάρκεια της εκπόνησης της εν λόγω πτυχιακής εργασίας.

Επίσης οφείλω ένα μεγάλο ευχαριστώ στον αδερφό μου Διονύσιο Χαρδαλούπα για τη βοήθεια του στη συγγραφή και στους γονείς μου για τη στήριξη που μου παρείχαν τόσα χρόνια.

Μετά τιμής

Ιάκωβος Χαρδαλούπας

Περιεχόμενα

Ευρετήριο εικόνων

Περίληψη

Η εμφάνιση των ψηφιακών επιστημών και η ευρεία χρήση των υπολογιστών επηρέασε σημαντικά τη μελέτη της γλώσσας, σε τέτοιο σημείο ώστε σήμερα να θεωρείται αδύνατη η γλωσσική επεξεργασία χωρίς την υποστήριξη των ηλ. υπολογιστών. Ο στόχος της Επεξεργασίας Φυσικής Γλώσσας άλλωστε είναι να γεφυρώσει το χάσμα μεταξύ της ανθρώπινης γλώσσας και των ηλ. Υπολογιστών, κι έτσι η μορφολογική επεξεργασία έχει προσφέρει σημαντική βοήθεια και διευκόλυνση για τη γλωσσική κατανόηση και παραγωγή των λέξεων.

Εφαρμογή της μορφολογικής επεξεργασίας αποτελεί η ανάπτυξη ενός μορφολογικού ηλεκτρονικού λεξικού για τη Νέα Ελληνική γλώσσα. Ένα τέτοιο λεξικό περιλαμβάνει πληροφορίες σχετικά με τη μορφολογία των λέξεων, όπως τα ρήματα, τα ουσιαστικά, τα επίθετα και τα επιρρήματα, και μπορεί να χρησιμοποιηθεί για την ανάλυση και την κατανόηση του κειμένου. Η ανάπτυξη ενός τέτοιου λεξικού απαιτεί τη συλλογή μεγάλου όγκου δεδομένων και την ανάλυσή τους για την αναγνώριση των διαφόρων μορφών και συντακτικών χαρακτηριστικών των λέξεων. Ενώ επίσης απαιτεί τη χρήση αυτόματων μεθόδων μάθησης μηχανής, όπως τους αλγόριθμους μηχανικής μάθησης.

Στην παρούσα εργασία θα μελετηθεί μια ήδη υπάρχουσα βάση δεδομένων της Νέας Ελληνικής γλώσσας και θα επιχειρηθεί να βελτιωθεί η δομή με την υλοποίησή της σε υπολογιστικά φύλλα, σκοπεύοντας να γίνει εύχρηστη και εύκολα διαχειρίσιμη.

Abstract

The emergence of digital sciences and the widespread use of computers significantly affected the study of language, to such an extent that today language processing is considered impossible without the support of computers. After all, the goal of Natural Language Processing is to bridge the gap between human language and computers, so morphological processing has provided significant help and facilitation for linguistic understanding and word production.

An application of morphological processing is the development of a morphological electronic dictionary for the Modern Greek language. Such a dictionary includes information about the morphology of words, such as verbs, nouns, adjectives and adverbs, and can be used to analyze and understand text. Developing such a dictionary requires collecting a large amount of data and analyzing it to identify the various forms and syntactic features of words. While, it also requires the use of automatic machine learning methods such as machine learning algorithms.

In this work, an already existing database of the Modern Greek language will be studied and an attempt will be made to improve the structure, by implementing it in spreadsheets, aiming to make it easy to use and easily manageable.

Εισαγωγή

Από το τέλος του 20ου αιώνα, η πληροφορική έχει σημαντικό αντίκτυπο στις γλωσσικές επιστήμες. Η εποχή χαρακτηρίστηκε από την εμφάνιση των ψηφιακών επιστημών και από την ευρεία χρήση των υπολογιστών. Γενικότερα επηρεάστηκε κάθε ανθρώπινη δραστηριότητα και ειδικότερα η μελέτη της γλώσσας.

Σήμερα η γλωσσική επεξεργασία είναι αδιανόητη χωρίς την υποστήριξη των υπολογιστών.

Η επίδραση της πληροφορικής στις γλωσσικές επιστήμες επικύρωσε διαφορετικές νέες θεωρίες για τις γλωσσικές λειτουργίες στα διάφορα επίπεδα γλωσσικής περιγραφής (φωνητική/φωνολογία, μορφοσύνταξη, σημασιολογία, πραγματολογία).

Η συνεργασία της πληροφορικής με τη γλωσσολογία ανήκει στον τομέα της Επεξεργασίας Φυσικής Γλώσσας, κλάδο της οποία αποτελεί η Μορφολογική Επεξεργασία.

Άλλωστε η γλωσσολογική ανάλυση αποτελείται από διακριτές διαδικασίες, όπως

1) Η Φωνητική-Φωνολογία, η οποία αναλύει τα ηχητικά και λειτουργικά συστατικά της γλώσσας (φθόγγους, φωνήματα).

2) Η Μορφολογία, η οποία αναλύει τα γραμματικά συστατικά των λέξεων.

3) Η Σύνταξη, η οποία αναλύει τον τρόπο ένωσης των λέξεων για να σχηματίσουν φράσεις και προτάσεις.

4) Η Σημασιολογία, η οποία μελετά τη σημασία των λέξεων (Lyons 1995 στο Παπακίτσος 2000)

5) Η Πραγματολογία, η οποία μελετά τον τρόπο με τον οποίο το περιβάλλον επιδρά στην ερμηνεία μιας πρότασης (Φιλιππάκη-Warburton 1992 στο Παπακίτσος 2000).

Η Μορφολογική Επεξεργασία είναι ένας κλάδος της Επεξεργασίας Φυσικής Γλώσσας που ασχολείται με την ανάλυση και την επεξεργασία της μορφολογικής δομής των λέξεων σε μια φυσική γλώσσα. Η μορφολογία αφορά τη μορφή και τη σύνθεση των λέξεων, όπως τα προθήματα, τα επιθήματα και οι καταλήξεις, καθώς και τους κανόνες για την απόκτηση νέων λέξεων μέσω παραγωγής και σύνθεσης.

Στη Μορφολογική Επεξεργασία χρησιμοποιούνται κανόνες και αλγόριθμοι για την αναγνώριση και ανάλυση των μορφολογικών χαρακτηριστικών των λέξεων, όπως η

κλίση, η αριθμητική πτώση, ο χρόνος και ο τρόπος σε γλωσσικά κατηγορήματα. Επίσης, η Μορφολογική Επεξεργασία ασχολείται με την αντιμετώπιση προβλημάτων όπως η αναγνώριση και ομαδοποίηση λέξεων σε μορφολογικές κατηγορίες και η αποκατάσταση της αρχικής μορφής μιας λέξης από την κατακερματισμένη μορφή της (stemming).

Η μορφολογική επεξεργασία συνδέεται στενά με την επεξεργασία φυσικής γλώσσας, η οποία χρησιμοποιεί υπολογιστικές τεχνικές με σκοπό την εκμάθηση, την κατανόηση και την παραγωγή περιεχομένου ανθρώπινης γλώσσας. Οι πρώιμες υπολογιστικές προσεγγίσεις στη γλωσσική έρευνα επικεντρώθηκαν στην αυτοματοποίηση της ανάλυσης της γλωσσικής δομής της γλώσσας και στην ανάπτυξη βασικών τεχνολογιών, όπως η αυτόματη μετάφραση, η αναγνώριση ομιλίας και η σύνθεση ομιλίας. Οι σημερινοί ερευνητές βελτιώνουν και χρησιμοποιούν τέτοια εργαλεία σε πραγματικές εφαρμογές, δημιουργώντας συστήματα προφορικού διαλόγου και μηχανές μετάφρασης ομιλίας σε ομιλία, εξόρυξη μέσων κοινωνικής δικτύωσης για πληροφορίες σχετικά με την υγεία ή τα οικονομικά και τον εντοπισμό απόψεων και συναισθημάτων για προϊόντα και υπηρεσίες.

Η παρούσα εργασία θα εστιάσει στο χώρο της Επεξεργασίας Φυσικής Γλώσσας και ιδιαίτερα στη Μορφολογική Επεξεργασία. Πιο συγκεκριμένα, αφορά σε πρόταση βελτίωσης της υπολογιστικής αντιμετώπισης της κλίσης της Νέας Ελληνικής Δημοτικής (Ell) (Παπακίτσος 2000), σχεδιάζοντας και υλοποιώντας ένα εύχρηστο μορφολογικό ηλεκτρονικό λεξικό της νέας ελληνικής γλώσσας. Πιο συγκεκριμένα, η εισαγωγή της νέας βάσης δεδομένων στο Excel σκοπεύει να κάνει τη διαχείριση της προηγουμένως δυσκίνητης και ξεπερασμένης συλλογής αρχείων κειμένου εύκολα διαχειρίσιμη και εύχρηστη. Αυτός ο μετασχηματισμός προβλέπεται ότι θα επιφέρει πολλά οφέλη και σημαντική βελτίωση στη χρηστικότητα και την προσβασιμότητα.

Θα δομηθεί σε 5 κεφάλαια. Αναλυτικά στο πρώτο κεφάλαιο αναλύεται η έννοια της επεξεργασίας φυσικής γλώσσας ενώ δίνεται ο ορισμός και η σημασία της Μορφολογικής Επεξεργασίας. Στο δεύτερο κεφάλαιο παρουσιάζεται η υπάρχουσα βάση δεδομένων του ηλεκτρονικού λεξικού, η κατηγοριοποίηση, κωδικοποίηση και οργάνωση των δεδομένων – λημμάτων. Στο τρίτο κεφάλαιο περιγράφεται η νέα δομή των αρχείων υλοποίησης σε φύλλα υπολογιστικά του Excel, παρουσιάζοντας τα βασικά εργαλεία που θα επιτρέψουν να πραγματοποιηθεί η συγκεκριμένη αλλαγή, δηλαδή τη γλώσσα προγραμματισμού python και τις βιβλιοθήκες που διαθέτει η εν λόγω γλώσσα.

Επιπλέον παρουσιάζονται δύο πολύ χρήσιμες για την κωδικοποίηση των δεδομένων βιβλιοθήκες, η Unicodedata και pandas.

Στο τέταρτο κεφάλαιο πραγματοποιείται σύγκριση της παλιάς μορφής με τη νέα μορφή της βάσης των δεδομένων, δίνοντας παραδείγματα της πολυπλοκότητας της παλιάς βάσης και της νέας βάσης, η οποία είναι πλέον βελτιωμένη κι εύχρηστη. Τέλος στο πέμπτο κεφάλαιο συζητούνται περαιτέρω ορισμένα ζητήματα τα οποία χρήζουν ανακεφαλαίωσης κι εξάγονται τα συμπεράσματα της παρούσας εργασίας.

Κεφάλαιο 1 Εισαγωγή στην επεξεργασία φυσικής γλώσσας

Στο κεφάλαιο αυτό γίνεται μια επισκόπηση στο πεδίο της επεξεργασίας φυσικής γλώσσας και της μορφολογικής επεξεργασίας, παρουσιάζοντας τις βασικές έννοιες και τεχνικές της αλλά και την ανάγκη για την οποία είναι απαραίτητα τα μορφολογικά λεξικά.

1.1 Η έννοια της επεξεργασίας φυσικής γλώσσας

Η Επεξεργασία Φυσικής Γλώσσας (NLP) είναι ένας τομέας της επιστήμης των υπολογιστών και της τεχνητής νοημοσύνης η οποία εστιάζει στην αλληλεπίδραση μεταξύ των υπολογιστών και της ανθρώπινης γλώσσας. Περιλαμβάνει την ανάπτυξη αλγορίθμων και μοντέλων που επιτρέπουν στους υπολογιστές να κατανοούν, να ερμηνεύουν και να δημιουργούν φυσική γλώσσα, με τρόπο που να έχει νόημα και χρησιμότητα για τον άνθρωπο (Goel 2017).

Η Επεξεργασία Φυσικής Γλώσσας περιλαμβάνει ένα ευρύ φάσμα εργασιών και εφαρμογών, όπως:

Κατανόηση κειμένου: Στην κατηγορία αυτή περιλαμβάνονται εργασίες όπως η ανάλυση προτάσεων, η εξαγωγή νοημάτων και σχέσεων μεταξύ λέξεων και ο προσδιορισμός οντοτήτων και των ιδιοτήτων τους. Επίσης περιλαμβάνει τεχνικές όπως η συντακτική και σημασιολογική ανάλυση, η αναγνώριση ονομαστικών οντοτήτων και η ανάλυση συναισθήματος.

Δημιουργία γλώσσας: Η NLP ασχολείται επίσης με τη δημιουργία γλώσσας που μοιάζει με ανθρώπινη, είτε πρόκειται για τη δημιουργία συνεκτικών προτάσεων, τη σύνταξη περιλήψεων ή την παραγωγή απαντήσεων φυσικής γλώσσας σε chatbots ή εικονικούς βοηθούς.

Μηχανική μετάφραση: Διαδραματίζει κρίσιμο ρόλο στην ανάπτυξη συστημάτων που μπορούν να μεταφράσουν αυτόματα κείμενο από τη μια γλώσσα στην άλλη. Στην κατηγορία αυτή περιλαμβάνεται η κατανόηση της δομής και της σημασίας των προτάσεων στη γλώσσα-πηγή και τη δημιουργία ισοδύναμων προτάσεων στη γλώσσα-στόχο.

Ανάκτηση πληροφοριών: Οι τεχνικές Επεξεργασίας Φυσικής Γλώσσας χρησιμοποιούνται στις μηχανές αναζήτησης για την κατανόηση των ερωτημάτων των

χρηστών και την ανάκτηση σχετικών πληροφοριών από τις τεράστιες ποσότητες δεδομένων κειμένου. Εδώ περιλαμβάνονται τεχνικές όπως η αναζήτηση εγγράφων, η κατανόηση ερωτημάτων και η κατάταξη συνάφειας (IBM 2023).

Αναγνώριση και σύνθεση ομιλίας: Η Επεξεργασία Φυσικής Γλώσσας επεκτείνεται στην επεξεργασία της προφορικής γλώσσας, συμπεριλαμβανομένων εργασιών όπως η αναγνώριση ομιλίας (μετατροπή προφορικών λέξεων σε γραπτό κείμενο) και η σύνθεση ομιλίας (δημιουργία προφορικού αποτελέσματος από γραπτό κείμενο).

