Κατάτμηση συνεχούς κειμένου με χρήση συσχετιστικού λεξικού, στατιστικών στοιχείων γλώσσας και τεχνικών οπισθοδρόμησης

Νικόλαος Βασιλάς, Anuj Sharma

1. Εισαγωγή

Με την διαρκώς αυξανόμενη εξάρτηση από τους Η/Υ για αποθήκευση, διαχείριση και επεξεργασία δεδομένων, το πρόβλημα της αποτελεσματικής και ακριβούς ανάκλησης βασισμένης σε ατελείς ή μολυσμένες από θόρυβο λέξεις κλειδιά γίνεται ολοένα και πιο σημαντικό. Ένα από τα πιο ενοχλητικά προβλήματα των σύγχρονων υπολογιστών είναι ο αδέξιος και αφύσικος χειρισμός που παρέχουν σε ασαφή και ατελή δεδομένα. Αυτό οφείλεται στο γεγονός ότι προκειμένου οι Η/Υ να αποθηκεύουν και να διαχειρίζονται δεδομένα με αποτελεσματικό τρόπο, τα δεδομένα αυτά πρέπει να μην περιέχουν πλεονασμούς. Σε αντίθεση, οι άνθρωποι από τη φύση τους επεξεργάζονται πλεονασματικές πληροφορίες και μπορούν εύκολα να ανεχθούν σφάλματα (θόρυβο), που αλλοιώνουν τα δεδομένα, εκμεταλλευόμενοι τον πλεονασμό τους (δηλαδή, τα συμφραζόμενα).

Ένα πρόβλημα που ανήκει στην παραπάνω κατηγορία είναι και αυτό της κατάτμησης συνεχούς κειμένου σε λέξεις, δηλαδή κειμένου χωρίς σημεία στίξης και κενά μεταξύ των λέξεων. Αυτό το πρόβλημα, αν και απαντάται συχνά στην κατάτμηση κειμένων που έχουν γραφεί κυρίως σε γλώσσες της ανατολικής Ασίας [1,2], δεν έχει απασχολήσει σημαντικά την ελληνική επιστημονική κοινότητα. Ο λόγος είναι ότι τα νέα ελληνικά, όπως και οι δυτικές γλώσσες, χρησιμοποιούν διαχωριστές μεταξύ των λέξεων (κενά, σημεία στίξης) με αποτέλεσμα η κατάτμηση να είναι δεδομένη. Όμως, υπάρχουν περιπτώσεις, όπως αυτή της συνεχούς μικρογράμματης ή και μεγαλογράμματης γραφής, που συναντάμε σε παλαιοχριστιανικά χειρόγραφα αλλά και σε αρχαία κείμενα (π.χ. πινακίδες της κλασικής ή της ελληνιστικής περιόδου) που δίνουν έναυσμα για ουσιαστική διερεύνηση των δυνατοτήτων αυτόματης κατάτμησης μέσω ηλεκτρονικού υπολογιστή.

Το κίνητρο αυτής της έρευνας προέρχεται από ένα εθνικό πρόγραμμα αναγνώρισης συνεχούς μικρογράμματης γραφής παλαιοχριστιανικών χειρογράφων από τη Μονή της Αγίας Αικατερίνης του Όρους Σινά [3]. Επειδή, η μετατροπή των χειρογράφων σε ψηφιακά κείμενα θα γίνεται με ένα σύστημα αυτόματης αναγνώρισης χαρακτήρων (OCR) το οποίο αναγκαστικά εισάγει λάθη αναγνώρισης, γίνεται η υπόθεση ότι ο πίνακας, με τις πιθανότητες ορθής αναγνώρισης των χαρακτήρων αλλά και τις πιθανότητες για κάθε είδος σφάλματος όπως, εισαγωγής, διαγραφής ή αντικατάστασης χαρακτήρων, είναι γνωστός και ότι προκύπτει από πειραματικά αποτελέσματα αναγνώρισης σε πραγματικά χειρόγραφα.

Η προτεινόμενη μεθοδολογία περιλαμβάνει:

α) τη χρήση ενός λεξικού αποθηκευμένου σε μνήμη πίνακα συσχέτισης με δομή ιεραρχικού αντίστροφου αρχείου ώστε να εξασφαλίζεται ταχύτατη αναζήτηση λέξεων και διόρθωση λαθών,

β) κατάταξη των ανακληθεισών λέξεων μέσω κριτηρίου που συνδυάζει την απόσταση Levenshtein με στατιστικά της ελληνικής γλώσσας, και

γ) κατάτμηση του συνεχούς κειμένου με οδηγό τις πιθανότερες λέξεις που αρχίζουν από την τρέχουσα θέση του δρομέα και με χρήση τεχνικών οπισθοδρόμησης στην περίπτωση που η τρέχουσα κατάτμηση οδηγείται σε αδιέξοδο.

2. Αποθήκευση Λεξικού σε Μνήμη Πίνακα Συσχέτισης

Στην ενότητα αυτή περιγράφεται η μέθοδος εξαγωγής χαρακτηριστικών και κωδικοποίησης των λέξεων που θα ακολουθηθεί σε όλη την εργασία καθώς και η συσχετιστική μνήμη για την αποθήκευση και συμπίεση του λεξικού. Επιπλέον, περιγράφεται και η μέθοδος ανάκλησης από τη συσχετιστική μνήμη. Το σύνολο αυτών των μεθόδων έχει ως αποτέλεσμα: α) οικονομία αποθηκευτικού χώρου, β) μεγάλη ταχύτητα ανάκλησης, και γ) δυνατότητες προσεγγιστικής ταύτισης (δηλαδή, ορθογραφικής διόρθωσης) όταν η λέξη-κλειδί περιέχει ορθογραφικά λάθη.