Θα πρέπει επίσης να αναφερθεί ότι η Επεξεργασία Φυσικής Γλώσσας χρησιμοποιεί διάφορες τεχνικές και μοντέλα, όπως η μηχανική μάθηση, η βαθιά μάθηση και οι στατιστικές μέθοδοι. Βασίζεται σε σχολιασμένους γλωσσικούς πόρους, όπως μεγάλα σώματα και γλωσσικές βάσεις δεδομένων, για την εκπαίδευση και την αξιολόγηση μοντέλων (Lauriola et al. 2022).

Ο στόχος της Επεξεργασίας Φυσικής Γλώσσας είναι να γεφυρώσει το χάσμα μεταξύ της ανθρώπινης γλώσσας και των υπολογιστών, επιτρέποντας στις μηχανές να κατανοούν, να ερμηνεύουν και να επικοινωνούν αποτελεσματικά με τους ανθρώπους στη φυσική γλώσσα, ανοίγοντας δυνατότητες για εφαρμογές σε τομείς όπως η ανάκτηση πληροφοριών, η ανάλυση συναισθημάτων, η απάντηση ερωτήσεων, η γλώσσα, η μετάφραση, και πολλά άλλα (Goel 2017).

1.2 Η έννοια της Μορφολογικής επεξεργασίας

Η Μορφολογική Επεξεργασία αναφέρεται στις γνωστικές και γλωσσικές διαδικασίες που εμπλέκονται στην ανάλυση και τον χειρισμό της δομής των λέξεων, ιδιαίτερα όσον αφορά τις νοηματικές ενότητες που ονομάζονται μορφήματα. Τα μορφήματα είναι οι μικρότερες γλωσσικές μονάδες που μεταφέρουν σημασιολογικές ή γραμματικές πληροφορίες (Schreuder & Baayen 1995).

Στη γλώσσα, οι λέξεις αποτελούνται από ένα ή περισσότερα μορφήματα που μπορούν να ταξινομηθούν σε διαφορετικούς τύπους. Για παράδειγμα, στη λέξη "unhappiness", τα μορφήματα "un-" και "-ness" μπορούν να αναγνωριστούν ως νοηματικές μονάδες με τις δικές τους σημασιολογικές και γραμματικές λειτουργίες. Η μορφολογική επεξεργασία περιλαμβάνει την αναγνώριση και τον χειρισμό αυτών των μορφών για την κατανόηση της σημασίας και της δομής των λέξεων.

Οι κύριες πτυχές της μορφολογικής επεξεργασίας περιλαμβάνουν:

Μορφολογική Ανάλυση: Περιλαμβάνει τη διάσπαση των λέξεων στα μορφήματα που τις αποτελούν και τον προσδιορισμό των γραμματικών τους ιδιοτήτων. Για παράδειγμα, στη λέξη "cats", τα μορφήματα "cat" και "-s" μπορούν να αναλυθούν για να προσδιοριστεί η βασική μορφή "cat" και ο δείκτης πληθυντικού αριθμού "-s".

Μορφολογική παραγωγή: Αναφέρεται στη διαδικασία συνδυασμού μορφημάτων για τη δημιουργία νέων λέξεων. Για παράδειγμα, η προσθήκη του προθήματος "un-" στη βασική λέξη "happy" δημιουργεί τη λέξη "unhappy" (Marantz 2013).

Μορφολογικοί κανόνες: Οι γλώσσες έχουν συγκεκριμένους κανόνες και μοτίβα για το συνδυασμό μορφημάτων. Αυτοί οι κανόνες διέπουν διαδικασίες όπως η κλίση (προσθήκη γραμματικών δεικτών) και η παραγωγή (δημιουργία νέων λέξεων προσθέτοντας επιθήματα). Για παράδειγμα, στα αγγλικά, ο πληθυντικός δείκτης "-s" προστίθεται σε ουσιαστικά για να υποδηλώνει περισσότερα από ένα (για παράδειγμα cat-cats), ενώ το παράγωγο επίθημα "-ness" μπορεί να προστεθεί στα επίθετα για να σχηματίσει αφηρημένα ουσιαστικά (για παράδειγμα happy-happiness).

Μορφολογική Αντίληψη: Αναφέρεται στη συνειδητή κατανόηση και γνώση ενός ατόμου σχετικά με τις μορφολογικές δομές και τις διαδικασίες στη γλώσσα. Η μορφολογική αντίληψη παίζει ρόλο στην ανάπτυξη του λεξιλογίου, στην αναγνωστική κατανόηση και στις δεξιότητες ορθογραφίας, καθώς βοηθά τα άτομα να αναγνωρίζουν τις σχέσεις μεταξύ των λέξεων και να αντλούν νόημα από τα μέρη της λέξης.

Κλιτική μορφολογία (inflectional morphology): Είναι γλωσσολογικός όρος που αναφέρεται στη διαδικασία τροποποίησης ενός λήμματος ή μιας λέξης σε διάφορες μορφές ή γραμματικές καταλήξεις, σύμφωνα με τα γραμματικούς κανόνες της γλώσσας. Αφορά δηλαδή μόνο την κατάληξη και όχι τα υπόλοιπα μορφήματα της λέξης.

Σε πολλές γλώσσες, οι λέξεις αλλάζουν μορφολογικά για να εκφράσουν διάφορες πτώσεις, γένη, αριθμούς, χρόνους, καταστάσεις. Οι μορφολογικές αλλαγές συνήθως επηρεάζουν την κατάληξη ή την εσωτερική δομή της λέξης, ενώ ο βασική ρίζα της λέξης μπορεί να παραμείνει αμετάβλητη. Για παράδειγμα, στην ελληνική γλώσσα, οι κανόνες μορφολογικής κλίσης μετατρέπουν τη βασική μορφή ενός επιθέτου όταν αυτό χρησιμοποιείται σε διαφορετικά γένη. Έτσι έχουμε «χαρούμενος» για το αρσενικό γένος, «χαρούμενη» για το θηλυκό και «χαρούμενο» για το ουδέτερο. Γίνεται

κατανοητό ότι η κλιτική μορφολογία ποικίλλει σημαντικά μεταξύ των γλωσσών, διότι οι διαφορετικές γλώσσες χρησιμοποιούν διαφορετικά συστήματα και μοτίβα κλίσης (Ralli 2002).

Η μορφολογική επεξεργασία είναι μια σημαντική πτυχή της γλωσσικής κατανόησης και παραγωγής. Συμβάλλει στην ικανότητά μας να κατανοούμε και να παράγουμε λέξεις, να αναγνωρίζουμε μορφές λέξεων και να αντλούμε νόημα από τη δομή των λέξεων. Η έρευνα στη μορφολογική επεξεργασία έχει επιπτώσεις στην κατάκτηση της γλώσσας, στην ανάπτυξη της ανάγνωσης και στην κατανόηση των γνωστικών μηχανισμών που διέπουν τη χρήση της γλώσσας (Carlisle 2000).

1.2 Η ανάγκη για μορφολογικά ηλεκτρονικά λεξικά

Τα μορφολογικά ηλεκτρονικά λεξικά είναι πολύτιμοι γλωσσικοί πόροι που παρέχουν πληροφορίες σχετικά με τις μορφολογικές ιδιότητες των λέξεων σε μια γλώσσα. Έχουν σχεδιαστεί για να βοηθούν σε διάφορες εργασίες κι εφαρμογές επεξεργασίας φυσικής γλώσσας. Οι λόγοι για τους οποίους είναι απαραίτητα τα μορφολογικά ηλεκτρονικά λεξικά είναι οι εξής (Hijzelendoorn & Cremers 2009):

Ανάλυση και δημιουργία λέξεων: Τα μορφολογικά λεξικά διευκολύνουν την ανάλυση και τη δημιουργία λέξεων, παρέχοντας πληροφορίες σχετικά με τη μορφολογική τους δομή, όπως για παράδειγμα τα προθήματα, τα επιθήματα και τις διάφορες μορφές ρίζας. Αυτές οι πληροφορίες είναι κρίσιμες για εργασίες όπως η μορφολογική ανάλυση, η εύρεση λημμάτων και ο σχηματισμός λέξεων.

Γλωσσική κατανόηση και αποσαφήνιση: Τα μορφολογικά λεξικά βοηθούν στην αποσαφήνιση λέξεων με πολλαπλές μορφές ή έννοιες. Παρέχοντας μορφολογική ανάλυση και επισήμανση, βοηθούν στην επίλυση ασαφειών σε εργασίες κατανόησης φυσικής γλώσσας, όπως η επισήμανση μέρους του λόγου και η σημασιολογική ανάλυση.

Επέκταση λεξιλογίου: Τα μορφολογικά λεξικά επιτρέπουν την επέκταση του λεξιλογίου παρέχοντας πληροφορίες σχετικά με τις σχέσεις παραγωγής μεταξύ των λέξεων. Μπορούν να βοηθήσουν στην αυτόματη δημιουργία νέων μορφών λέξεων, εφαρμόζοντας μορφολογικούς κανόνες και μετασχηματισμούς.

Εκμάθηση και Διδασκαλία Γλωσσών: Τα μορφολογικά λεξικά είναι ωφέλιμα για τους μαθητές και τους καθηγητές γλωσσών. Παρέχουν πληροφορίες για τη δομή και το σχηματισμό των λέξεων, βοηθώντας στην απόκτηση λεξιλογίου, στην κατανόηση των οικογενειών λέξεων και στην κατανόηση γραμματικών προτύπων.

Υπολογιστική Γλωσσολογία κι Εφαρμογές Επεξεργασίας Φυσικής Γλώσσας: Στην υπολογιστική γλωσσολογία και την επεξεργασία φυσικής γλώσσας, τα μορφολογικά λεξικά χρησιμεύουν ως βασικοί πόροι για την ανάπτυξη γλωσσικών μοντέλων, συστημάτων μηχανικής μετάφρασης, ορθογραφικών ελέγχων, συστημάτων ανάκτησης πληροφοριών και άλλων εφαρμογών επεξεργασίας γλώσσας. Συμβάλλουν στην ακρίβεια και την αποτελεσματικότητα αυτών των συστημάτων, παρέχοντας μορφολογικές πληροφορίες και κανόνες.

Πηγή για Γλωσσική Έρευνα: Τα μορφολογικά λεξικά χρησιμεύουν ως πολύτιμοι πόροι για τη γλωσσική έρευνα, δίνοντας τη δυνατότητα στους ερευνητές να διερευνήσουν μορφολογικά πρότυπα, να αναλύσουν τις διαδικασίες σχηματισμού λέξεων και να μελετήσουν τη γλωσσική παραλλαγή και αλλαγή. Παρέχουν μια δομημένη αποθήκη μορφολογικών δεδομένων για γλωσσική ανάλυση και σύγκριση.

Συνολικά, τα μορφολογικά ηλεκτρονικά λεξικά διαδραματίζουν ζωτικό ρόλο στην ενίσχυση της αποτελεσματικότητας και της ακρίβειας διαφόρων εργασιών επεξεργασίας γλώσσας, υποστηρίζοντας την εκμάθηση γλωσσών και διευκολύνοντας τη γλωσσική έρευνα. Παρέχουν πολύτιμες πληροφορίες σχετικά με τις μορφολογικές ιδιότητες των λέξεων, επιτρέποντας την καλύτερη κατανόηση, παραγωγή και χειρισμό λέξεων σε υπολογιστικά και γλωσσικά πλαίσια (Maxwell & Poser 2004).

Κεφάλαιο 2 Παρουσίαση της Μορφολογικής Βάσης Δεδομένων

Στο κεφάλαιο αυτό παρουσιάζεται μια επισκόπηση της δημιουργίας και της διαχείρισης της παλιάς μορφολογικής βάσης δεδομένων. Εξηγούνται οι αρχές της οργάνωσης και της δομής της βάσης δεδομένων, καθώς και οι διάφορες πληροφορίες που αποθηκεύονται για κάθε λέξη, όπως η μορφολογική πληροφορία, οι κλίσεις, οι απαραίτητες συντακτικές πληροφορίες και άλλες σχετικές παράμετροι. Επιπλέον, παρουσιάζονται οι τεχνικές και οι μεθοδολογίες που χρησιμοποιούνται για τη δημιουργία και τη διαχείριση της μορφολογικής βάσης δεδομένων, περιλαμβάνοντας τη συλλογή και την επεξεργασία των δεδομένων ή την ανάπτυξη των αλγορίθμων.

2.1 Το λεξικό ως βάση δεδομένων

Το Ηλεκτρονικό Λεξικό αποτελεί μια βάση δεδομένων η οποία περιλαμβάνει τα χαρακτηριστικά των λέξεων στη Νέα Ελληνική γλώσσα. Αυτή η βάση δεδομένων υποστηρίζει τους μορφολογικούς επεξεργαστές, που είναι υπεύθυνοι για την ανάλυση και την επεξεργασία της μορφής των λέξεων.

Η ανάπτυξη του Ηλεκτρονικού Λεξικού πραγματοποιήθηκε σε δύο στάδια. Αρχικά, πραγματοποιήθηκε ένα πιλοτικό στάδιο σε μικρή κλίμακα, όπου εισήχθησαν 500 λήμματα. Αυτό το στάδιο είχε ως στόχο να ελεγχθούν οι δυνατότητες οργάνωσης και σχεδίασης που χρησιμοποιήθηκαν.

Στη συνέχεια, η ανάπτυξη συνεχίστηκε σε μεγάλη κλίμακα, όπου εισήχθηκε ένα μεγάλο πλήθος λημμάτων. Αυτό το στάδιο είχε ως στόχο να εκτιμηθούν οι συνολικές επιδόσεις του συστήματος και να γίνουν οι απαραίτητες ρυθμίσεις για τη βελτίωση του.

Κατά τη διάρκεια της ανάπτυξης, επιλέχθηκαν τα δεδομένα που θα εισαχθούν στο Λεξικό, ο τρόπος κωδικοποίησης καθώς και ο τρόπος οργάνωσης τους.

2.2 Κατηγοριοποίηση δεδομένων

Τα λήμματα της βάσης δεδομένων αποτελούν οι κατηγορίες που ακολουθούν:

- Προθήματα, (παρα-, κατα-, υπο-) κι επιθήματα (ιζ-, -ευ-, -ικ-)
- ρίζες/θέματα (χρον-, λογ-, γραφ-)

- καταλήξεις (-ος, -ω, -ομαι)
- άκλιτα (ποτε, ισως, και)

Ιδιαιτερότητες παρουσιάζουν τα σύνθετα θέματα (υπογραφ-, καταλογ-), τα άκλιτα, οι λέξεις με ανώμαλη κλίση (το ρήμα είμαι), τα οποία καταχωρούνται ως ολόκληρες λέξεις και δεν δημιουργούνται διακριτές κλιτικές κατηγορίες. Ιδιαίτερη περίπτωση αποτελούν τα συνθετικά επιρρήματα «αστραπηδόν», «βροχηδόν», «σωρηδόν», τα οποία περιέχουν το επίθημα «-δόν» και κάποιο ουσιαστικό καθώς και οι συντετμημένες μορφές ρημάτων «φέρτον», «πέστα», «θάρθω», «νάρθω», οι οποίες δεν ακολουθούν τους συνηθισμένους κανόνες της ελληνικής γλώσσας. Ωστόσο όλες οι προαναφερθείσες περιπτώσεις δεν αντιμετωπίστηκαν ως ξεχωριστές κατηγορίες.

Από την άλλη, τα λήμματα της ελληνικής γλώσσας έχουν ορθογραφική και φωνολογική μορφή. Η ορθογραφική μορφή αναφέρεται στον τρόπο με τον οποίο γράφονται τα λήμματα, ενώ η φωνολογική μορφή αφορά τις πτυχές της προφοράς και τονισμού (Παπακίτσος 2000).

Οι ιδιότητες των λημμάτων είναι οι εξής:

- *Σημασιολογικές ιδιότητες*: Αναφέρονται στη σημασία και τη σημασιολογική κατηγορία του λήμματος. Αυτές μπορεί να είναι χρονικές (ρήματα παρελθόντα), τοπικές (ονόματα τοποθεσιών), ποσοτικές (αριθμητικά) και άλλες κατηγορίες.
- *Μορφοσυντακτικές ιδιότητες*: Αναφέρονται στα γραμματικά χαρακτηριστικά των λημμάτων. Αυτές περιλαμβάνουν το γένος (αρσενικό, θηλυκό, ουδέτερο), την πτώση (ονομαστική, γενική, αιτιατική, κλπ.), τον αριθμό (ενικός, πληθυντικός) και άλλες γραμματικές ιδιότητες.
- *Φωνολογικές ιδιότητες*: Αφορούν τον τρόπο προφοράς και τονισμού των λημμάτων.

Για την αποθήκευση των λημμάτων σε βάση δεδομένων χρησιμοποιούνται διάφορα χαρακτηριστικά και ιδιότητες. Τα κύρια χαρακτηριστικά που κωδικοποιούνται είναι:

Ορθογραφική μορφή: Η κωδικοποίηση περιλαμβάνει την ορθογραφική μορφή του λήμματος, δηλαδή τον τρόπο που γράφεται.

Μορφοσυντακτικές ιδιότητες: Κωδικοποιούνται οι μορφοσυντακτικές ιδιότητες των λημμάτων, όπως το γένος, η πτώση, ο αριθμός και άλλες γραμματικές πληροφορίες.

Τονισμός: Ο τονισμός του λήμματος κωδικοποιείται για να δείξει την έμφαση στην προφορά της λέξης.

Σχέσεις (αλλόμορφα): Καταγράφονται οι σχέσεις των λημμάτων μεταξύ τους, όπως οι μορφολεξικοί και οι μεταπλαστικοί νόμοι. Αυτό βοηθά στην ανάκτηση και αντιστοίχιση των λημμάτων κατά την αναζήτηση και την επεξεργασία των δεδομένων.

Παραγωγή, σύνθεση και κλίση: Αποθηκεύονται πληροφορίες σχετικά με την παραγωγή και την κλίση δηλαδή ο τρόπος που κλίνεται μία λέξη και πως από το θέμα της, παράγονται άλλες λέξεις.

Στο πλαίσιο της σχεδίασης της βάσης δεδομένων, κάθε λήμμα συνοδεύεται από ένα μοναδικό αριθμό-κλειδί, το οποίο λειτουργεί ως ταυτότητα για το λήμμα. Η δημιουργία αυτού του κλειδιού ακολουθεί μια τυπική διαδικασία και προσφέρει τη δυνατότητα μοναδικής αναγνώρισης για λήμματα με κοινή μορφή αλλά διαφορετικές ιδιότητες. Για παράδειγμα, τα λήμματα «οδηγ-ώ» και «οδηγ-ός» μπορεί να έχουν διαφορετικές ιδιότητες, αλλά μπορούν να συσχετιστούν με ένα μοναδικό αριθμό-κλειδί.

Το κλειδί παρέχει μια ομοιόμορφη περιγραφή του λήμματος, ανεξάρτητα από το μέγεθός του. Συνήθως αποτελείται από λίγα ψηφία, 2 ψηφιολέξεις έναντι των περίπου 6 ψηφίων που αντιστοιχούν στον μέσο όρο μεγέθους των λημμάτων. Επιπλέον, το κλειδί συνδέει το λήμμα με τις υπόλοιπες ιδιότητές του, καθώς η κύρια μνήμη ενδέχεται να μην επαρκεί για την αποθήκευση όλων των δεδομένων.

Το διεθνές πρότυπο λεξικών βάσεων δεδομένων EAGLES, περιλαμβάνει όλες τις ιδιότητες των λημμάτων που αναφέρθηκαν ανωτέρω, πλην των κλειδιών.

2.3 Κωδικοποίηση δεδομένων

Κατά την ανάπτυξη της Ηλεκτρονικής βάσης δεδομένων, σημαντικό ζήτημα αποτελεί η κωδικοποίηση των δεδομένων-λημμάτων, η οποία θα επηρεάζει την αποτελεσματικότητα και την απόδοση της αναζήτησης.

Κατά συνέπεια είναι απαραίτητο να αποφασιστεί η κατάλληλη μορφή για την καταχώρηση των λημμάτων, η κωδικοποίηση των ιδιοτήτων και σχέσεων δηλαδή ο τρόπος παρουσίασης των μορφοσυντακτικών, σημασιολογικών και φωνολογικών ιδιοτήτων, καθώς και ο τρόπος αποθήκευσης και διαχείρισης των δεδομένων.

Στη βάση δεδομένων, οι τιμές των ιδιοτήτων κωδικοποιούνται συνήθως ως αριθμοί ή χαρακτήρες. Τα λήμματα κωδικοποιούνται ως ASCII χαρακτήρες που αντιπροσωπεύουν την ορθογραφική τους μορφή.

Οι ιδιότητες των λημμάτων τοποθετήθηκαν ανάλογα με το είδος τους, όπως ρίζες/θέματα, επιθήματα, προθήματα, καταλήξεις και άκλιτα. Η κατηγοριοποίηση των επιθημάτων πραγματοποιήθηκε σε κατηγορίες όπως παραγωγικά, κλιτικά, αριθμητικά, παραθετικά και άλλα, βάσει των γραμματικών χαρακτηριστικών τους.

Γενικά, η κωδικοποίηση αυτή επιτρέπει την αποθήκευση και την αναζήτηση των λημμάτων και των ιδιοτήτων τους με αποδοτικό τρόπο, προσφέροντας τη δυνατότητα αναγνώρισης και αναζήτησης των δεδομένων (Παπακίτσος 2000).

2.4 Σχεδίαση του Ηλεκτρονικού Λεξικού

Η σωστή σχεδίαση της ηλεκτρονικής βάσης έχει ως στόχο την αποτελεσματική υποστήριξη του μορφολογικού αναλυτή. Αρχικά, ο μορφολογικός αναλυτής ανακαλύπτει τη συμβολοσειρά ενός λήμματος, το οποίο αποτελεί το μόρφημα. Έπειτα, μέσω του λήμματος καταλήγει στο κλειδί του. Η διπλή κατεύθυνση του βέλους επιτρέπει την αντίστροφη διαδικασία, δηλαδή από το κλειδί να φτάνει κανείς στη συμβολοσειρά.

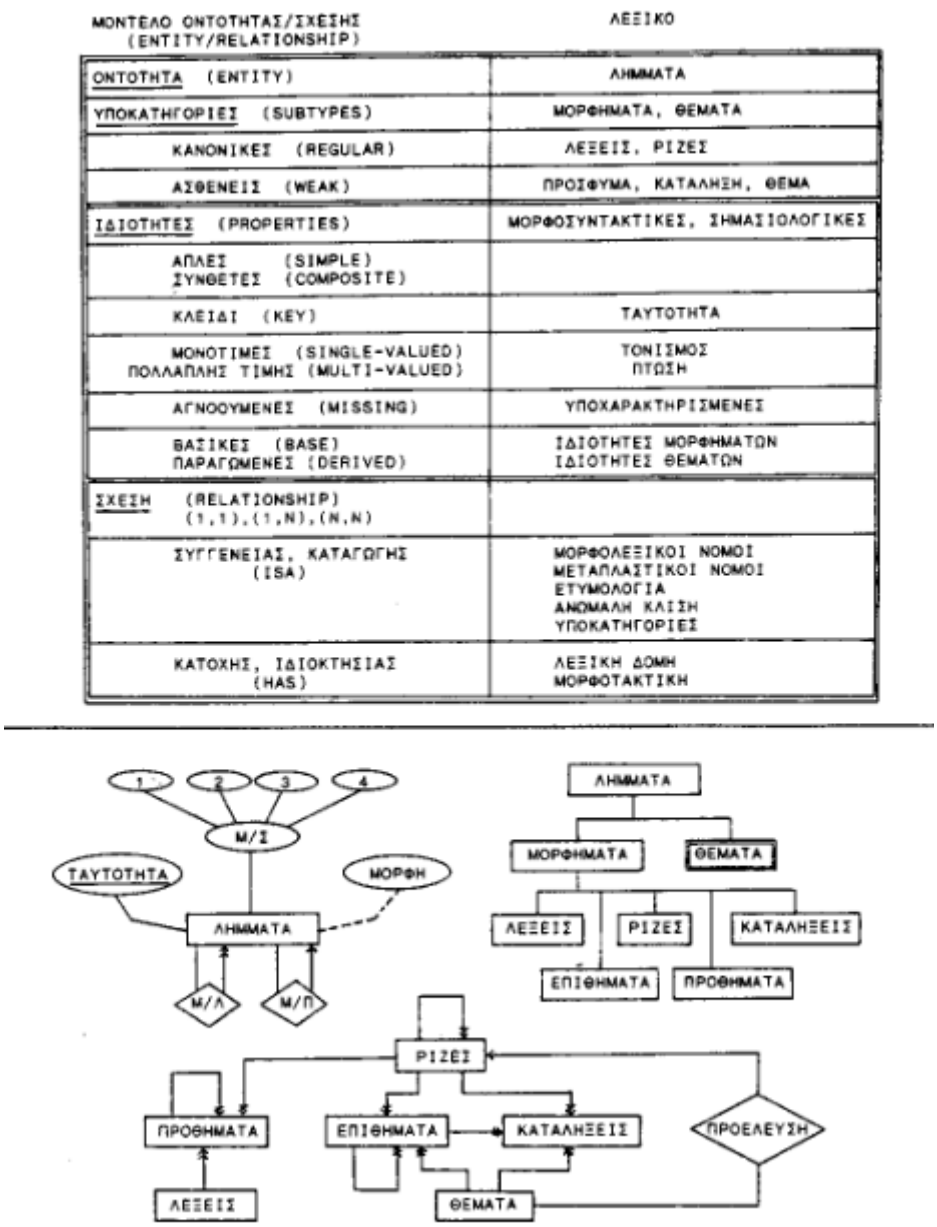
Το κλειδί του λήμματος λειτουργεί ως δείκτης προς τις ιδιότητές του, που περιλαμβάνουν το χαρακτηρισμό του μορφήματος, καθώς και τα παράγωγα ή σύνθετα του λήμματος. Τα παράγωγα και σύνθετα ταξινομούνται σε κατηγορίες, όπως παραγωγικά, κλιτικά, αριθμητικά, παραθετικά, βάσει των γραμματικών τους χαρακτηριστικών.

Στο σημείο αυτό θα πρέπει να αναφερθεί ότι η λειτουργία της βάσης ορίζει τις προδιαγραφές σχεδίασης του συστήματος, δηλαδή τα δεδομένα κωδικοποιούνται με τέτοιον τρόπο ώστε να μειώνεται το λογισμικό του μορφολογικού αναλυτή και να εκτελείται γρήγορα η αναζήτηση στη βάση. Επιπλέον στόχο αποτελεί η επέκταση της βάσης για μελλοντικές χρήσεις. Με βάση αυτά, εξετάστηκαν τα γνωστά μοντέλα σχεδίασης βάσεων δεδομένων, προκειμένου να χρησιμοποιηθεί το πιο κατάλληλο.

Για την οργάνωση και την υλοποίηση της βάσης δεδομένων επιλέχθηκαν τα δενδρικά λεξικά (ROOT/CONTINUATION DATABASES). Ενώ για τις ανάγκες της

Επεξεργασίας Φυσικής Γλώσσας καταλληλότερο μοντέλο κρίθηκε το σημασιολογικό μοντέλο σχέσεων/οντοτήτων (SEMANTIC MODELLING-ENTITY/RELATIONSHIP). Η καταλληλότητα αυτού του μοντέλου γίνεται φανερή όταν συγκρίνουμε τις έννοιες του μοντέλου (αριστερή στήλη) με τις έννοιες του ΗΛ (δεξιά στήλη) από όπου προκύπτει η μονοσήμαντη αντιστοιχία.

Οι έννοιες της ηλεκτρονικής βάσης δεδομένων και οι σχέσεις τους οργανώνονται σύμφωνα με το μοντέλο E/R. Στην πάνω δεξιά γωνία ταξινομούνται τα λήμματα σε 5 κατηγορίες μορφημάτων και στα θέματα. Στην πάνω αριστερή πλευρά βρίσκεται η σχεδίαση του λήμματος (ΛΗΜΜΑΤΑ) με τις ιδιότητές του (ΤΑΥΤΟΤΗΤΑ, ΜΟΡΦΗ, Μ/Σ: μορφοσυντακτικές ιδιότητες), εκ των οποίων η ΤΑΥΤΟΤΗΤΑ αποτελεί κλειδί (είναι υπογραμμισμένη), καθώς και τις σχέσεις του με άλλα λήμματα (Μ/Λ, Μ/Π). Στο κάτω μέρος παρουσιάζονται λεπτομερέστερα οι διάφορες σχέσεις που έχουν τα λήμματα μεταξύ τους.