2.1 Εξαγωγή χαρακτηριστικών και κωδικοποίηση λέξεων

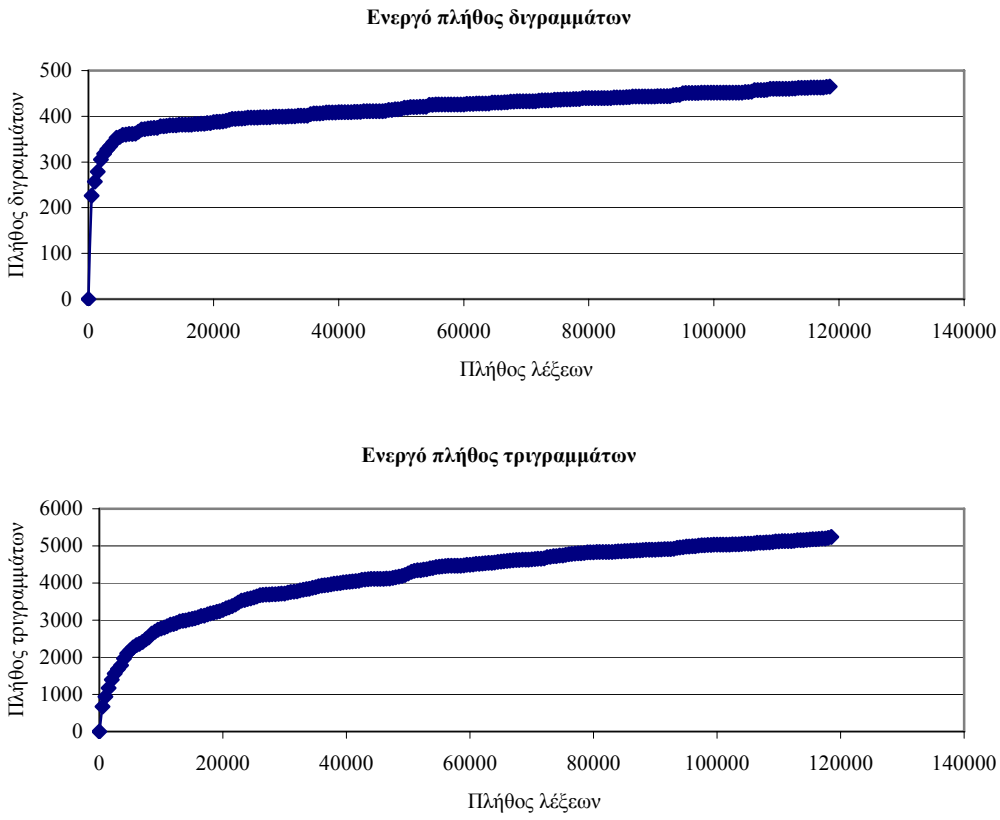
Η έρευνα των τελευταίων δεκαετιών πάνω στη διόρθωση ορθογραφικών λαθών με χρήση λεξικού έχει καταδείξει ότι η κωδικοποίηση των λέξεων με τη μέθοδο των ν-γραμμάτων οδηγεί σε καλύτερα αποτελέσματα όσον αφορά την ποιότητα και την ταχύτητα ανάκλησης [4,5]. Η κωδικοποίηση μιας λέξης με ν-γράμματα είναι μια μέθοδος εξαγωγής χαρακτηριστικών με το κάθε χαρακτηριστικό να σχηματίζεται από μια ομάδα ν γειτονικών γραμμάτων της λέξης. Για παράδειγμα, τα μονογράμματα, διγράμματα και τριγράμματα (δηλαδή για ν = 1, 2 ή 3 αντίστοιχα) της λέξης ΚΕΙΜΕΝΟ παρουσιάζονται στον Πίνακα 1. Το σύμβολο '−' που εμφανίζεται πριν το 'Κ', αντιστοιχεί στο κενό που προηγείται της λέξης.

Πίνακας 1. Αναπαράσταση λέξης με ν-γράμματα.

	Λέξη: ΚΕΙΜΕΝΟ						
Μονογράμματα:	Κ	Ε	Ι	Μ	Ε	Ν	Ο
Διγράμματα:	−Κ	ΚΕ	ΕΙ	ΙΜ	ΜΕ	ΕΝ	ΝΟ
Τριγράμματα:	−−Κ	−ΚΕ	ΚΕΙ	ΕΙΜ	ΙΜΕ	ΜΕΝ	ΕΝΟ

Ένα στατιστικό στοιχείο που μπορεί κανείς να εκμεταλλευτεί (π.χ. στην ανίχνευση και διόρθωση λαθών) αφορά τη συχνότητα εμφάνισης των ν-γραμμάτων σε μια γλώσσα. Όταν, για παράδειγμα, εμφανίζεται κάποιο ν-γραμμα που δεν είναι δυνατόν να υπάρχει σε μια συγκεκριμένη γλώσσα (π.χ. το τρίγραμμα ‘ααα’) τότε γνωρίζουμε ότι έχει γίνει λάθος. Αν και στην παρούσα εργασία δεν χρησιμοποιούνται τα στατιστικά στοιχεία των ν-γραμμάτων, έχει ενδιαφέρον να παρατηρήσει κανείς ότι το πλήθος των διγραμμάτων και τριγραμμάτων που εμφανίζονται σε λεξικό 118.603 λέξεων είναι αρκετά μικρότερο του θεωρητικού μέγιστου (βλ. Σχ. 1).

Ας θεωρήσουμε ότι μια συμβολοσειρά σχηματίζεται από τους χαρακτήρες του αλφαβήτου μιας γλώσσας. Ένας κλασικός τρόπος για να ελέγξουμε αν αντιστοιχεί σε λέξη της γλώσσας είναι η σύγκρισή της με ένα λεξικό. Στην περίπτωση που υπάρχει *ακριβής ταύτιση* με κάποια λέξη του λεξικού τότε η αναγνώριση τελειώνει επιτυχώς. Ειδάλλως, θεωρούμε ότι υπάρχουν λανθασμένοι χαρακτήρες στη συμβολοσειρά και ακολουθούμε μια διαδικασία *προσεγγιστικής ταύτισης* σύμφωνα με την οποία η συμβολοσειρά και οι λέξεις του λεξικού πρώτα κωδικοποιούνται σε αριθμητικά διανύσματα του ιδίου μήκους και στη συνέχεια συγκρίνονται με χρήση κάποιας μετρικής απόστασης. Η πλησιέστερη, σύμφωνα με την επιλεγμένη μετρική, λέξη του λεξικού θα είναι και η απάντηση της διαδικασίας ταύτισης.



Σχ. 1. Το πλήθος των διαφορετικών διγραμμάτων και τριγραμμάτων που εμφανίζονται σε ένα λεξικό 118.603 λέξεων.

Στο Σχ. 2 παρουσιάζεται η κωδικοποίηση σε αριθμητικό δυαδικό διάνυσμα της λέξης ΚΕΙΜΕΝΟ χρησιμοποιώντας μονογράμματα. Το μήκος του διανύσματος θα είναι 24, δηλαδή όσο και το μέγεθος του αλφαβήτου, με την κάθε θέση στο διάνυσμα να αντιστοιχεί και σε ένα μονόγραμμα. Για διγράμματα ή τριγράμματα το μήκος του διανύσματος θα είναι 25^2 ή 25^3 αντίστοιχα, καθώς θεωρούμε ότι το αλφάβητο προσαυξάνεται με το σύμβολο του κενού. Τα στοιχεία του δυαδικού διανύσματος υποδηλώνουν την ύπαρξη (`1`) ή απουσία (`0`) του αντίστοιχου ν-γράμματος.

```
KEIMENO→(0 0 0 0 1 0 0 0 1 1 0 1 1 0 1 0 0 0 0 0 0 0 0 0)
```

Σχ. 2. Παράδειγμα κωδικοποίησης λέξης με μονογράμματα.

Εκτεταμένες έρευνες στο παρελθόν [4] έδειξαν ότι η κωδικοποίηση των λέξεων με ν-γράμματα παρουσιάζει τρία σημαντικά πλεονεκτήματα έναντι άλλων, π.χ. ASCII, κωδικοποιήσεων. Το πρώτο πλεονέκτημα είναι ότι δημιουργεί αριθμητικά διανύσματα σταθερού μήκους (ανεξάρτητα από το πλήθος των χαρακτήρων σε μια λέξη) διευκολύνοντας με αυτόν τον τρόπο την εύρεση των *χαρακτηριστικών* δύο συμβολοσειρών που ταυτίζονται ώστε να υπολογισθεί η ομοιότητά τους. Το

δεύτερο πλεονέκτημα αφορά την ικανοποίηση του βασικού κριτηρίου κωδικοποίησης: *όμοιες λέξεις πρέπει να κωδικοποιούνται σε κοντινά διανύσματα στον χώρο των χαρακτηριστικών*. Αυτό είναι προφανές καθώς τα απλά ορθογραφικά λάθη επηρεάζουν μόνο ένα μικρό ποσοστό των ν-γραμμάτων (χαρακτηριστικών) μιας λέξης. Τέλος, το τρίτο πλεονέκτημα είναι ότι η κωδικοποίηση με ν-γράμματα δημιουργεί *αραιά* διανύσματα χαρακτηριστικών (το πλήθος των μη μηδενικών στοιχείων είναι μικρότερο ή ίσο από το μήκος της λέξης) επιτρέποντας αποτελεσματικές αναπαραστάσεις του λεξικού και ταχύτατη ανάκληση λέξεων.

2.2 Η Μνήμη Πίνακα Συσχέτισης (CMM)

Μια συσχετιστική μνήμη είναι ένα σύστημα που συσχετίζει ζεύγη διανυσμάτων εισόδου και εξόδου. Κύριο χαρακτηριστικό των συσχετιστικών μνημών είναι η ανοχή στον θόρυβο, δηλαδή η δυνατότητα ανάκλησης της σωστής εξόδου ακόμη και αν η είσοδος είναι μολυσμένη με θόρυβο. Στην γενική περίπτωση που οι έξοδοι είναι διαφορετικές από τις εισόδους οι μνήμες αυτές ονομάζονται *μνήμες ετεροσυσχέτισης* ενώ αν οι έξοδοι κωδικοποιούν την κατηγορία στην οποία ανήκει η είσοδος τότε ονομάζονται *ταξινομητές*.

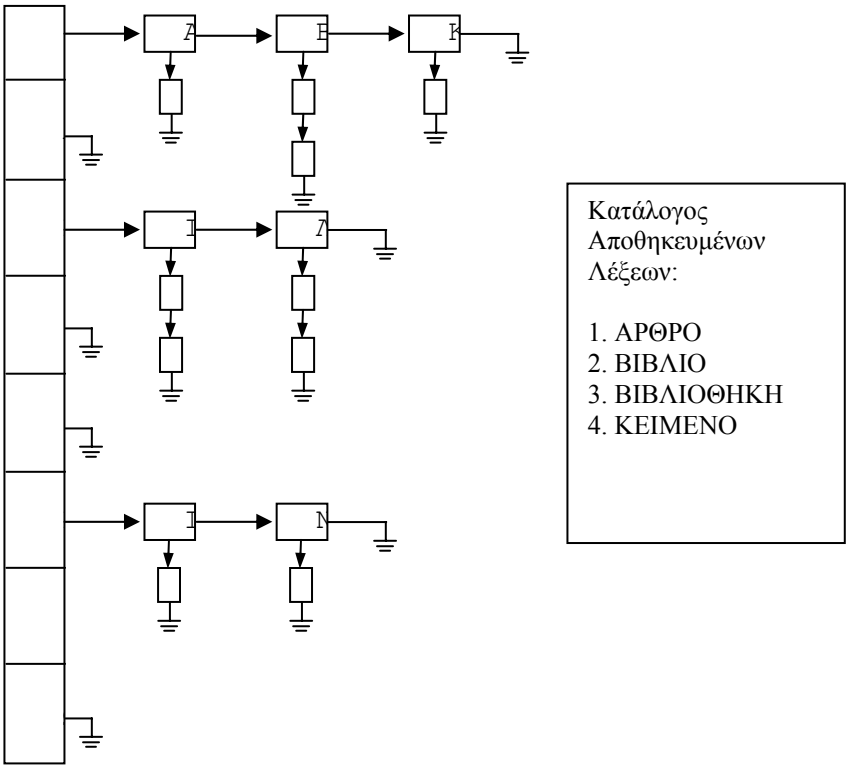
Ας θεωρήσουμε ένα λεξικό P λέξεων. Μετά την εξαγωγή ν-γραμμάτων, οι λέξεις κωδικοποιούνται σε δυαδικά διανύσματα στήλης $\mathbf{s_i}$ (i = 1, 2, ... P), που ονομάζονται διανύσματα εισόδων. Έστω, επίσης, ότι σε κάθε $\mathbf{s_i}$ αντιστοιχίζεται ένα διάνυσμα στήλης $\mathbf{r_i}$, που ονομάζεται διάνυσμα εξόδου. Η μνήμη πίνακα συσχέτισης (*Correlation Matrix Memory* ή CMM) ανήκει στις γραμμικές μνήμες [4,6] και περιγράφεται από την εξίσωση εξωτερικού γινομένου:

$$M = R\, S^T$$

όπου S και R είναι οι πίνακες εισόδου και εξόδου με στήλες τα $\mathbf{s_i}$ και $\mathbf{r_i}$ αντίστοιχα και S^T είναι ο ανάστροφος του S. Στην παρούσα εργασία, η κάθε λέξη του λεξικού, μαζί με τις παραλλαγές της, που προκύπτουν από ορθογραφικά λάθη, θα αντιστοιχεί σε μια κατηγορία. Συνεπώς, αν κωδικοποιήσουμε την κάθε μια από τις P κατηγορίες με μια στήλη του μοναδιαίου PxP πίνακα I, τότε R = I και $M = S^T$. Το λεξικό που αποθηκεύεται σε μια συσχετιστική μνήμη θα ονομάζεται στο εξής *συσχετιστικό λεξικό*.

Η ανάκληση από το συσχετιστικό λεξικό επιτυγχάνεται σε δύο στάδια: α) ταξινόμηση της λέξης εισόδου σε μια κατηγορία, και β) ανάκτηση της λέξης από πίνακα αναζήτησης (look-up table) μέσω άμεσης δεικτοδότησης, καθώς η κατηγορία θα είναι ο αύξων αριθμός της λέξης στον πίνακα αναζήτησης. Η ταξινόμηση της λέξης **s** επιτυγχάνεται σε δύο φάσεις: α) υπολογισμό της εξόδου της συσχετιστικής μνήμης **r** μέσω της σχέσης $\mathbf{r} = M\mathbf{s}$, και β) εύρεση της θέσης του μέγιστου στο διάνυσμα **r**. Η κατηγορία στην οποία θα ανήκει η είσοδος υποδηλώνεται από τη θέση του μέγιστου στο διάνυσμα εξόδου καθώς, $r = S^T \mathbf{s} = (\mathbf{s}_1{}^T\mathbf{s}, \mathbf{s}_2{}^T\mathbf{s}, \ldots \mathbf{s}_P{}^T\mathbf{s})^T$, με τα στοιχεία του να υπολογίζουν τα εσωτερικά γινόμενα (δηλαδή την ομοιότητα) των αποθηκευμένων διανυσμάτων εισόδου με τη νέα είσοδο. Το στοιχείο του **r** με τη μέγιστη ομοιότητα θα πρέπει να καθορίσει την κατηγορία στην οποία ανήκει η νέα, πιθανώς μολυσμένη από θόρυβο, είσοδος.

Μια αποτελεσματική αναπαράσταση της μνήμης πίνακα συσχέτισης που εκμεταλλεύεται την αραιότητα των διανυσμάτων χαρακτηριστικών είναι μέσω της δομής *ιεραρχικού αντίστροφου αρχείου* (ΙΑΑ). Στη δομή ΙΑΑ αντί να αποθηκεύονται τα ν-γράμματα της κάθε λέξης, αποθηκεύονται οι αύξοντες αριθμοί (κατηγορίες) των λέξεων που περιέχουν κάποιο ν-γραμμα. Επιπλέον, προκειμένου να επιταχυνθεί η προσθήκη νέων λέξεων που μπορεί να εισαγάγουν και νέα ν-γράμματα, στη δομή ΙΑΑ τα ν-γράμματα αποθηκεύονται σε διαφορετικές ομάδες ανάλογα με τον πρώτο χαρακτήρα (βλ. Σχ. 3).