Εικόνα 1 Μοντέλο E/R, αναπαράσταση των σχέσεων και των ιδιοτήτων των λημμάτων (Παπακίτσος 2000)

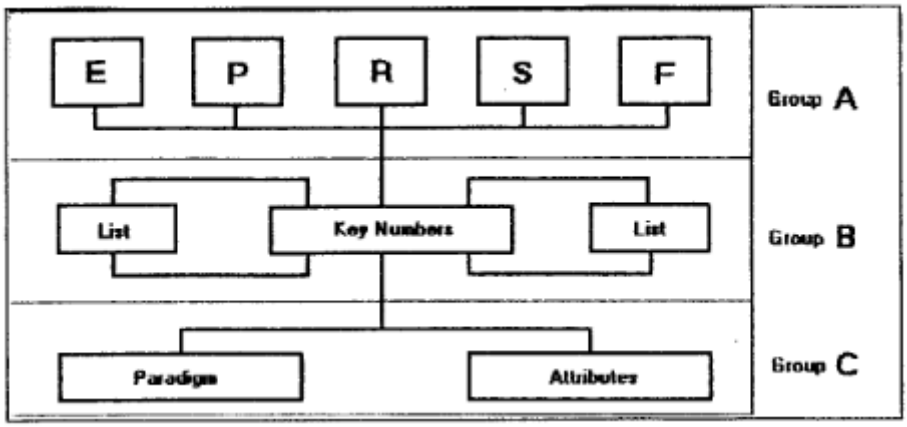
2.5 Οργάνωση της Ηλεκτρονικής Βάσης δεδομένων

Η οργάνωση της βάσης δεδομένων έγινε με τέτοιο τρόπο ώστε να παρέχεται η δυνατότητα ελέγχου των παραγόμενων αναλύσεων, αλλά ταυτόχρονα να είναι ευέλικτος και οικονομικά αποδοτικός.

Τα δεδομένα του λεξικού κατηγοριοποιήθηκαν σε 3 ομάδες (Papakitsos 1997 στο Παπακίτσος 2000) όπως προκύπτει από την Εικόνα 2.

Αυτός ο τρόπος θα πρέπει να είναι αποτελεσματικός για τις εξαιρέσεις ή τις πληροφορίες που κωδικοποιούνται δύσκολα υπό τη μορφή ιδιοτήτων (ιδιότητες για το γένος, τον αριθμό ή την πτώση μιας λέξης).

Η πρώτη ομάδα αρχείων περιλαμβάνει τις συμβολοσειρές των μορφημάτων και μπορεί να αποθηκεύονται σε πέντε ξεχωριστά αρχεία ή σε ένα αρχείο, ανάλογα με τις χρησιμοποιούμενες δομές δεδομένων (Παπακίτσος 2000).



Εικόνα 2 Οι 3 ομάδες των δεδομένων του Λεξικού (Παπακίτσος, 2000)

Τα μορφήματα συνοδεύονται με δείκτες (Key Numbers) που αναφέρονται στις ιδιότητές τους, όπως το Paradigm (παράδειγμα κλίσης) και τα Attributes (χαρακτηριστικά). Επίσης, υπάρχουν λίστες (List) που περιλαμβάνουν τα προθήματα που προηγούνται και τα επιθήματα που ακολουθούν το συγκεκριμένο μόρφημα. Αυτές οι λίστες μπορεί να περιέχουν και άλλα κλειδιά που αναδεικνύουν τα προσφύματα που συνοδεύουν το συγκεκριμένο λήμμα, ή δείκτες που κωδικοποιούν διάφορες σχέσεις. Η δεύτερη ομάδα περιέχει κλειδιά και δείκτες ενώ η τρίτη περιέχει τα χαρακτηριστικά.

Κάθε είσοδος του λεξικού αντιστοιχίζεται με έναν μονοσήμαντο κωδικό αριθμό, που του επιτρέπει να έχει εύκολη μελλοντική επέκταση των ιδιοτήτων του λήμματος και να εξυπηρετεί καλύτερα τόσο την ανάλυση όσο και τη σύνθεση του κειμένου. Αυτός ο κωδικός αριθμός βοηθά επίσης στην αποτελεσματική διαχείριση των πληροφοριών γενικότερα.

Ο μονοσήμαντος κωδικός αριθμός είναι μοναδικός για κάθε λήμμα και του επιτρέπει να αποθηκευτεί και να ανακτηθεί εύκολα από τη βάση δεδομένων. Αυτός ο κωδικός αριθμός μπορεί να χρησιμοποιηθεί ως αναγνωριστικό για το συγκεκριμένο λήμμα και να συνοδεύεται από διάφορες ιδιότητες και πληροφορίες που αφορούν το λήμμα.

Ο μονοσήμαντος κωδικός αριθμός βοηθάει στην επέκταση των ιδιοτήτων του λήμματος, καθώς μπορεί να προστεθούν νέες πληροφορίες και γνωρίσματα στον κωδικό χωρίς να επηρεάζεται η αναζήτηση και η επεξεργασία του.

Ως ένα ενιαίο σύνολο, οι πληροφορίες που καταχωρούνται για κάθε λήμμα περιλαμβάνουν τις ονομασίες υλοποίησης κι έχουν την ακόλουθη δομή (Εικόνα 3):

- *Μέγιστο Μέγεθος* (Maximum): Είναι η μέγιστη κατηγορία στην οποία ανήκει το λήμμα. Αυτή η πληροφορία βρίσκεται στην πάνω αριστερή γωνία του λήμματος.
- *Μορφή* (Form): Αναφέρεται στη συμβολοσειρά που αντιπροσωπεύει το λήμμα.
- *Κατηγορία* (Class): Αναφέρεται στην κατηγορία του λήμματος, για παράδειγμα, ουσιαστικό, ρήμα, επίθετο.
- *Δείκτης Next*: Είναι ένας δείκτης που οδηγεί στο επόμενο λήμμα με την ίδια μορφή. Αυτό είναι χρήσιμο κυρίως για την αντιμετώπιση των μεταπλαστικών νόμων, δηλαδή των κανόνων που επιτρέπουν τη μεταβολή της μορφής ενός λήμματος.
- *Κλειδί* (Code): Είναι ένα κλειδί που οδηγεί σε υπόλοιπες ιδιότητες του λήμματος. Αυτές οι ιδιότητες περιλαμβάνουν πληροφορίες σχετικά με τη σημασιολογία, τις κλίσεις, τις παραλλαγές, τις συνώνυμες λέξεις.

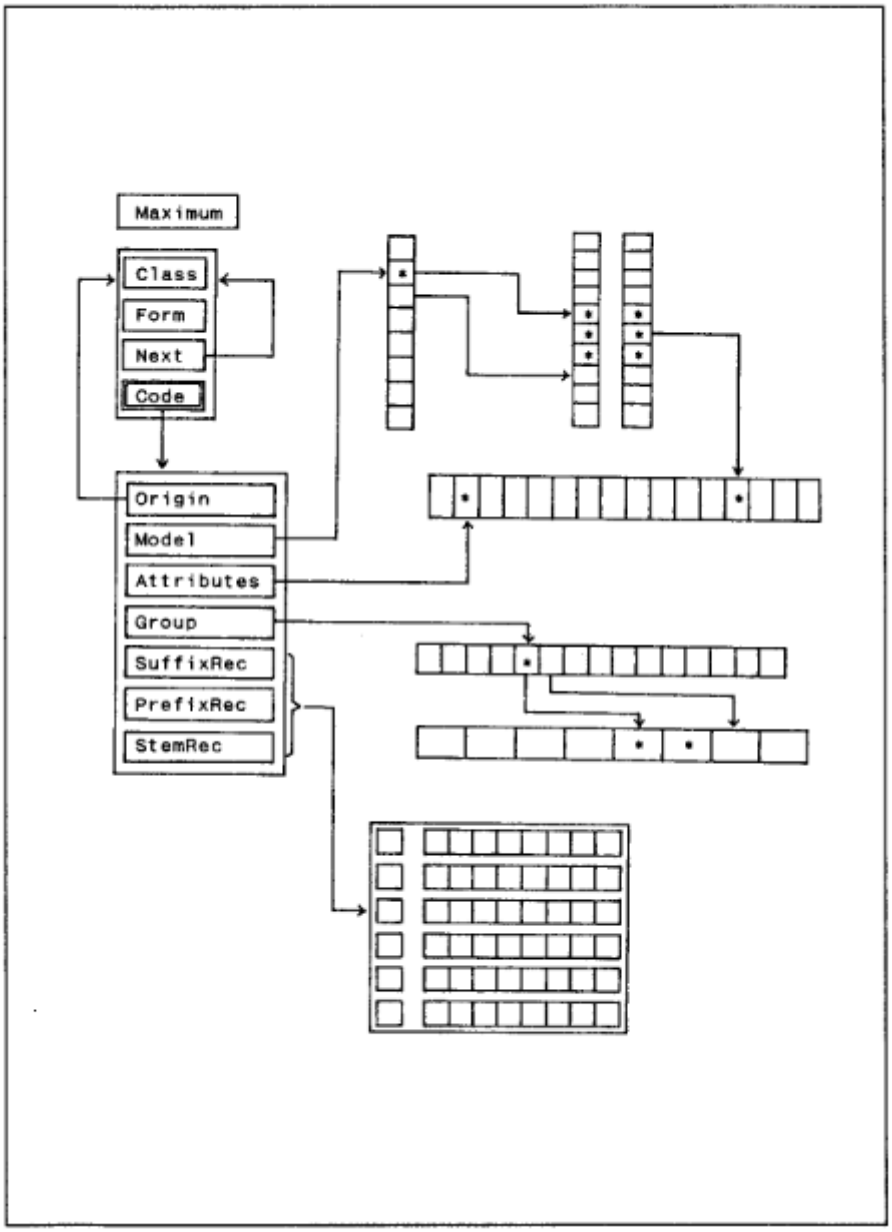
Οι ιδιότητες που αναφέρονται κάτω από το Κλειδί – Code είναι οι εξής:

- *Προέλευση* (Origin): Αυτή η ιδιότητα οδηγεί πίσω στη θέση της συμβολοσειράς, για να διευκολύνει τη γεννήτρια κειμένου να ανακτήσει την αρχική μορφή του λήμματος.
- *Μοντέλο* (Model): Αυτή η ιδιότητα είναι ένας δείκτης προς το κλιτικό υπόδειγμα του λήμματος, που βρίσκεται σε έναν πίνακα θέσης (άνω δεξιά).
- *Ιδιότητες* (Attributes): Αυτή η ιδιότητα οδηγεί σε μορφοσυντακτικές ιδιότητες του λήμματος.
- *Ομάδα* (Group): Αυτή η ιδιότητα οδηγεί σε έναν πίνακα με δείκτες προς τα αλλόμορφα του λήμματος, σχετικά με τους μορφολεξικούς νόμους.

Επιπλέον, υπάρχει μια ομάδα τριών δεικτών:

SuffixRec, PrefixRec και *StemRec*: Αυτοί οι δείκτες οδηγούν σε διάφορους πίνακες ίδιας δομής και υλοποιούν την παραγωγή και τη σύνθεση του λήμματος με άλλα

επιθήματα, προθήματα ή ρίζες/θέματα, για τη δημιουργία άλλων θεμάτων. Οι εν λόγω δείκτες δηλώνουν την συγκεκριμένη θέση του πίνακα δεδομένων που οδηγούν, εξασφαλίζοντας ταχύτατη πρόσβαση.



Εικόνα 3 Οργάνωση των δεδομένων του λεξικού (Παπακίτσος 2000)

Όσον αφορά τη φυσική οργάνωση, το λεξικό έχει δύο μορφές: Στατική και Δυναμική. Το στατικό λεξικό αποτελεί ένα είδος αποθετηρίου εφόσον είναι μόνιμα αποθηκευμένο στον σκληρό δίσκο, ενώ το δυναμικό αντιστοιχεί στο ενεργό μέρος το λεξικού, το οποίο φορτώνεται προσωρινά στην Κύρια Μνήμη (Main Memory) του υπολογιστή για να αυξηθεί η ταχύτητα επεξεργασίας.

Στο στατικό λεξικό πραγματοποιείται η διαχείριση από το Σύστημα Διαχείρισης Βάσης Δεδομένων, η οποία περιλαμβάνει εισαγωγές, διαγραφές και μεταβολές στα δεδομένα του λεξικού. Αυτό συμβαίνει επειδή η λειτουργία του λεξικού στη δυναμική του μορφή είναι πολύπλοκη και προσαρμοσμένη στην επεξεργασία των δεδομένων.

Το στατικό λεξικό περιέχει τα παρακάτω 5 αρχεία τύπου κειμένου, από τα οποία το καθένα αντιστοιχεί σε ξεχωριστή κατηγορία μορφήματος, όπου το κάθε μόρφημα βρίσκεται καταχωρημένο με τις ιδιότητες του:

- Prefix.txt,
- roots.txt,
- Endings.txt,
- Suffix.txt
- Words.txt,

Στο Endings.txt καταχωρούνται οι καταλήξεις, ανάλογα με τα κλιτικά υποδείγματα.

Για να μετατραπεί το στατικό λεξικό σε δυναμικό, χρησιμοποιείται το Σύστημα Ημιαυτόματης Διαχείρισης. Το δυναμικό λεξικό αποτελείται από δύο μέρη, από τα οποία το πρώτο είναι το σταθερό δυναμικό λεξικό που περιλαμβάνει μεγάλα αρχεία τύπου Turbo Pascal (GROUPS.TPF, LEMMATA.TPF και GRPPART.TPF), παραμένει αποθηκευμένο στον σκληρό δίσκο, χρησιμοποιώντας τα *TPF κατά τη διάρκεια της επεξεργασίας. Το δεύτερο μέρος είναι το ενεργό δυναμικό λεξικό το οποίο περιλαμβάνει έξι αρχεία τύπου Turbo Pascal (MAXIMA.TPF, MORPHS.TPF, ATTRIBS.TPF, MODELS), τα οποία μεταφέρονται σε πίνακες στην Κύρια Μνήμη μέσω της Μονάδας Loader στη μεταβλητή TLexicon. Αυτά τα αρχεία περιλαμβάνουν τις βασικές ιδιότητες των λημμάτων, (μορφή, κλειδί) επιταχύνοντας την αναγνώριση. Το συγκεκριμένο κομμάτι είναι κρίσιμο για τις επιδόσεις των μορφολογικών επεξεργαστών, καθώς περιέχει τις συμβολοσειρές και τα κλειδιά των λημμάτων που απαιτούνται για την αναζήτηση και την επεξεργασία τους. Επιπλέον, η επιλογή της δομής δεδομένων που θα φιλοξενήσει αυτές τις πληροφορίες είναι μια σημαντική παράμετρος απόδοσης. Η κατάλληλη επιλογή μπορεί να έχει σημαντική επίδραση στην ταχύτητα και την αποδοτικότητα των μορφολογικών επεξεργαστών, καθώς επιτρέπει γρήγορη πρόσβαση κι επεξεργασία των δεδομένων που απαιτούνται για την ανάλυση των λημμάτων (Παπακίτσος 2000).

2.6 Οι δομές των δεδομένων του Ηλεκτρονικού Λεξικού

Η τρέχουσα ηλεκτρονική βάση του λεξικού περιέχει μορφήματα και μπορεί να περιλαμβάνει από 8.000 έως 30.000 λήμματα, τα οποία επαρκούν για την κάλυψη τουλάχιστον του αριθμού λέξεων που υπάρχουν και στα απλά ηλεκτρονικά λεξικά. Το λεξικό μορφημάτων είναι πιο πολύπλοκο από ένα απλό ηλεκτρονικό λεξικό, ενώ παρέχει μεγαλύτερη ευελιξία, καθώς μπορεί ευκολότερα να αντιμετωπίσει νεολογισμούς. Ωστόσο υφίσταται η πρόκληση να βρεθούν αποτελεσματικότερες δομές δεδομένων για την αποθήκευση των συμβολοσειρών.

Η αποθήκευση των συμβολοσειρών, των κλειδιών τους, του τύπου του μορφήματος, των τιμών των ιδιοτήτων και των κλιτικών υποδειγμάτων γίνεται στην Κύρια Μνήμη κι έχουν δοκιμαστεί 13 διαφορετικές δομές δεδομένων.

Οι δομές των δεδομένων μπορεί να είναι αρχεία κειμένου ASCII (TXT), αρχεία τύπου με πίνακες (A??), δυναμικές λίστες (L??) και δυαδικά δέντρα (BTS) με σειριακή (?SS), δυαδική (?BS) ή τριαδική αναζήτηση (?TS). Σε πίνακες με λιγότερα από 100 στοιχεία, η σειριακή και η δυαδική αναζήτηση έχουν την ίδια μέση ταχύτητα ανεύρεσης, ενώ μία δυναμική λίστα είναι 2 με 4 φορές πιο αργή από έναν πίνακα. Η θέση κάθε στοιχείου της δομής περιείχε τη συμβολοσειρά και το κλειδί της, και οι δομές είχαν 100 και 200 θέσεις. Δοκιμάζοντας την απόδοση των διαφορετικών δομών, αναδείχθηκε ως ταχύτερη μέθοδος η δυαδική αναζήτηση πίνακα (ABS) με εξαίρεση τη μέθοδο κατακερματισμού (Hashing-HAS), ωστόσο επειδή ο αριθμός των στοιχείων κάθε δομής δεν ξεπέρασε τα 20, προτιμήθηκε η σειριακή αναζήτηση με κριτήριο την απλότητά της.

Για να επιτευχθεί καλύτερη ταχύτητα και διαχείριση μνήμης δοκιμάστηκαν και χρησιμοποιήθηκαν υβριδικές δομές, διότι ο μονοδιάστατος πίνακας προσφέρει ταχύτητα αναζήτησης απαιτώντας πολλή μνήμη. Αντίθετα η δυναμική λίστα επιτυγχάνει καλύτερη διαχείριση μνήμης με χαμηλή ταχύτητα. Μία υβριδική δομή των δύο, με δυναμικές λίστες μονοδιάστατων πινάκων προσφέρει καλύτερη επίδοση ταχύτητας-μνήμης. Επίσης σημαντικό θέμα διαχείρισης αποτελούν και οι ρουτίνες διαχείρισης των εν λόγω δομών. Η πληροφορία του ηλεκτρονικού λεξικού κωδικοποιείται σε μία σύνθετη εγγραφή της μορφής λήμμα-κλειδί-τύπος-ιδιότητες για την οποία απαιτείται μονοδιάστατος πίνακας εγγραφών. Όταν μεταβάλλεται το πεδίο εγγραφής, μεταβάλλεται και η ρουτίνα διαχείρισης.

Για να αντιμετωπιστούν οι διαφορετικοί τύποι πεδίων και να οργανωθεί ο όγκος των δεδομένων, προτιμήθηκε η χρήση μονοδιάστατων πινάκων, δηλαδή κάθε πεδίο αντιστοιχεί σε ένα στοιχείο του μονοδιάστατου πίνακα. Για παράδειγμα, αν έχουμε ένα πεδίο τύπου ψηφιολέξης (byte), θα υπάρχει ένα στοιχείο στον μονοδιάστατο πίνακα που αντιπροσωπεύει αυτό το πεδίο. Αντίστοιχα, για ένα πεδίο αριθμητικού τύπου (word), θα υπάρχουν δύο συνεχόμενα στοιχεία στον μονοδιάστατο πίνακα που αντιπροσωπεύουν αυτό το πεδίο. Οι ρουτίνες αναζήτησης περιορίστηκαν σε δύο, μία για τύπο ψηφιολέξης (byte) και μία για αριθμητικό τύπο (word).

Για την αποθήκευση των συμβολοσειρών, του τύπου και των κλειδιών τους, δοκιμάστηκαν επιπλέον οι παρακάτω δομές:

Πίνακες κατακερματισμού (hash-tables): Χρησιμοποιήθηκαν πίνακες κατακερματισμού για την αποθήκευση των συμβολοσειρών, με το κλειδί να χρησιμοποιείται για την αναζήτηση. Οι πίνακες κατακερματισμού είναι μια αποδοτική δομή για τη γρήγορη αναζήτηση σε μεγάλα σύνολα δεδομένων.

Δενδρικά λεξικά (root lexica ή tries): Χρησιμοποιήθηκαν δενδρικά λεξικά για την αποθήκευση των συμβολοσειρών. Οι δενδρικές δομές είναι κατάλληλες για την αναζήτηση σε συμβολοσειρές με βάση το πρόθημα.

Πινακοειδή λεξικά: Έγινε δοκιμή πινακοειδούς λεξικού για την αποθήκευση των συμβολοσειρών. Αυτή η δομή συνδυάζει τα πλεονεκτήματα ενός πίνακα και ενός δένδρου για γρήγορη αναζήτηση (Παπακίτσος 2000).

Κεφάλαιο 3 Βελτίωση της προηγούμενης λειτουργίας

Σε αυτό το κεφάλαιο παρουσιάζεται η μετατροπή της παλιάς βάσης των δεδομένων του ηλεκτρονικού λεξικού, το οποίο από ένα σύνολο αρχείων κειμένου μετατρέπεται σε ένα διαχειρίσιμο αρχείο Excel. Με τη βοήθεια της γλώσσας προγραμματισμού python και των βιβλιοθηκών της Unicodedata και pandas επιτυγχάνεται η εν λόγω μετατροπή, η οποία προσφέρει μια εντυπωσιακή προοπτική για την αποτελεσματική διαχείριση και ανάλυση των δεδομένων, κατά συνέπεια μεγάλη αποδοτικότητα στην επεξεργασία φυσικής γλώσσας.

3.1 Η γλώσσα προγραμματισμού python

Η Python είναι μια υψηλού επιπέδου γλώσσα προγραμματισμού που δημιουργήθηκε από τον Guido van Rossum το 1991 και είναι πολύ δημοφιλής λόγω της απλότητας και της ευαναγνωσιμότητάς της.

Η Python χρησιμοποιεί ευέλικτη σύνταξη και υποστηρίζει πολλά παραδείγματα προγραμματισμού, όπως διαδικαστικό, αντικειμενοστρεφή και λειτουργικό προγραμματισμό. Είναι μια γλώσσα προγραμματισμού γενικής χρήσης και μπορεί να χρησιμοποιηθεί για την ανάπτυξη διάφορων εφαρμογών, από γραμμές εντολών έως ιστοσελίδες κι εφαρμογές επιχειρησιακού λογισμικού (Van Rossum 2007).

Μια από τις κύριες δυνατότητες της Python είναι η ύπαρξη μιας εκτεταμένης βιβλιοθήκης που περιλαμβάνει πολλά εργαλεία και πακέτα, καλύπτοντας διάφορους τομείς όπως επιστημονική υπολογιστική, τεχνητή νοημοσύνη, ανάλυση δεδομένων, ανάπτυξη ιστού. Λόγω των εξειδικευμένων βιβλιοθηκών χρησιμοποιείται σε πολλά περιβάλλοντα και μπορεί να προσαρμοστεί σε κάθε τύπο χρήσης. Χρησιμοποιείται ιδιαίτερα ως γλώσσα σεναρίου για την αυτοματοποίηση απλών αλλά κουραστικών εργασιών.

Ο αριθμός των τυπικών λειτουργικών μονάδων βιβλιοθήκης μπορεί να αυξηθεί με συγκεκριμένες ενότητες γραμμένες σε C ή Python. Η τυπική βιβλιοθήκη είναι ιδιαίτερα καλά σχεδιασμένη για τη σύνταξη εφαρμογών χρησιμοποιώντας το Διαδίκτυο, με μεγάλο αριθμό υποστηριζόμενων τυπικών μορφών και πρωτοκόλλων (όπως MIME και HTTP). Παρέχονται επίσης ενότητες για τη δημιουργία GUI και τον χειρισμό κανονικών εκφράσεων. Η Python περιλαμβάνει επίσης ένα πλαίσιο δοκιμών μονάδων

(unittest, πρώην PyUnit πριν από την έκδοση 2.1) για τη δημιουργία περιεκτικών σειρών δοκιμών.

Αν και κάθε προγραμματιστής μπορεί να υιοθετήσει τις δικές του συμβάσεις για τη σύνταξη κώδικα Python, ο Guido van Rossum έχει διαθέσει έναν οδηγό, που αναφέρεται ως "PEP 8". Δημοσιεύθηκε το 2001, διατηρείται ακόμα για να προσαρμοστεί στις αλλαγές της γλώσσας. Η Google προσφέρει επίσης έναν οδηγό.

Η Python έχει αρκετές διαθέσιμες ενότητες για την κατασκευή λογισμικού με GUI. Το πιο διαδεδομένο είναι το Tkinter. Αυτή η ενότητα είναι κατάλληλη για πολλές εφαρμογές και μπορεί να θεωρηθεί επαρκής στις περισσότερες περιπτώσεις. Ωστόσο, έχουν δημιουργηθεί και άλλα modules για να είναι δυνατή η σύνδεση της Python με άλλες βιβλιοθήκες λογισμικού («εργαλειοθήκη»), για περισσότερες λειτουργίες, για καλύτερη ενσωμάτωση με το λειτουργικό σύστημα που χρησιμοποιείται ή απλώς για τη δυνατότητα χρήσης της Python με την αγαπημένη της βιβλιοθήκη. Πράγματι, ορισμένοι προγραμματιστές βρίσκουν τη χρήση του Tkinter πιο κουραστική από άλλες βιβλιοθήκες. Αυτές οι άλλες ενότητες δεν αποτελούν μέρος της τυπικής βιβλιοθήκης κι επομένως πρέπει να ληφθούν χωριστά.

Οι κύριες λειτουργικές μονάδες που παρέχουν πρόσβαση σε βιβλιοθήκες GUI είναι οι Tkinter και Pmw (Python megawidgets)50 για Tk, wxPython για wxWidgets, PyGTK για GTK, PyQt και PySide για Qt και τέλος FxPy για το FOX Toolkit. Υπάρχει επίσης μια προσαρμογή της βιβλιοθήκης SDL: Pygame, μια σύνδεση του SFML: PySFML, καθώς και μια βιβλιοθήκη γραμμένη ειδικά για την Python: Pyglet (en).

Τέλος διαθέτει βιβλιοθήκες όπως η NLTK (Natural Language Toolkit) που μπορούν να εκτελέσουν εργασίες επεξεργασίας φυσικής γλώσσας. Οι συγκεκριμένες βιβλιοθήκες μπορούν να χρησιμοποιηθούν για να εκτελέσουν λειτουργίες όπως λημματοποίηση, αναζήτηση συνωνύμων ή μορφολογική ανάλυση, για να βελτιωθεί η λειτουργικότητα εύρεσης και ανάκτησης λέξεων στο λεξικό.

Οι εφαρμογές λογισμικού και οι γλώσσες προγραμματισμού, συμπεριλαμβανομένης της Python, συχνά ενσωματώνουν τη βιβλιοθήκη δεδομένων Unicode ή ένα υποσύνολο αυτής για την υποστήριξη του κατάλληλου χειρισμού, επεξεργασίας και απόδοσης χαρακτήρων Unicode. Η βιβλιοθήκη επιτρέπει στους προγραμματιστές να έχουν πρόσβαση σε λεπτομερείς πληροφορίες σχετικά με χαρακτήρες, να εκτελούν επικυρώσεις χαρακτήρων, να εφαρμόζουν αλγόριθμους ταξινόμησης και αναζήτησης

και να διασφαλίζουν ακριβή αναπαράσταση κειμένου σε διαφορετικά συστήματα γραφής.

3.2 Η βιβλιοθήκη Unicode data

Το Unicode Standard αποτελεί πρότυπο τεχνολογίας πληροφοριών για τη συνεπή κωδικοποίηση, αναπαράσταση και χειρισμό κειμένου που εκφράζεται στα περισσότερα από τα συστήματα γραφής του κόσμου. Το πρότυπο, το οποίο διατηρείται από την Unicode Consortium, ορίζει από την τρέχουσα έκδοση (15.0) 149.186 χαρακτήρες που καλύπτουν 161 σύγχρονα και ιστορικά συστήματα γραφής, καθώς και σύμβολα, χιλιάδες emoji (συμπεριλαμβανομένων των χρωμάτων), και μη οπτικούς κωδικούς ελέγχου και μορφοποίησης (Garfinkel 2012).

Η επιτυχία της Unicode στην ενοποίηση συνόλων χαρακτήρων οδήγησε στην ευρεία και κυρίαρχη χρήση της στη διεθνοποίηση και τον εντοπισμό του λογισμικού υπολογιστών. Το πρότυπο έχει εφαρμοστεί σε πολλές πρόσφατες τεχνολογίες, συμπεριλαμβανομένων των σύγχρονων λειτουργικών συστημάτων, XML, JSON και των πιο σύγχρονων γλωσσών προγραμματισμού, μερικές φορές μόνο σε μορφή UTF-8.