Σχ. 3. Παράδειγμα δομής ΙΑΑ για 2-γράμματα, με τέσσερις αποθηκευμένες λέξεις.

Λόγω της παραπάνω κωδικοποίησης των λέξεων, στο συσχετιστικό λεξικό όλες οι λέξεις δεικτοδοτούνται από τα ν-γράμματά τους. Για το λεξικό των 118.603 λέξεων που χρησιμοποιήθηκε στην εργασία αυτή, ο χρόνος δημιουργίας του συσχετιστικού λεξικού σε PC Athlon στα 2.8GHz με 256MB RAM ήταν της τάξης των 5 sec ενώ ο μέσος χρόνος για την ανάκληση μιας λέξης ήταν της τάξης των 15 msec.

3. Μεθοδολογία Κατάτμησης Συνεχούς Κειμένου

Η μεθοδολογία κατάτμησης χρησιμοποιεί: α) το συσχετιστικό λεξικό για την ταχύτατη ανάκληση των Ν πλησιέστερων, ως προς τα κοινά ν-γράμματα με τη λέξη-κλειδί, λέξεων, β) συνδυασμό της μετρικής Levenshtein (για καλύτερη εκτίμηση της ομοιότητας) με το πλήθος των κοινών ν-γραμμάτων, τα μήκη των λέξεων και στατιστικά της γλώσσας με σκοπό την επιλογή της προτεινόμενης λέξης, και γ) δυνατότητες οπισθοδρόμησης και επιλογής άλλης προτεινόμενης λέξης όταν η κατάτμηση οδηγείται σε αδιέξοδο.

3.1 Η μετρική Levenshtein

Η πιο διαδεδομένη μετρική που χρησιμοποιείται σήμερα κυρίως στην ακαδημαϊκή έρευνα, για προσεγγιστική ταύτιση συμβολοσειρών, βασίζεται στην έννοια της απόστασης Levenshtein [4,5]. Η απόσταση Levenshtein μεταξύ δύο συμβολοσειρών είναι ο ελάχιστος αριθμός στοιχειωδών λειτουργιών ορθογραφικής διόρθωσης (δηλαδή, απλή αντικατάσταση, εισαγωγή ή διαγραφή χαρακτήρα) που απαιτούνται για τον μετασχηματισμό της μιας συμβολοσειράς στην άλλη. Για

παράδειγμα, LevDist(ΒΙΒΛΙΟ, ΒΙΒΛΙΟΘΗΚΗ) = 4 καθώς απαιτούνται τέσσερις εισαγωγές χαρακτήρων για να μετασχηματισθεί η πρώτη στη δεύτερη ενώ LevDist(ΒΙΒΛΙΟ, ΚΕΙΜΕΝΟ) = 5 (αντικατάσταση του πρώτου 'Β' με το 'Κ', εισαγωγή του 'Ε' και τρεις αντικαταστάσεις των 'Β', 'Λ' και 'Ι' με τα 'Μ', 'Ε' και 'Ν' αντίστοιχα).

Η μετρική Levenshtein ή οι παραλλαγές της που εισάγουν και την έννοια του κόστους σε κάθε στοιχειώδη ορθογραφική λειτουργία, δεν μπορούν να χρησιμοποιηθούν απ'ευθείας για την ανάκληση λέξεων από μεγάλα λεξικά. Ο λόγος είναι ότι ο αλγόριθμος δυναμικού προγραμματισμού για τον υπολογισμό της απόστασης Levenshtein δύο λέξεων με μήκη k_1 και k_2 χαρακτήρων απαιτεί χρόνο $O(k_1k_2)$ ενώ αν το λεξικό περιέχει P λέξεις τότε ο χρόνος θα είναι της τάξης του $O(P\ k^2)$ όπου k το μέσο μήκος των λέξεων. Η πολυπλοκότητα αυτή είναι απαγορευτική για λεξικά πολλών δεκάδων χιλιάδων λέξεων όπως αυτό που χρησιμοποιούμε στην παρούσα εργασία.

Η τεχνική που θα ακολουθήσουμε για να παρακάμψουμε το παραπάνω πρόβλημα είναι η ταχύτατη ανάκληση, σε πρώτο στάδιο, ενός συνόλου N λέξεων (υπολεξικό) μέσω της συσχετιστικής μνήμης και στη συνέχεια η καλύτερη αποτίμηση του βαθμού ομοιότητας των N λέξεων μέσω της απόστασης Levenshtein. Αν το υπολεξικό είναι αρκετά μεγάλο ώστε να περιέχει τη σωστή λέξη, τότε το αποτέλεσμα θα είναι το ίδιο (ως προς τη σωστή λέξη) με το να υπολογίζαμε την απόσταση Levenshtein της λέξης-κλειδί με κάθε λέξη του λεξικού.

3.2 Συλλογή στατιστικών ελληνικής γλώσσας

Εκτός της απόστασης Levenshtein για την επιλογή της λέξης του υπολεξικού με τη μεγαλύτερη ομοιότητα με το κλειδί, σε ένα πιθανοτικό μοντέλο, η επιλογή της λέξης μπορεί να γίνει χρησιμοποιώντας τα στατιστικά της γλώσσας. Για παράδειγμα, μπορεί κανείς να θεωρήσει ότι η πιο πιθανή λέξη είναι αυτή που απαντάται με τη μεγαλύτερη συχνότητα στη γλώσσα. Επίσης, είναι δυνατόν να ληφθούν συχνότητες εμφάνισης δύο διαδοχικών λέξεων (δηλαδή, το σύστημα να έχει μνήμη) ώστε το ζεύγος <προηγούμενη λέξη, προτεινόμενη λέξη από το υπολεξικό> να έχει τη μεγαλύτερη πιθανότητα.

Η συλλογή στατιστικών για την ελληνική γλώσσα έγινε με χρήση μιας μεγάλης συλλογής νεοελληνικών κειμένων. Για τον σκοπό αυτό δημιουργήθηκε ένα αρχείο που περιείχε 2.412.210 λέξεις από 70 τεύχη της διμηνιαίας εφημερίδας «ΤΟ ΚΑΠΟΔΙΣΤΡΙΑΚΟ» του Πανεπιστημίου Αθηνών.

3.3 Κατάτμηση κειμένου

Έστω ότι M συμβολίζει το μήκος της μεγαλύτερης λέξης του λεξικού. Η κατάτμηση του συνεχούς κειμένου αρχίζει με τον δρομέα στη θέση του πρώτου χαρακτήρα. Στη συνέχεια εξετάζονται διαδοχικά οι M συμβολοσειρές που έχουν αρχή τη θέση του δρομέα και μήκη M, M-1, ..., 2, 1, χαρακτήρων αντίστοιχα. Έστω S_i η συμβολοσειρά με μήκος i (i = 1, 2, ..., M).

Η μεθοδολογία κατάτμησης περιλαμβάνει τρεις φάσεις. Στην πρώτη φάση, για κάθε λέξη-κλειδί S_i, ανακαλούνται οι N λέξεις του υπολεξικού W_{ij} (i = 1, ..., N, j = 1, ..., M) με τη μεγαλύτερη ομοιότητα όσον αφορά τα κοινά ν-γράμματα. Αυτό έχει ως αποτέλεσμα την ανάκληση N·M λέξεων. Στη δεύτερη φάση, επιλέγεται η προτεινόμενη λέξη για την κατάτμηση έτσι ώστε να μεγιστοποιεί ένα συνολικό κριτήριο καταλληλότητας που συνδυάζει τα ακόλουθα έξι επιμέρους κριτήρια:

α) την κανονικοποιημένη απόσταση Levenshtein,

$$x_1 = \text{LevDist}(S_i, W_{ij})/\text{MaxLevDist}$$

όπου ο αριθμητής είναι η απόσταση των S_i, W_{ij} και *MaxLevDist* είναι η μέγιστη απόσταση μεταξύ δύο λέξεων του λεξικού. Η *MaxLevDist* είναι μικρότερη ή ίση με το μήκος της μεγαλύτερης λέξης του λεξικού.

β) το κανονικοποιημένο πλήθος των κοινών ν-γραμμάτων,

$$x_2 = [Ngrams(S_i) - CommonNgrams(S_i, W_{ij})]/Ngrams(S_i)$$

όπου *CommonNgrams*(S_i, W_{ij}) συμβολίζει το πλήθος των κοινών ν-γραμμάτων των S_i, W_{ij} και *Ngrams*(S_i) είναι το πλήθος των ν-γραμμάτων της S_i.

γ) τη συχνότητα εμφάνισης της λέξης, κανονικοποιημένη ως προς τη μέγιστη συχνότητα για λέξεις του ιδίου μήκους,

$$x_3 = [MaxWordFreq(|W_{ij}|) - WordFreq(W_{ij})]/MaxWordFreq(|W_{ij}|)$$

όπου $|W_{ij}|$ συμβολίζει το μήκος της W_{ij}.

δ) την από κοινού συχνότητα εμφάνισης της προηγούμενης λέξης PW, της κατάτμησης, με την W_{ij}, κανονικοποιημένη ως προς τη μέγιστη από κοινού συχνότητα εμφάνισης ζεύγους λέξεων με πρώτη λέξη την PW,

$$x_4 = [\text{MaxWordPairFreq(PW)} - \text{WordPairFreq(PW, } W_{ij})]/\text{MaxWordPairFreq(PW)}$$

ε) το κανονικοποιημένο, ως προς το μέγιστο μήκος λέξης, απόλυτο μήκος της W_{ij}

$$x_5 = [MaxWordLength - WordLength(W_{ij})]/MaxWordLength$$

στ) την κανονικοποιημένη απόλυτη τιμή της διαφοράς του μήκους της W_{ij} από το μήκος της S_i,

$$x_6 = |\text{WordLength}(S_i) - \text{WordLength}(W_{ij})|/\max\{\text{WordLength}(S_i), \text{WordLength}(W_{ij})\}$$

όπου ο παρονομαστής επιλέγει το μεγαλύτερο μήκος μέσω της max{ }.

Η κανονικοποίηση εξασφαλίζει ότι τα x_i, i = 1, …, 6, είναι στο διάστημα [0, 1), με το 0 να αντιστοιχεί στη μέγιστη ομοιότητα μεταξύ των S_i και W_{ij}, εκτός του κριτηρίου (ε) που έχει σχεδιαστεί για να δίνει προτεραιότητα στις μεγαλύτερου μήκους λέξεις. Η συνολική καταλληλότητα Z δίνεται από τη σχέση:

$$Z = \sqrt{\frac{w_1 z_1^2 + w_2 z_2^2 + w_3 z_3^2 + w_4 z_4^2 + w_5 z_5^2 + w_6 z_6^2}{w_1 + w_2 + w_3 + w_4 + w_5 + w_6}}$$

όπου τα z_i, i = 1, …, 6, είναι οι επιμέρους βαθμοί καταλληλότητας και w_i, είναι οι σχετικές βαρύτητες κατά τον συνδυασμό τους. Οι επιμέρους βαθμοί καταλληλότητας είναι πραγματικοί αριθμοί στο διάστημα [0, 1) (συνεπώς και ο Z θα είναι ένας πραγματικός αριθμός στο [0, 1)) και δίνονται από τις συναρτήσεις κανονικής κατανομής:

$$z_i = e^{-x_i^2/2\sigma_i^2} \quad ; \qquad \forall i = 1, 2, \ldots, 6$$

με σ_i τις τυπικές τους αποκλίσεις. Η τυπική απόκλιση παρέχει ένα μέτρο της σχετικής απόστασης της αποθηκευμένης λέξης W_{ij} από την είσοδο S_i, ως προς κάποιο επιμέρους κριτήριο, έτσι ώστε η επίδρασή του στο ολικό κριτήριο καταλληλότητας να είναι σημαντική. Επειδή δε όλα τα x_i είναι κανονικοποιημένα στο ίδιο διάστημα, οι τυπικές αποκλίσεις μπορούν να επιλεγούν και ίσες με μια κοινή τυπική απόκλιση σ. Λέξεις που απέχουν πολύ από την είσοδο ως προς κάποιο κριτήριο θα έχουν $z_i \to 0$, ενώ αυτές που μοιάζουν με την είσοδο θα έχουν $z_i \to 1$. Η συμπεριφορά αυτή οφείλεται στην μορφή της κανονικής συνάρτησης.

Τέλος, στην τρίτη φάση, ελέγχεται αν η προτεινόμενη λέξη ικανοποιεί ορισμένα εμπειρικά κριτήρια όπως, οι συχνότητες εμφάνισης (τόσο της προτεινόμενης λέξης όσο και σε συνδυασμό με την προηγούμενη λέξη) και το συνολικό πλήθος χαρακτήρων της προηγούμενης και της προτεινόμενης λέξης, να είναι μεγαλύτερα από κάποια κατώφλια. Στην περίπτωση που τα κριτήρια ικανοποιούνται, η λέξη θεωρείται "καλή" υποψήφια για την κατάτμηση. Ειδάλλως, θα θεωρείται ως "κακή" υποψήφια λέξη. Αν στην ομάδα των Κ τελευταίων λέξεων το ποσοστό των καλών λέξεων ξεπερνά κάποιο κατώφλι, τότε η κατάτμηση θεωρείται επιτυχής και ο δρομέας

μετακινείται στη θέση του πρώτου χαρακτήρα μετά την νέα λέξη της κατάτμησης. Στην αντίθετη περίπτωση, η κατάτμηση θεωρείται ανεπιτυχής και ο δρομέας οπισθοδρομεί στην πρώτη από τις K τελευταίες λέξεις που έχει χαρακτηρισθεί κακή ώστε να επιλεγεί άλλη προτεινόμενη λέξη. Αν βρεθεί νέα λέξη τότε αντικαθιστά την παλιά και η κατάτμηση συνεχίζει στο υπόλοιπο κείμενο. Αν δεν βρεθεί κάποια άλλη κατάλληλη λέξη τότε η παλιά λέξη παραμένει ως είχε και ο δρομέας τοποθετείται στην επόμενη κακή λέξη, κ.ο.κ.