Το Unicode είναι ένα πρότυπο κωδικοποίησης χαρακτήρων που είναι συγχρονισμένο με το ISO/IEC 10646 και περιλαμβάνει ένα ευρύ φάσμα χαρακτήρων από διάφορες γλώσσες και συμβόλων. Το Unicode δεν περιορίζεται μόνο στην κωδικοποίηση των χαρακτήρων, αλλά παρέχει επίσης λεπτομερείς πληροφορίες σχετικά με τη σύνθεση, την αποσύνθεση, την απόδοση και την εμφάνιση των χαρακτήρων. Περιλαμβάνει κανόνες για τη διαχείριση του διπλού κατευθυντικού κειμένου και προσφέρει αρχεία αναφοράς και γραφήματα για την υποστήριξη των προγραμματιστών και σχεδιαστών στη σωστή χρήση του χαρακτηριστικού συνόλου χαρακτήρων.

Το Unicode μπορεί να αποθηκευτεί χρησιμοποιώντας διάφορες κωδικοποιήσεις που μετατρέπουν τους χαρακτήρες σε διακριτές ακολουθίες byte. Το Unicode περιλαμβάνει διάφορες κωδικοποιήσεις, αλλά οι πιο κοινές είναι οι UTF-8, UTF-16 και UTF-32. Η UTF-8 είναι μια πολύ ευέλικτη κωδικοποίηση που μπορεί να αναπαραστήσει οποιονδήποτε χαρακτήρα του Unicode και είναι συμβατή με την ASCII κωδικοποίηση. Η UTF-16 είναι μια διπλάσια κωδικοποίηση που χρησιμοποιεί 16-bit για την αναπαράσταση χαρακτήρων, ενώ η UTF-32 χρησιμοποιεί 32-bit για κάθε χαρακτήρα.

Η βιβλιοθήκη "unicodedata" είναι μια βιβλιοθήκη της γλώσσας προγραμματισμού Python που παρέχει λειτουργίες για την ανάκτηση πληροφοριών σχετικά με τους Unicode χαρακτήρες. Όπως παρουσιάστηκε στα προηγούμενα, η Unicode είναι μια πρότυπη μέθοδος κωδικοποίησης χαρακτήρων που περιλαμβάνει ένα ευρύ φάσμα γλωσσών και συμβόλων από διάφορες γλωσσικές ομάδες.

Από τη βιβλιοθήκη "unicodedata" παρέχονται συναρτήσεις για την πληροφόρηση σχετικά με τα διάφορα χαρακτηριστικά των Unicode χαρακτήρων, όπως η κατηγορία χαρακτήρα, η κατάσταση κεφαλαίου-πεζού, οικογένεια γραμματοσειράς, αριθμητικές τιμές και άλλα, πληροφορίες χρήσιμες για τις επεξεργασίες των Unicode χαρακτήρων, δηλαδή την ανάγνωση, την επεξεργασία και την απεικόνιση κειμένου.

Περιέχει δηλαδή εκτενείς πληροφορίες για τους μεμονωμένους χαρακτήρες Unicode, συμπεριλαμβανομένων των ιδιοτήτων, των αντιστοιχίσεων και των σχέσεών τους.

Μερικές από τις κύριες συναρτήσεις που παρέχονται από τη βιβλιοθήκη "unicodedata" περιλαμβάνουν τις εξής:

unicodedata.category(): Επιστρέφει την κατηγορία ενός Unicode χαρακτήρα.

unicodedata.name(): Επιστρέφει το όνομα ενός Unicode χαρακτήρα.

Η βιβλιοθήκη δεδομένων Unicode διανέμεται συχνά με τη μορφή αρχείων Unicode Character Database (UCD), όπως UnicodeData.txt και DerivedCoreProperties.txt. Αυτά τα αρχεία παρέχουν ολοκληρωμένα δεδομένα για κάθε χαρακτήρα Unicode, συμπεριλαμβανομένου του σημείου κώδικα, του ονόματος, της κατηγορίας, του συστήματος γραφής και των διαφόρων ιδιοτήτων χαρακτήρων (Tauber 2019).

3.3 Η βιβλιοθήκη Tkinter

Η Tkinter είναι η αρχική δωρεάν γραφική βιβλιοθήκη για τη γλώσσα Python, που επιτρέπει τη δημιουργία γραφικών διεπαφών. Προέρχεται από μια προσαρμογή της βιβλιοθήκης γραφικών Tk που γράφτηκε για την Tcl.

Η Tkinter είναι μια βιβλιοθήκη ανοιχτού κώδικα, φορητή γραφική διεπαφή χρήστη (GUI) που έχει σχεδιαστεί για χρήση σε δέσμες ενεργειών Python. Η Tkinter βασίζεται στη βιβλιοθήκη Tk, τη βιβλιοθήκη γραφικών που χρησιμοποιείται από τους Tcl/Tk και Perl, η οποία με τη σειρά της υλοποιείται στη C. Επομένως, μπορεί να ειπωθεί ότι η Tkinter υλοποιείται χρησιμοποιώντας πολλαπλά επίπεδα.

Στα πλεονεκτήματα της συγκεκριμένης βιβλιοθήκης συγκαταλέγονται η πολυεπίπεδη προσέγγιση, η προσβασιμότητα, η φορητότητα, η διαθεσιμότητα.

Η πολυεπίπεδη προσέγγιση που χρησιμοποιείται στον σχεδιασμό της Tkinter, της δίνει όλα τα πλεονεκτήματα της βιβλιοθήκης TK. Ως εκ τούτου, τη στιγμή της δημιουργίας, η Tkinter κληρονόμησε τα πλεονεκτήματα μιας εργαλειοθήκης γραφικών που είχε χρόνο να ωριμάσει. Έτσι οι πρώιμες εκδόσεις της Tkinter έγιναν πολύ πιο σταθερές και αξιόπιστες απ' ότι αν είχαν ξαναγραφτεί από την αρχή. Επιπλέον, η μετατροπή από Tcl/Tk σε Tkinter ήταν πραγματικά μηδαμινή, επομένως οι προγραμματιστές Tk μπορούσαν να μάθουν να χρησιμοποιούν την Tkinter πολύ εύκολα.

Σχετικά με την προσβασιμότητα, η εκμάθηση της Tkinter είναι σχετικά διαισθητική, γρήγορη και ανώδυνη. Η υλοποίηση της Tkinter κρύβει λεπτομερείς και περίπλοκες κλήσεις σε απλές κι εύχρηστες μεθόδους. Αυτό είναι μια συνέχεια του τρόπου σκέψης της Python, καθώς η γλώσσα υπερέχει στην ταχεία κατασκευή πρωτοτύπων.

Όσον αφορά τη φορητότητα, τα σενάρια Python που χρησιμοποιούν Tkinter δεν απαιτούν τροποποιήσεις για τη μεταφορά από τη μια πλατφόρμα στην άλλη. Η Tkinter είναι διαθέσιμη για όλες τις πλατφόρμες για τις οποίες υλοποιείται η Python, δηλαδή τα Microsoft Windows, X Windows και Macintosh. Αυτό της δίνει ένα μεγάλο πλεονέκτημα σε σχέση με τις περισσότερες ανταγωνιστικές βιβλιοθήκες, οι οποίες συχνά περιορίζονται σε μία ή δύο πλατφόρμες. Επιπλέον, η Tkinter παρέχει την εγγενή εμφάνιση και αίσθηση της συγκεκριμένης πλατφόρμας στην οποία εκτελείται.

Η Tkinter περιλαμβάνεται πλέον σε όλες τις διανομές Python, επομένως, δεν απαιτούνται πρόσθετες μονάδες για την εκτέλεση σεναρίων που χρησιμοποιούν την Tkinter.

Τέλος, η πολυεπίπεδη προσέγγιση που υιοθετείται στον σχεδιασμό της Tkinter μπορεί να έχει ορισμένα μειονεκτήματα όσον αφορά την ταχύτητα εκτέλεσης. Το θέμα της ταχύτητας μπορεί να είναι πρόβλημα με τα παλαιότερα και πιο αργά μηχανήματα, εφόσον οι περισσότεροι σύγχρονοι υπολογιστές είναι αρκετά γρήγοροι για να αντιμετωπίσουν την επιπλέον επεξεργασία σε εύλογο χρονικό διάστημα. Ωστόσο, όταν η ταχύτητα είναι κρίσιμης σημασίας, πρέπει να λαμβάνεται μέριμνα για την εγγραφή κώδικα που είναι όσο το δυνατόν πιο αποτελεσματικός.

3.4 Η βιβλιοθήκη docx

Η python-docx είναι μια βιβλιοθήκη Python για τη δημιουργία και την ενημέρωση αρχείων Microsoft Word (.docx). Χρησιμοποιείται για τη δημιουργία και την επεξεργασία των εγγράφων Word (αρχεία .docx). Μέσω αυτής της βιβλιοθήκης, μπορεί να δημιουργηθούν, να τροποποιηθούν ή να γίνουν επεξεργάσιμα αρχεία Word με ποικίλες λειτουργίες, όπως προσθήκη κειμένου, εικόνων, πινάκων, στυλ, σχόλια και άλλα.

Η συγκεκριμένη βιβλιοθήκη αποτελεί μια ευέλικτη στη χρήση διεπαφή προγραμματισμού που επιτρέπει τη δημιουργία / επεξεργασία εγγράφων Word μέσω του κώδικα Python. Με τη βοήθειά της, αυτοματοποιούνται οι διάφορες λειτουργίες της δημιουργίας των εγγράφων Word, παράγοντας αναφορές, εκθέσεις ή εγχειρίδια χρήσης για ποικίλες εφαρμογές.

Για να χρησιμοποιηθεί η βιβλιοθήκη python-docx, πρέπει να εγκατασταθεί μέσω του pip το εργαλείο διαχείρισης πακέτων της Python. Έπειτα από την εγκατάσταση, μπορεί να εισαχθεί η βιβλιοθήκη στον κώδικα και να ξεκινήσει η επεξεργασία των εγγράφων Word.

3.5 Η βιβλιοθήκη os

Η βιβλιοθήκη os αποτελεί ενσωματωμένη βιβλιοθήκη της Python και παρέχει λειτουργίες για τη διαχείριση του λειτουργικού συστήματος, όπως το σύστημα αρχείων και οι επικοινωνίες με το περιβάλλον του υπολογιστή. Διαθέτει πολλές χρήσιμες συναρτήσεις για τη διαχείριση διαδρομών αρχείων, τη δημιουργία και τη διαγραφή αρχείων και φακέλων, τον έλεγχο των δικαιωμάτων πρόσβασης αρχείων, την εκτέλεση εντολών συστήματος.

Αποτελεί ένα πολύ χρήσιμο εργαλείο για την εκτέλεση λειτουργιών με το λειτουργικό σύστημα. Παρέχει ένα πλήρες φάσμα λειτουργιών για εργασία με αρχεία, καταλόγους, διεργασίες, χρήστες, διαδρομές αρχείων και καταλόγου και πληροφορίες λειτουργικού συστήματος.

Με τη βοήθεια της βιβλιοθήκης os, είναι δυνατόν να προγραμματιστούν αλληλεπιδράσεις με το σύστημα αρχείων, να γίνουν διαχειρίσιμες διαδρομές και

ονόματα αρχείων, να ελεγχθούν αρχεία ή φάκελοι, άδειες πρόσβασης αρχείων ή να εκτελεστούν οι εντολές συστήματος.

3.6 Η βιβλιοθήκη openpyxl

Η βιβλιοθήκη openpyxl είναι ανοικτού κώδικα βιβλιοθήκη με την οποία γίνονται επεξεργάσιμα αρχεία Excel στη γλώσσα προγραμματισμού Python. Με τη βοήθεια της openpyxl, είναι δυνατό να αναγνωσθούν, να γραφούν και να γίνουν επεξεργάσιμα αρχεία Excel στη μορφή .xlsx.

Η εν λόγω βιβλιοθήκη παρέχει πολλές δυνατότητες για την εργασία με αρχεία Excel, όπως για παράδειγμα την προσθήκη και την επεξεργασία φύλλων εργασίας, τη δημιουργία και τη μετατροπή κελιών, την ανάγνωση και την εγγραφή δεδομένων, τη διαμόρφωση της μορφής κελιών, ή την προσθήκη γραφικών και διαγραμμάτων.

3.7 Η νέα δομή των δεδομένων σε υπολογιστικά φύλλα Excel

Χρησιμοποιώντας την ευέλικτη δυνατότητα της Python, μπορεί να σχεδιαστεί εκ νέου και να εφαρμοστεί η νέα δομή των αρχείων υλοποίησης σε υπολογιστικά φύλλα Excel, ώστε να είναι η λειτουργία αναζήτησης και αλληλεπίδρασης πιο εύχρηστη.

Συγκεκριμένα, τα δεδομένα του ηλεκτρονικού λεξικού βρίσκονται σε πρωτογενή μορφή της οποίας η κωδικοποίηση δεν είναι πλέον αναγνώσιμη από τα σύγχρονα προγράμματα. Για να γίνει δυνατό να διαβαστούν τα δεδομένα από τα σύγχρονα προγράμματα, πρέπει να αλλάξει η κωδικοποίησή τους σε μια πιο σύγχρονη. Σε αυτή τη διαδικασία μπορεί να χρησιμοποιηθεί η Python, η οποία παρέχει διάφορες βιβλιοθήκες κι εργαλεία για τη διαχείριση και την επεξεργασία διαφορετικών μορφών κωδικοποίησης.

Η κωδικοποίηση των δεδομένων σε μια πιο σύγχρονη μορφή θα γίνει με τη βοήθεια του Excel. Δηλαδή θα μεταφερθούν και θα μορφοποιηθούν τα δεδομένα σε τακτικές δυάδες σε ένα αρχείο υπολογιστικού φύλλου του Excel. Η δημιουργία αυτή του υπολογιστικού φύλλου θα καταστήσει την ανάγνωση των δεδομένων πιο εύκολη για τους χρήστες ενώ θα καταστήσει την πρόσβασή τους πιο εύκολη. Στην ουσία, η δημιουργία της νέας βάσης δεδομένων (Excel) δίνει στον χρήστη τη δυνατότητα να τροποποιεί και να προσθέτει ή και να αφαιρεί με ευκολία και ταχύτητα από αυτήν. Με

αυτόν τον τρόπο ο χρήστης μπορεί να χρησιμοποιήσει τη βάση δεδομένων για οποιαδήποτε επεξεργασία φυσικής γλώσσας.

3.8 Η αποδοτικότητα της νέας βάσης στην επεξεργασία φυσικής γλώσσας

Η αποτελεσματικότητα της νέας βάσης δεδομένων του ηλεκτρονικού λεξικού του Excel στην επεξεργασία φυσικής γλώσσας (NLP) εξαρτάται από διάφορους παράγοντες, όπως τον σχεδιασμό ή την υλοποίηση της βάσης δεδομένων, το μέγεθος την πολυπλοκότητα των δεδομένων και τις εργασίες NLP που εκτελούνται.