4. Πειραματικά αποτελέσματα

Στην πιλοτική εφαρμογή κατάτμησης νεοελληνικών κειμένων, χρησιμοποιήθηκε ένα λεξικό 112.575 ελληνικών λέξεων. Στο λεξικό αυτό προστέθηκαν και 6.028 νέες λέξεις από μια μεγάλη συλλογή 70 τευχών της διμηνιαίας εφημερίδας *Το Καποδιστριακό* του Πανεπιστημίου Αθηνών που χρησιμοποιήθηκε για την εξαγωγή των στατιστικών της γλώσσας. Χάριν ευκολίας, τα 70 τεύχη συνενώθηκαν σε ένα μεγάλο αρχείο με το τελικό κείμενο να περιλαμβάνει 2.412.210 λέξεις.

Το παραπάνω κείμενο υπέστη μια προεπεξεργασία προκειμένου να απαλειφθούν: α) όλες οι αγγλικές λέξεις, β) όλα τα ακρωνύμια, και γ) όλα τα κύρια ονόματα. Στη συνέχεια απαλείφθηκαν όλοι οι τόνοι και μετατράπηκαν όλα τα κεφαλαία σε μικρά γράμματα. Τέλος, όλες οι νέες λέξεις του κειμένου προστέθηκαν στο λεξικό έτσι ώστε το τελικό μέγεθος του λεξικού ήταν 118.603 λέξεων.

Προκειμένου να ικανοποιηθεί το βασικό κριτήριο κωδικοποίησης και επειδή ένα απλό λάθος χαρακτήρα μπορεί να επηρεάσει έως και 2ν ν-γράμματα, για να αποφύγουμε μεγάλες αλλοιώσεις στα διανύσματα ν-γραμμάτων κυρίως των λέξεων με λίγα γράμματα, αποφασίσθηκε ο διαχωρισμός του λεξικού σε τρία μικρότερα λεξικά ανάλογα με το μήκος των λέξεων. Το πρώτο λεξικό περιελάμβανε λέξεις μικρού μήκους (από 1 έως και 4 γράμματα), το δεύτερο περιελάμβανε λέξεις μεσαίου μήκους (από 5 έως και 8 γράμματα), και το τρίτο περιελάμβανε λέξεις μεγάλου μήκους (από 9 και πλέον γράμματα). Κάθε ένα από τα παραπάνω τρία λεξικά αποθηκεύθηκε σε συσχετιστική μνήμη με δομή IAA. Η κωδικοποίηση των λέξεων έγινε με μονογράμματα, διγράμματα και τριγράμματα για τις μικρές, μεσαίες και μεγάλες λέξεις αντίστοιχα.

Όπως προαναφέρθηκε, οι χρόνοι για τη δημιουργία των τριών συσχετιστικών μνημών και ο μέσος χρόνος ανάκλησης από τις συσχετιστικές μνήμες ήταν της τάξης των 5 sec και 15 msec αντίστοιχα. Ειδικά για την περίπτωση που η λέξη-κλειδί είναι στο όριο των κατηγοριών (π.χ. 4, 5, 8 ή 9 γραμμάτων), ο μέσος χρόνος περιλαμβάνει και την περίπτωση αναζήτησης και στο αμέσως επόμενο ή προηγούμενο λεξικό καθώς ένα λάθος εισαγωγής ή διαγραφής χαρακτήρα αλλοιώνει το μήκος της λέξης.

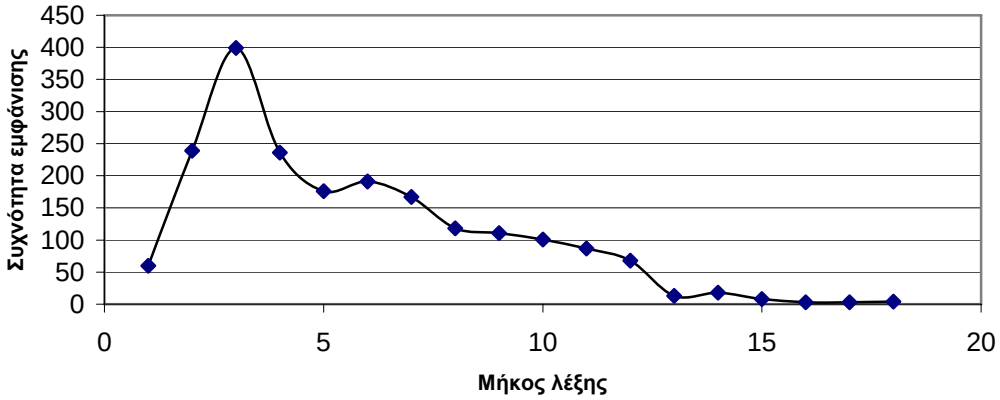
Η υπόθεση που έγινε σχετικά με το υποθετικό OCR σύστημα είναι ότι τα λάθη αναγνώρισης χαρακτήρων είναι ισοπίθανα, κάτι που βέβαια δεν ισχύει σε ένα πραγματικό σύστημα. Η μόνη διαφορά είναι ότι κατά τον υπολογισμό της απόστασης Levenshtein θα πρέπει κανείς να λάβει υπόψη του το κόστος που αναφέρεται στην κάθε στοιχειώδη ορθογραφική διόρθωση. Στην παρούσα εργασία, το κόστος θεωρείται το ίδιο για κάθε στοιχειώδη διόρθωση.

Τα στατιστικά στοιχεία της γλώσσας εξήχθησαν από το κείμενο των 2.412.210 λέξεων. Αυτά χρησιμοποιήθηκαν τόσο για την αποτίμηση της καταλληλότητας των προτεινόμενων λέξεων όσο και για τον έλεγχο των κριτηρίων οπισθοδρόμησης (βλ. 2^{η} και 3^{η} φάση αντίστοιχα, της ενότητας 3.3) κατά τη διαδικασία της κατάτμησης.

Η ανάκληση από τις συσχετιστικές μνήμες επιστρέφει έναν κατάλογο το πολύ 10 λέξεων που έχουν τουλάχιστον κατά 30% κοινά ν-γράμματα με τη λέξη-κλειδί. Προκειμένου να επιλεγεί μια λέξη από τον κατάλογο, όλες οι λέξεις πρέπει να αποτιμηθούν και να ταξινομηθούν. Η αποτίμηση των λέξεων γίνεται σύμφωνα με τη καταλληλότητα Z της ενότητας 3.3 όπου οι τυπικές αποκλίσεις των επιμέρους κριτηρίων καταλληλότητας επιλέχθηκαν ως $\sigma_i = 1$, (i = 1, 2, …, 6). Οι βαρύτητες επιλέχθηκαν εμπειρικά (μετά από πλήθος δοκιμών) ως εξής: $w_1 = 8$, $w_2 = 10$, $w_3 = 3$, $w_4 = 6$, $w_5 =$

5, w_6 = 5. Συνεπώς, παρατηρούμε ότι τα κριτήρια με την μεγαλύτερη επίδραση στην ολική αποτίμηση Z ήταν η συχνότητα εμφάνισης των λέξεων και η απόσταση Levenshtein. Τέλος, τα κατώφλια για τα τρία εμπειρικά κριτήρια ελέγχου της οπισθοδρόμησης, επιλέχθηκαν ως εξής: η συχνότητα εμφάνισης της προτεινόμενης λέξης να είναι τουλάχιστον 7.8%, η συχνότητα εμφάνισης του ζεύγους <προτεινόμενη λέξη, προηγούμενη λέξη> να είναι τουλάχιστον 14.7% και, τέλος, το συνολικό μήκος του ζεύγους να είναι τουλάχιστον 8 χαρακτήρων.

Το δοκιμαστικό κείμενο που επιλέχθηκε για κατάτμηση προέρχεται από ένα τεύχος της παραπάνω συλλογής 2002 λέξεων με κατανομή των λέξεων ως προς το μήκος τους σύμφωνα με το Σχ.4 (άνω των 900 λέξεων έχουν μήκος έως 4 γράμματα). Αφού υπέστη προεπεξεργασία για την απαλοιφή των κενών χαρακτήρων, των σημείων στίξης και των τόνων και για την μετατροπή των κεφαλαίων σε μικρά, στη συνέχεια υπέστη αλλοιώσεις σε επίπεδο χαρακτήρων ώστε να προσομοιάσει ένα υποθετικό σύστημα OCR. Θεωρώντας ότι το OCR σύστημα είχε την ίδια πιθανότητα αναγνώρισης για κάθε χαρακτήρα καθώς και την ίδια πιθανότητα για κάθε είδος λάθους, παρήχθησαν κείμενα συνεχούς γραφής με 0%, 0.6%, 1.2% και 2.4% λάθη (σε επίπεδο χαρακτήρα). Μετά τη διαδικασία κατάτμησης, τα ποσοστά αναγνώρισης των λέξεων για κάθε ένα από τα τέσσερα πιλοτικά κείμενα είναι 93.6%, 89.4%, 86.9% και 80.5% αντίστοιχα.



Σχ. 4. Η κατανομή των λέξεων του δοκιμαστικού κειμένου ως προς το μήκος τους.

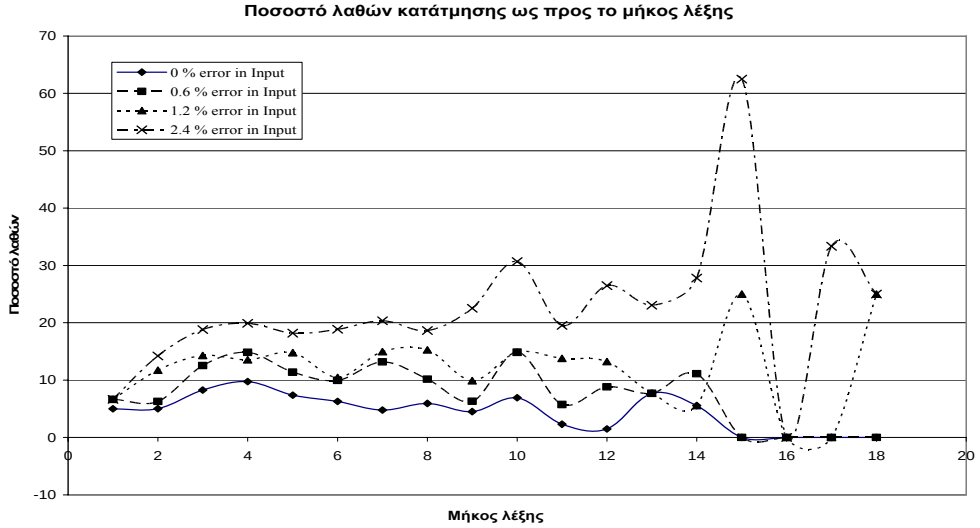
Τα αποτελέσματα της κατάτμησης για τα τέσσερα πιλοτικά κείμενα παρουσιάζονται στο Σχ. 5. Αν e(n) συμβολίζει το πλήθος των λέξεων μήκους n με λάθη κατάτμησης και t(n) συμβολίζει το συνολικό πλήθος των λέξεων μήκους n, τότε στο μεν Σχ. 5α δίνονται τα ποσοστά λανθασμένων λέξεων ανά μήκος λέξης p(n) = [e(n)/t(n)]x100% ενώ στο Σχ. 5β δίνεται η κατανομή των λαθών ανάλογα με το μήκος των λέξεων. Πιο αναλυτικά, από το Σχ. 5α παρατηρούμε ότι τα ποσοστά λαθών κατά την κατάτμηση περίπου τριπλασιάζονται όταν υπεισέλθει αλλοίωση μόλις 2.4% των χαρακτήρων. Τούτο οφείλεται στο ότι σε επίπεδο λέξεων τα λάθη θα είναι της τάξης του 15% για ένα μέσο μήκος λέξεων περίπου 6 χαρακτήρων.

Από το Σχ. 5α είναι φανερό ότι τόσο οι μικρού όσο και οι μεσαίου και μεγάλου μήκους λέξεις είναι αρκετά ευάλωτες (ποσοστιαία) στα λάθη. Επειδή, όμως, οι μικρές λέξεις αποτελούν συνήθως και την πλειονότητα σε ένα κείμενο (βλ. Σχ. 4) πιστεύουμε ότι το μεγαλύτερο βάρος σε μελλοντική έρευνα πρέπει να δοθεί στην βελτίωση και λεπτομερειακή ρύθμιση των παραμέτρων του συστήματος κατάτμησης, κυρίως ως προς τις μικρές λέξεις.

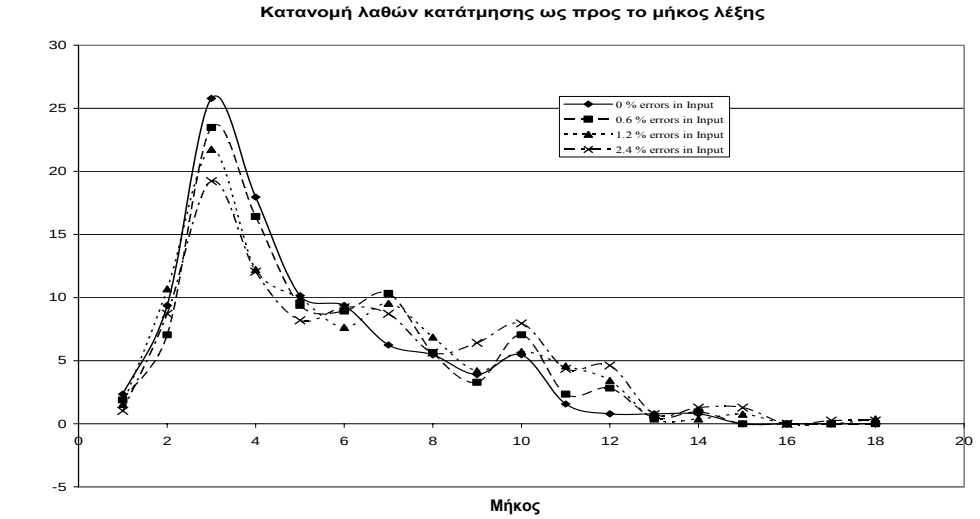
Τέλος, στο Σχ. 6 παρουσιάζεται ένα δείγμα του αρχικού συνεχούς κειμένου χωρίς λάθη χαρακτήρων (στα πλαίσια εμφανίζονται οι λέξεις που δεν αναγνωρίσθηκαν) παράλληλα με το

αποτέλεσμα της κατάτμησης (τα λάθη εμφανίζονται υπογραμμισμένα). Συνενώσεις μικρότερων λέξεων (π.χ. «αυτοπαρουσιάζονται») ή διαχωρισμός μεγάλων σε μικρότερες λέξεις (π.χ. «στοπ αν επι στη μι ο») είναι συνηθισμένες αιτίες σφαλμάτων. Αποφυγή αυτού του είδους των σφαλμάτων θα μπορούσε να γίνει μέσω συντακτικής ανάλυσης.