Το Excel είναι μια εφαρμογή υπολογιστικών φύλλων και μπορεί να χειριστεί δεδομένα κειμένου και να εκτελέσει βασικές λειτουργίες που σχετίζονται με το κείμενο, κατά συνέπεια είναι αρκετά αποτελεσματικό εργαλείο για προηγμένες εργασίες NLP.

Η εισαγωγή της νέας βάσης δεδομένων στο Excel μπορεί να φέρει επανάσταση στη διαχείριση της προηγουμένως δυσκίνητης βάσης δεδομένων, η οποία αποτελούσε συλλογή αρχείων κειμένου. Αυτός ο μετασχηματισμός φαίνεται να επιφέρει πολλά οφέλη και σημαντικές βελτιώσεις στη χρηστικότητα και την προσβασιμότητα.

Ο μετασχηματισμός της πρώην βάσης δεδομένων που βασιζόταν σε διαφορετικά αρχεία κειμένου σε μια εύκολα προσβάσιμη και διαχειρίσιμη μορφή Excel βελτιώνει σημαντικά τον τρόπο επεξεργασίας και ανάλυσης των δεδομένων. Η φιλική προς τον χρήστη διεπαφή, η εκτεταμένη λειτουργικότητα και η συμβατότητα με άλλα εργαλεία καθιστούν τη νέα βάση δεδομένων του Excel απαραίτητο στοιχείο στον τομέα της επεξεργασίας φυσικής γλώσσας. Η αποτελεσματικότητά του, η ευκολία και η ευελιξία του δίνουν τη δυνατότητα στους χρήστες να ξεκλειδώσουν τις πραγματικές δυνατότητες των γλωσσικών τους δεδομένων και να οδηγήσουν σε σημαντικές ιδέες και ανακαλύψεις.

Κεφάλαιο 4 Συζήτηση

Στο κεφάλαιο αυτό συζητείται η εξέλιξη και η μετάβαση από την παλιά μορφή της βάσης δεδομένων στη νέα βάση δεδομένων Excel, αναδεικνύοντας τις προηγούμενες αδυναμίες της παλιάς μορφής και αναλύοντας τις δυνατότητες που προσφέρει το Excel.

Όπως παρουσιάστηκε στο κεφάλαιο 2, η παλιά βάση δεδομένων αποτελούνταν από διαφορετικά αρχεία txt τα οποία ήταν προσβάσιμα μέσω του Libre Office. Γίνεται κατανοητό ότι από τη χρήση αυτής της παλιάς δομής προέκυπταν προκλήσεις και δυσκολίες στη διαχείριση και την αξιοποίηση των δεδομένων.

Όσον αφορά τη μορφολογική διεργασία και την ανάλυσή της, παρουσιάζεται μια σειρά από προβλήματα τα οποία χρειάζεται να αντιμετωπιστούν κατά την υλοποίηση του συστήματος μορφολογικής επεξεργασίας (Παπακίτσος 2000).

- Για παράδειγμα, πρώτα-πρώτα στην υπολογιστική εφαρμογή έπρεπε να αντιμετωπιστεί ο προβληματισμός που υπήρχε μεταξύ του ορισμού της λέξης και της υπαρκτής λέξης. Καθώς το σύστημα σχεδιάστηκε με σκοπό την επεξεργασία των λέξεων στη γραπτή τους μορφή. Έτσι, ως λέξη ορίστηκε ο αριθμός χαρακτήρων που περιέχονται ανάμεσα σε χαρακτήρες κενού ή σημείων στίξης ή άλλων ειδικών χαρακτήρων. Οι χαρακτήρες της λέξης πρέπει να ανήκουν στο Ελληνικό ή Λατινικό αλφάβητο και στην αντίστοιχη ASCII μορφή τους ή να είναι ψηφία (0..9). Ως υπαρκτή λέξη ορίστηκε εκείνη η λέξη η οποία μπορεί να βρεθεί σε ένα κοινό ορθογραφικό ή ερμηνευτικό λεξικό της Νέας Ελληνικής, τα οποία χρησιμοποιήθηκαν για τη δημιουργία της βάσης.
  Μάλιστα οι ανεξάρτητοι από το περιβάλλον νόμοι επαναγραφής της Λεξικής Δομής δεν επαρκούσαν για την υπολογιστική κάλυψη του φαινομένου (Papakitsos 1997 στο Παπακίτσος 2000).
  Αυτή η ανεπάρκεια προέρχεται από το γεγονός ότι στους νόμους παραγωγής και σύνθεσης της μορφολογίας δεν υπάρχει προκαθορισμένη συνθήκη τερματισμού. Αυτό σημαίνει ότι οι διεργασίες μπορεί να εκτελούνται συνεχώς χωρίς ορισμένο σημείο διακοπής. Έτσι η συγκεκριμένη έλλειψη μπορούσε να οδηγήσει στην ανάπτυξη ατέρμονων βρόχων, με αποτέλεσμα το πρόγραμμα να μην μπορεί να διακοπεί ομαλά.
  Οι επιπτώσεις αυτού του προβλήματος ήταν ιδιαίτερα εμφανείς στη γεννήτρια λέξεων, καθώς μπορούσε να προκύψει υπερπαραγωγή λέξεων. Αυτό σημαίνει

ότι η γεννήτρια μπορούσε να παράγει ατέρμονες ακολουθίες λέξεων που δεν ήταν πραγματικές ή δεν είχαν κατάληξη.

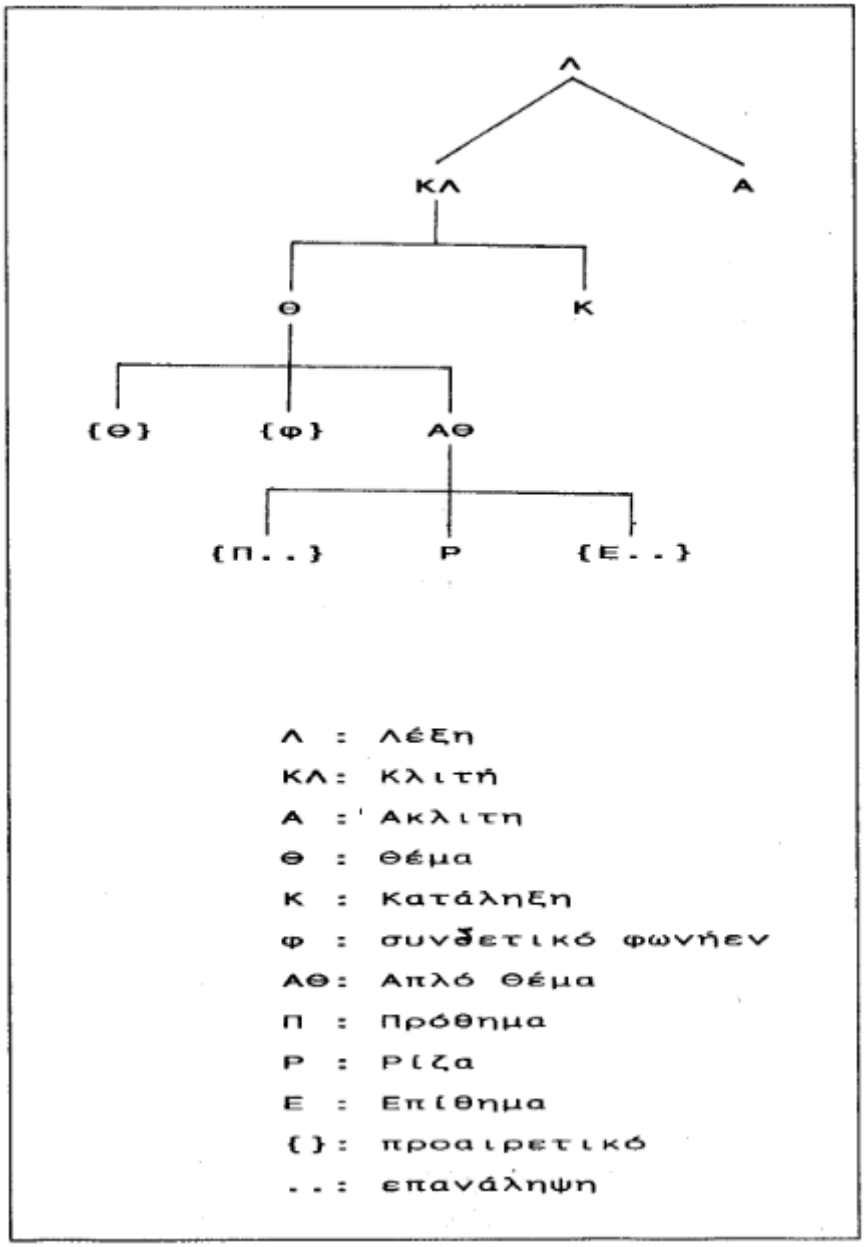
Επιπλέον, αυτό το πρόβλημα μπορούσε να επηρεάσει επίσης τον αυτόματο αναλυτή, ο οποίος θα μπορούσε να αντιμετωπίσει δυσκολίες στον προσδιορισμό του τερματισμού μιας λέξης ή θα μπορούσε να εξάγει λανθασμένα αποτελέσματα λόγω της ατέρμονης φύσης της διαδικασίας.

Για να αποφευχθούν αυτές οι επιπτώσεις, ήταν σημαντικό να υπάρχει καλή διαχείριση της μορφολογικής επεξεργασίας και να ληφθούν υπόψη οι συνθήκες τερματισμού των διεργασιών.

Ωστόσο, στην πραγματικότητα δεν παρατηρήθηκαν λέξεις με έναν πολύ μεγάλο αριθμό μορφημάτων και μάλλον η διαδικασία ήταν πεπερασμένη. Οι νόμοι επαναγραφής ήταν κατάλληλοι για την περιγραφή της δημιουργίας μιας λέξης. Παρόλα αυτά, για την ανάλυση μιας λέξης ήταν αναγκαίος ένας τρόπος που θα έδινε έμφαση στη δομή της, η οποία σχηματιζόταν από τους νόμους επαναγραφής, που επιτρέπουν να λειτουργήσει και η αντίστροφη διαδικασία.

Ήταν δηλαδή αναγκαία η ανάπτυξη της μεθόδου ανάλυσης που θα επέτρεπε την αντίστροφη μετατροπή μιας λέξης στη δομή της πριν από την εφαρμογή των νόμων επαναγραφής. Αυτή η διαδικασία στην ουσία θα επέτρεπε την αποτελεσματική ανάλυση της λέξης και την επαναφορά της στην αρχική της μορφή.

Στις Εικόνες 4 και 5 παρουσιάζεται η δομή των λέξεων στην παλιά βάση δεδομένων του ηλεκτρονικού λεξικού.



Εικόνα 4 Η Δομή των Λέξεων στην παλιά βάση δεδομένων (Παπακίτσος 2000)

Εφαρμόζοντας τους παραπάνω νόμους, καταλήγουμε στη δημιουργία λέξεων που έχουν τη μορφή της Εικόνας 5, από την οποία προκύπτει η δομή των λέξεων της Νέας Ελληνικής. Αυτή η περιγραφή αναφέρεται τόσο στον αριθμό όσο και στη σχετική θέση των μορφημάτων στο σύνολο της λέξης, ενώ οι συντελεστές αυτοί προσδιορίστηκαν μέσω ανάλυσης ελεύθερου κειμένου και είναι ικανοί να υποστηρίξουν έναν μορφολογικό αναλυτή.

$\lambda=\Sigma\mu_m$	λ: λέξη μ: μόρφημα κάθε τύπου m: πλήθος μορφημάτων (>0)
$\lambda=\Sigma\theta_n+\Sigma\kappa_I$	θ: θέμα n: πλήθος θεμάτων (>0) κ: κατάληξη I: {0,1}
$\theta=\Sigma\pi_k+\rho+\Sigma\varepsilon_l$	π: πρόθημα k: πλήθος προθημάτων (>=0) ρ: ρίζα ε: επίθημα l: πλήθος επιθημάτων (>=0)

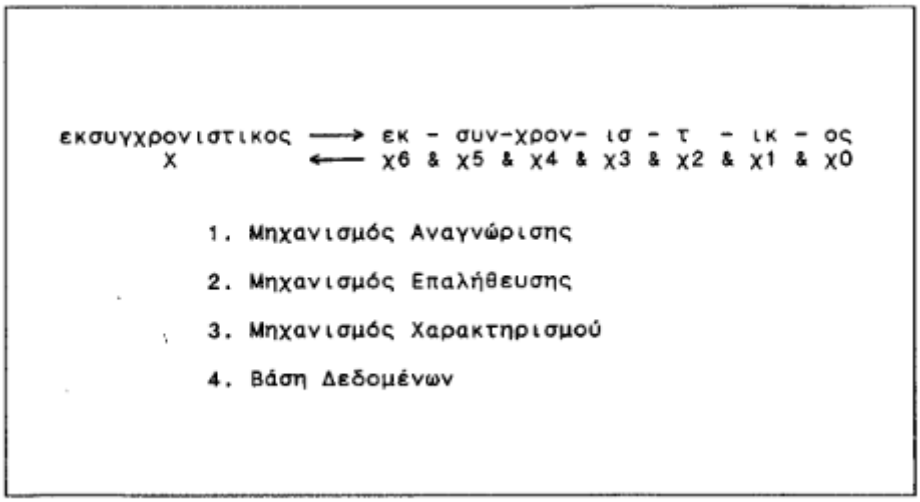
Εικόνα 5 Περιγραφή της δομής των λέξεων της παλιάς βάσης δεδομένων (Παπακίτσο, 2000)

- Όσον αφορά την υπολογιστική διαχείριση, για να αποδώσει ο μορφολογικός επεξεργαστής το σύνολο των χαρακτηριστικών συγκεκριμένης λέξης από τα σύνολα των χαρακτηριστικών των στοιχείων της πρέπει να αποτελείται από 4 μέρη τα οποία είναι (Εικόνα 6):
  - ο *μηχανισμός αναγνώρισης*, που αναγνωρίζει ένα μόρφημα στο εσωτερικό της λέξης,
  - ο *μηχανισμός επαλήθευσης*, που επαληθεύει τη σωστή ανεύρεση του μορφήματος,
  - ο *μηχανισμός χαρακτηρισμού*, που θα αποδώσει τις ιδιότητες 'χ' από τις 'χ0'-'χ6', και τέλος
  - η *βάση δεδομένων*, που περιέχει τις απαραίτητες πληροφορίες για την παραπάνω διαδικασία (Ηλεκτρονικό Λεξικό). Στη βάση δεδομένων, κάθε μόρφημα θα πρέπει να είναι καταχωρημένο και να συνδέεται με τα χαρακτηριστικά του (για παράδειγμα [συν, χ5], [ικ-χ1].