(α)



(β)



Σχ.5. Αποτελέσματα κατάτμησης για 4 κείμενα εισόδου με διαφορετικά ποσοστά αλλοιωμένων χαρακτήρων.

Συνεχές κείμενο χωρίς OCR λάθη	Η έξοδος του συστήματος κατάτμησης
…στοαρθρο αυτοπαρουσιαζονται πληροφοριες για τις ιδιοτητες των ακτινοβολιων με ιδιαιτερη εμφαση στην κινητη τηλεφωνια οπου	στο αρθρο **αυτοπαρουσιαζονται** πληροφοριες για τις ιδιοτητες των ακτινοβολιων με ιδιαιτερη εμφαση στην κινητη τηλεφωνια οπου

εστωνακτινοβολιωνμειδιαιτερημεφασηστηνκινητητηλεφωνιαοπ ουπαρουσιαζεταιμεμετρησειςηεντασητηςακτινοβολιαςαποτισκερ αιεςβασηςκαιαποτακινητατηλεφωναδιατυπωνονταιπροτασειςγια τηντοποθετησητωνκεραιωνβασηςωστεναευρισκονταισεαπολυτη ασφαλειαοικατοικοιπουζουνγυρωαποτισκεραιεςκαιτελοςδινοντα ιοδηγιεςγιαασφαληχρησητωνκινητωντηλεφωνωνμετηχρησηειδι κηςπροστασιαςτηνοποιαεχουμεεπινοησειοιπληροφοριεςπουπαρ ουσιαζονταιεχουνπροκυψειαποτισδραστηριοτητεςτηςερευνητικη ςμαςομαδαςστοντομεαβιολογιαςκυτταρουκαιτουτμηματοςβιολο γιαςστοπανεπιστημιοαθηνωνηοποιαασχολειταιερευνητικακαιεκ παιδευτικαεπισειρανετωνμετιςεπιπτωσειςτηςμηακτινοβολιαςστ ηζωντανηυληπωςλειτουργειτοσυστημακινητηςτηλεφωνιαςσπανι ααναρωτιεταιοκατοχοςκινητουτηλεφωνουπωςγινεταικαιμπορειν αεπικοινωνειαποοπουδηποτεστηνκυριολεξιααποθεωρητικηςπλε υραςφυσικαισωςνομιζειοτιτοκινητολειτουργειοπωςεναραδιοφω νοπουπραγματιμπορειναπιασεισταθμουςσχεδονπαντουαυτοειναι σωστοκατατοημισυγιατιτοκινητοειναιπαραλληλακαιακτινοβολι α…	παρουσιαζεται με μετρησεις η ενταση της ακτινοβολιας απο τις κεραιες βασης και απο τα κινητα τηλεφωνα διατυπωνονται προτασεις για την τοποθετηση των κεραιων βασης ωστε να ευρισκονται σε απολυτη ασφαλεια οι κατοικοι που ζουν γυρω απο τις κεραιες και τελος δινονται οδηγιες για ασφαλη χρηση των κινητων τηλεφωνων με τη χρηση ειδικης προστασιας την οποια εχουμε επινοησει οι πληροφοριες που παρουσιαζονται εχουν προκυψει απο τις δραστηριοτητες της ερευνητικης μας ομαδας στον τομεα βιολογιας κυτταρου και του τμηματος βιολογιας **στοπ αν επι στη μι ο** αθηνων η οποια ασχολειται ερευνητικα και εκπαιδευτικα επι **σειρα νετων** με τις επιπτωσεις της μη ακτινοβολιας στη ζωντανη υλη πως λειτουργει το συστημα κινητης τηλεφωνιας σπανια αναρωτιεται ο κατοχος κινητου τηλεφωνου πως γινεται και μπορει να επικοινωνει απο οπουδηποτε στην κυριολεξια απο θεωρητικης πλευρας **φυσικαι σωων ο μι ζει** οτι το κινητο λειτουργει οπως ενα ραδιοφωνο που πραγματι μπορει να πιασει σταθμους σχεδον παντου αυτο ειναι σωστο κατα το ημισυ γιατι το κινητο ειναι παραλληλα και ακτινοβολια

Σχ.6. Παράδειγμα κατάτμησης συνεχούς κειμένου χωρίς OCR λάθη.

5. Συμπεράσματα – Μελλοντική Έρευνα

Το σύστημα κατάτμησης που παρουσιάσθηκε στην εργασία αυτή αναγνωρίζει και διαχωρίζει με αρκετά μεγάλη επιτυχία λέξεις σε συνεχές κείμενο ακόμη και στην περίπτωση λαθών στους χαρακτήρες του κειμένου. Για τον σκοπό αυτό χρησιμοποιείται ένα συσχετιστικό λεξικό για ταχύτατη προσεγγιστική ανάκληση, αποτίμηση της καταλληλότητας των προτεινόμενων λέξεων και δυνατότητες οπισθοδρόμησης όταν δεν ικανοποιούνται ορισμένα εμπειρικά κριτήρια κατάτμησης. Η εφαρμογή της παραπάνω μεθοδολογίας σε πιλοτικά συνεχή κείμενα, με ή χωρίς λάθη χαρακτήρων, έδωσε ικανοποιητικά αποτελέσματα κατάτμησης.

Από το Σχ. 5β, παρατηρούμε ότι οι μικρές λέξεις (έως 4 γραμμάτων) δυσκολεύουν περισσότερο την κατάτμηση. Συνεπώς, περαιτέρω βελτίωση και καλύτερη ρύθμιση των παραμέτρων του συστήματος κατάτμησης, κυρίως ως προς τις μικρές λέξεις, αναμένεται να βελτιώσει σημαντικά τα συνολικά ποσοστά αναγνώρισης των λέξεων. Άλλωστε, αυτός θα είναι και ο στόχος κατά την εφαρμογή της προτεινόμενης μεθοδολογίας στην κατάτμηση πατερικών κειμένων συνεχούς μικρογράμματης γραφής της Μονής Αγίας Αικατερίνης του Όρους Σινά.

Σημειώσεις

1. R. Sproat and C. Shih, “A Statistical Method for Finding Word Boundaries in Chinese Text”, *Comp. Proc. Chinese and Oriental Lang.* 4 (1990) 336-351.
2. J. M. Ponte and W. B. Croft, “Useg: A Retargetable Word Segmentation Procedure for Information Retrieval”, *Symp. Doc. Anal. Inform. Retrieval* (SDAIR’96), 1996.
3. B. Gatos, I. Pratikakis and S. Perantonis, “An Adaptive Binarisation Technique for Low Quality Historical Documents”, *IAPR W. on Doc. Anal. Sys.* (DAS’04), LNCS 3163, Florence, Italy, (2004) 102-113.
4. V. Cherkassky, N. Vassilas, G. Brodt and H. Wechsler, “Conventional and Associative Memory Approaches to Automatic Spelling Correction”, *Int’l J. Eng. Appl. of Artificial Intelligence* 5 no 3 (1992) 223-237.
5. J. Zobel and P. Dart, “Finding Approximate Matches in Large Lexicons”, *Software Practice and Experienc* 25 no 3 (1995) 331-345.
6. T. Kohonen, *Self-Organization and Associative Memories*, Springer, NY, 1984.