Η κυριότερη διαφορά μεταξύ των διαφόρων μοντέλων μορφολογικής επεξεργασίας βρίσκεται στον μηχανισμό αναγνώρισης και στη βάση δεδομένων. Ο μηχανισμός

αναγνώρισης της Λειτουργικής Αποσύνθεσης είναι αυτός που χρησιμοποιήθηκε στο παρόν σύστημα.

Στο σημείο αυτό πρέπει να σημειωθεί ότι η κύρια διαφορά μεταξύ των μοντέλων μορφολογικής επεξεργασίας βρίσκεται στον μηχανισμό αναγνώρισης και στη βάση δεδομένων που χρησιμοποιούνται και ότι στο συγκεκριμένο λεξικό έγινε χρήση του μηχανισμού αναγνώρισης της Λειτουργικής Αποσύνθεσης (Παπακίτσος 2000).



Εικόνα 6 Ο μηχανισμός ενός αυτόματου μορφολογικού αναλυτή στην παλιά βάση δεδομένων (Παπακίτσος 2000)

Η εισαγωγή τώρα της νέας βάσης δεδομένων σε μορφή Excel μεταμόρφωσε το προγενέστερο σύνολο αρχείων κειμένου που ήταν δύσκολα προσβάσιμο σε ένα αρχείο εύκολης διαχείρισης δεδομένων. Αυτή η αλλαγή δημιούργησε σημαντικά οφέλη και αξιοσημείωτες δυνατότητες.

Πρώτα απ' όλα, η μετάβαση σε μια βάση δεδομένων σε μορφή Excel προσφέρει ευκολία στην πρόσβαση και τη διαχείριση των δεδομένων. Με τη χρήση φύλλων εργασίας, τα κελιά και τις λειτουργίες του Excel, οι χρήστες μπορούν να οργανώσουν, να φιλτράρουν και να αναζητήσουν πληροφορίες με ευκολία. Επιπλέον, η δυνατότητα προσθήκης προσαρμοσμένων μακροεντολών και σεναρίων επεξεργασίας επιτρέπει την αυτοματοποίηση εργασιών και την αύξηση της παραγωγικότητας.

Θα πρέπει επίσης να σημειωθεί ότι η νέα βάση δεδομένων σε μορφή Excel παρέχει ευελιξία στην αναλυτική επεξεργασία των δεδομένων. Η μορφή βάσης δεδομένων

Excel ενισχύει επίσης τη συνεργασία και την κοινή χρήση δεδομένων μεταξύ των χρηστών. Με την αποθήκευση των πληροφοριών σε μια κεντρική και τυποποιημένη μορφή, γίνεται ευκολότερη η ανταλλαγή δεδομένων, η συνεργασία σε έργα και η διατήρηση της ακεραιότητας των δεδομένων. Η δυνατότητα προστασίας και ασφάλειας της βάσης δεδομένων, χρησιμοποιώντας τις ενσωματωμένες δυνατότητες ασφαλείας του Excel, διασφαλίζει περαιτέρω την εμπιστευτικότητα και την ακεραιότητα των δεδομένων.

Επιπλέον, η μορφή βάσης δεδομένων Excel προσφέρει απρόσκοπτη ενοποίηση με άλλο λογισμικό και συστήματα, χάριν στη συμβατότητα του Excel με διάφορα εργαλεία ανάλυσης δεδομένων, γλώσσες προγραμματισμού και πλαίσια αναφοράς, κατά συνέπεια βελτιστοποιημένες ροές εργασιών. Το γεγονός αυτό ανοίγει νέες δυνατότητες για προηγμένη επεξεργασία δεδομένων, παρέχοντας μια ολοκληρωμένη λύση για τη διαχείριση και την ανάλυση γλωσσικών δεδομένων.

Για να επιτευχθεί η παραπάνω μετατροπή, αναπτύχθηκε ο κώδικας εφαρμογής GUI (γραφικού περιβάλλοντος) που χρησιμοποίησε τη βιβλιοθήκη Tkinter στη γλώσσα προγραμματισμού Python. Η εφαρμογή στην ουσία δημιουργεί μια απλή διεπαφή για τη δημιουργία μιας βάσης δεδομένων από ένα αρχείο εξαγωγής. Παρατίθεται εξ ολοκλήρου παρακάτω:

```
# Import the library
from tkinter import *
from tkinter import filedialog
from docx import Document
import os
import unicodedata
from openpyxl import Workbook


# Create an instance of window
win=Tk()
win.title('Database Maker')
```

```
# Set the geometry of the window
win.geometry("400x150")


# Create a label
Label(win, text="Choose database export file type", font='Arial 16 bold').pack(pady=15)


def remove_control_characters(s):
    return ''.join(c for c in s if unicodedata.category(c)[0] != 'C')
```

- Αρχικά εισάγονται οι απαραίτητες βιβλιοθήκες για τη δημιουργία του γραφικού περιβάλλοντος, την επιλογή αρχείων, την επεξεργασία αρχείων Word (.docx), τη διαχείριση αρχείων και τη δημιουργία βιβλίων Excel.
- Δημιουργείται το παράθυρο της εφαρμογής και ορίζεται ο τίτλος και οι διαστάσεις του παραθύρου.
- Δημιουργείται μια ετικέτα με κείμενο που λέει "Choose database export file type" και τοποθετείται στο παράθυρο της εφαρμογής.
- Ορίζεται μια συνάρτηση για την αφαίρεση των χαρακτήρων ελέγχου, χρησιμοποιείται δηλαδή για να αφαιρέσει τους χαρακτήρες ελέγχου από μια συμβολοσειρά.

```
# Function to open a file in the system
def open_file_word():
    filepath = filedialog.askopenfilename(title="Choose database export file type", filetypes=(("text    files","*.txt"), ("all files","*.*")))
    path = os.path.split(filepath)[0]
    os.chdir(path)
    for file in os.listdir():
        if file.endswith('.TXT'):
            file_path=f"{path}/{file}"
```

```
        h=Document()
        f=open(file_path,mode='r',encoding='cp737')
        for i in f:
            if i[0] == ':':
                w=remove_control_characters(i)
                paragraph=h.add_paragraph(w)
            else:
                w=remove_control_characters(i)
                paragraph=h.add_paragraph(w)
        f.close()
        h.save(os.path.splitext(file_path)[0]+'.docx')


def open_file_excel():
    filepath = filedialog.askopenfilename(title="Choose database export file type", filetypes=(("text    files","*.txt"), ("all files","*.*")))
    path = os.path.split(filepath)[0]
    os.chdir(path)
    for file in os.listdir():
        if file.endswith('.TXT'):
            file_path=f"{path}/{file}"
            wb = Workbook()
            ws = wb.active
            print(os.path.splitext(file_path)[0].split("/")[-1])
            f=open(file_path,mode='r',encoding='cp737')
            ws.title = os.path.splitext(file_path)[0].split("/")[-1]
            for i in f:
```

```
            if i[0] == ':':
                w=remove_control_characters(i)


                ws.append(w.split())
            else:
                w=" "+remove_control_characters(i)
                ws.append(w.split(" "))
        wb.save(os.path.splitext(file_path)[0].split("/")[-1]+".xlsx")
        #df = pd.read_csv(file_path,delimiter=' ',encoding='cp737',header=1)
    # can replace with:
    # df = pd.read_csv('input.tsv', sep='\t') for tab delimited
        #df.to_excel(os.path.splitext(file_path)[0]+'.xlsx')


# Create a button to trigger the dialog
button = Button(win, text="Word", command=open_file_word)
button.pack()
button1 = Button(win, text="Excel", command=open_file_excel)
button1.pack()


win.mainloop()
```

- Ορίζονται συναρτήσεις για το άνοιγμα αρχείων Word και Excel. Οι εν λόγω συναρτήσεις καλούνται όταν ο χρήστης επιλέγει να ανοίξει ένα αρχείο Word ή Excel αντίστοιχα.
- Οι συναρτήσεις συνδέονται με κουμπιά και συγκεκριμένα δημιουργούνται δύο κουμπιά ("Word" και "Excel") τα οποία καλούν τις αντίστοιχες συναρτήσεις όταν πατηθούν.
- Τέλος εκτελείται η εφαρμογή.

Κεφάλαιο 5 Συμπεράσματα

Στην παρούσα εργασία αναπτύχθηκε η βελτίωση της βάσης ηλεκτρονικού λεξικού.

Η μετάβαση από τα διαφορετικά αρχεία txt σε μία βάση δεδομένων Excel προσφέρει απαράμιλλη ευκολία στην πρόσβαση και τον χειρισμό των δεδομένων του ηλεκτρονικού λεξικού.

Με τη χρήση φύλλων εργασίας, κελιών και των ισχυρών λειτουργιών του Excel, οι χρήστες μπορούν να οργανώσουν, να φιλτράρουν και να αναζητήσουν πληροφορίες χωρίς κόπο. Η δυνατότητα προσθήκης προσαρμοσμένων μακροεντολών και σεναρίων επεξεργασίας επιτρέπει περαιτέρω την αυτοματοποίηση εργασιών, αυξάνοντας την παραγωγικότητα σε νέα ύψη.

Τα συμπεράσματα από την εν λόγω μετάβαση, θα μπορούσαν να συνοψιστούν στα εξής:

- *Βελτίωση της λειτουργικότητας*: Η τροποποίηση του ηλεκτρονικού λεξικού και η μετάβαση στη νέα βάση δεδομένων του Excel έχουν βελτιώσει σημαντικά τη λειτουργικότητα και την αποδοτικότητα του λεξικού. Οι χρήστες τώρα μπορούν να αναζητούν και να ανακτούν πληροφορίες με μεγαλύτερη ευκολία και ταχύτητα, χρησιμοποιώντας τις πλούσιες δυνατότητες του Excel για φιλτράρισμα, ταξινόμηση και υπολογισμούς.
- *Ενίσχυση της αξιοπιστίας των δεδομένων*: Η νέα βάση δεδομένων του Excel παρέχει ένα αξιόπιστο και οργανωμένο περιβάλλον για την αποθήκευση και διαχείριση των δεδομένων του λεξικού. Η χρήση του Excel επιτρέπει την εφαρμογή αυστηρών ελέγχων ποιότητας, τον έλεγχο της συνέπειας και την αποτροπή των λαθών στην καταχώρηση κι επεξεργασία των δεδομένων.
- *Δυνατότητα ανάλυσης δεδομένων*: Η μετάβαση στη βάση δεδομένων του Excel ανοίγει νέες δυνατότητες γι’ ανάλυση δεδομένων. Οι χρήστες μπορούν να αξιοποιήσουν τα ισχυρά εργαλεία ανάλυσης δεδομένων του Excel, όπως συγκεντρωτικούς πίνακες, για να αποκτήσουν βαθύτερες πληροφορίες και να δημιουργήσουν ουσιαστικές αναφορές με βάση τα δεδομένα του λεξικού.
- *Απλοποιημένη συντήρηση και ενημερώσεις*: Η χρήση του Excel ως βάσης δεδομένων απλοποιεί τη συντήρηση και τις ενημερώσεις του ηλεκτρονικού λεξικού. Με τη γνώριμη διεπαφή του Excel και τις διαισθητικές δυνατότητες επεξεργασίας, η προσθήκη, η τροποποίηση και η διαγραφή καταχωρήσεων

δεδομένων μπορεί να γίνει εύκολα και αποτελεσματικά, διασφαλίζοντας ότι το λεξικό παραμένει ενημερωμένο.

- *Απόλυτη ενσωμάτωση σε άλλες εφαρμογές*: Τόσο η συμβατότητα όσο και η ευρεία χρήση του Excel καθιστούν εύκολη την ενσωμάτωση του ηλεκτρονικού λεξικού με άλλες εφαρμογές ή συστήματα. Δηλαδή τα δεδομένα του λεξικού μπορούν να εισαχθούν ή να εξαχθούν απρόσκοπτα σε διάφορες μορφές, διευκολύνοντας τη διαλειτουργικότητα με άλλα εργαλεία.

Βιβλιογραφία

Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. Reading and writing, 12, 169-190.

Garfinkel, S. L. (2012). Programming Unicode. Login, 37(2), 1-13.

Goel, B. (2017). Developments in The Field of Natural Language Processing. International Journal of Advanced Research in Computer Science, 8(3).

Hijzelendoorn, M., & Cremers, C. (2009). An object-oriented and fast lexicon for semantic generation. arXiv preprint arXiv:0905.3318.

Lauriola, I., Lavelli, A., & Aiolli, F. (2022). An introduction to deep learning in natural language processing: Models, techniques, and tools. Neurocomputing, 470, 443-456.

Marantz, A. (2013). No escape from morphemes in morphological processing. Language and cognitive processes, 28(7), 905-916.

Maxwell, M., & Poser, W. (2004). Morphological interfaces to dictionaries. In Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries (pp. 65-68).

McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. Python for high performance and scientific computing, 14(9), 1-9.

Papakitsos, E., Grigoriadou, M., & Philokyprou, G. (2002). Modelling a Morpheme-based Lexicon for Modern Greek. Literary and linguistic computing, 17(4), 475-490.

Ralli, A. (2002). The role of morphology in gender determination: evidence from Modern Greek.

Schreuder, R., & Baayen, R. H. (1995). Modeling morphological processing. Morphological aspects of language processing, 2, 257-294.

Stankovic, R., Krstev, C., Lazic, B., & Škoric, M. (2018). Electronic dictionaries-from file system to lemon based lexical database. In Proceedings of LREC (pp. 18-W23).

Tauber, J. K. (2019). Character encoding of classical languages. 2019). Digital classical philology: Ancient Greek and Latin in the digital revolution, 137-158.

Van Rossum, G. (2007, June). Python Programming Language. In USENIX annual technical conference (Vol. 41, No. 1, pp. 1-36).

Παπακίτσος, Ε. (2000). Συμβολή στη μορφολογική επεξεργασία της Νέας Ελληνικής: λειτουργική αποσύνθεση, καρτεσιανό ηλεκτρονικό λεξικό (Doctoral dissertation, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών (ΕΚΠΑ). Σχολή Θετικών Επιστημών. Τμήμα Πληροφορικής).

Ιστοσελίδες

IBM (2023). https://www.ibm.com/topics/natural-language-processing