Ένα Υβριδικό Μοντέλο για την Δυναμική Κανονικοποίηση Ελληνικών Κειμένων

Γεώργιος Καρμπέρης, Γεώργιος Κουρουπέτρογλου και Γεράσιμος Ξύδας
Τμήμα Πληροφορικής και Τηλεπικοινωνιών
Πανεπιστήμιο Αθηνών
{grad0350, koupe, gxydas}@di.uoa.gr

Περίληψη: *Στην παρούσα εργασία παρουσιάζεται ένα νέο υβριδικό σύστημα κανονικοποίησης κειμένου, που μπορεί να χειρίζεται αποτελεσματικά αριθμητικές και συνεπτυγμένες γραπτές εκφράσεις (ΓΕ) σε κείμενα της ελληνικής γλώσσας. Το μοντέλο ανιχνεύει τις μη κανονικοποιημένες ΓΕ, τις ταξινομεί, αίρει τις αμφισημίες και παράγει τις γραμματικά και συντακτικά αντίστοιχες σωστές λεκτικές μορφές. Η μέθοδος βασίζεται σε συνδυασμό κανόνων, πινάκων αντιστοίχησης και εφαρμογής επεξεργασίας φυσικής γλώσσας (μορφολογική, συντακτική και σημασιολογική ανάλυση). Η εργασία αποτελεί μια προσπάθεια μεθοδολογικής γενικής επίλυσης του προβλήματος που μπορεί να εφαρμοστεί σε οποιοδήποτε κείμενο, όχι συγκεκριμένου θεματικού πεδίου, στηρίζεται σε μια μη-εποπτευόμενη προσέγγιση και μπορεί να επεκταθεί και σε άλλες περιπτώσεις μη-λεκτικών μορφών. Στο εργαστήριο Ομιλίας του Πανεπιστημίου Αθηνών σχεδιάστηκε μια εφαρμογή που υλοποιεί το μοντέλο αυτό σε λειτουργία πραγματικού χρόνου η οποία ενσωματώθηκε με επιτυχία στο σύστημα μετατροπής κειμένου σε ομιλία ΔΗΜΟΣΘΈΝΗΣ. Η αξιολόγηση του συστήματος με μια συλλογή κειμένων, αποτελούμενη από 20Κ λέξεις και 374 ΓΕ από το διαδίκτυο, καθώς και επιστημονικών εργασιών, έδειξε την αποτελεσματικότητα του μοντέλου, καθώς το ποσοστό ορθής κανονικοποίησης που επιτεύχθηκε ήταν 97,1%.*

Abstract*: In this work we present a new hybrid system for the normalization of texts that can handle with success numerical and alphabetical written expressions, known as Non-Standard Words (NSW) in the Greek language. The proposed model identifies NSWs, classifies and disambiguates them and produces their morphological and syntactical correct corresponding written forms. The method is based on a combination of rules, lookup tables and natural language processing (morphological, syntactic and*

semantic). This work is an effort for providing a methodological generic solution to the problem of normalization that can be applied in any text, not domain specific. It is based on an unsupervised approach and can be extended to other cases of NSWs. In the Speech Group of the University of Athens, we developed an application that implements this model in real-time conditions. This application was further incorporated as a module in DEMOSTHeNES Speech Composer. The system was tested using a 20K-word corpus and 374 sentences containing NSWs that was selected from the Web and from scientific works. The evaluation showed the efficiency of the model, as the success rate of correct normalization we achieved was 97.1%.

Λέξεις-Κλειδιά*:* κανονικοποίηση κειμένου, κανονικοποίηση μη κανονικών λέξεων, ανάπτυγμα συντμήσεων, ακρώνυμων και αριθμητικών, μετατροπή κειμένου σε ομιλία

1. Εισαγωγή

Πέρα από τις κοινές λέξεις και τα ονόματα τα πραγματικά κείμενα περιέχουν γραπτές εκφράσεις (ΓΕ) με μη κανονικές λέξεις που γενικά μπορούν να διακριθούν σε τρεις κατηγορίες: α) αλφαβητικές εκφράσεις (βραχυγραφίες, συμπτύξεις, ακρώνυμα, κ.ά), β) αριθμητικές εκφράσεις (κοινοί αριθμητικά, τακτικοί αριθμοί, ημερομηνίες, ώρες, κ.ά) σε αραβική, ελληνική και λατινική συμβολογραφία και γ) άλλες ΓΕ (ηλεκτρονικές διευθύνσεις, μικτές αλφαριθμητικές, κ.ά). Η κανονικοποίηση κειμένων έχει σκοπό να αντικαταστήσει τις παραπάνω ΓΕ με την κατάλληλη αλληλουχία συναφών εννοιολογικά λέξεων που περιέχονται σε ένα κοινό λεξικό. Το πόσο συχνά απαντώνται αριθμητικές και συνεπτυγμένες ΓΕ εξαρτάται από τον τύπο του κειμένου. Για την αγγλική γλώσσα, η συχνότητα εμφάνισης αριθμητικών και συνεπτυγμένων ΓΕ είναι: στα μυθιστορήματα 1.5%, στον τύπο 4.9%, στην ηλεκτρονική αλληλογραφία 10.7%, στις συνταγές 13.7%, στα διαδικτυακά chats 20.1% και στα εμπιστευτικά έγγραφα 27.9% [Black et al, 1999]. Στο σύνολο των μη κανονικοποιημένων περιπτώσεων ΓΕ 44% είναι αλφαβητικές εκφράσεις και 36% αριθμητικές [Black et al. 1999]. Η κανονικοποίηση των κειμένων βρίσκει εφαρμογή κυρίως σε εφαρμογές μετατροπής κειμένου σε ομιλία (ΜΚΟ), αναγνώρισης ομιλίας και ανάκτησης πληροφοριών. Τα περισσότερα από αυτά τα συστήματα χρησιμοποιούν κανονικοποίηση με περιορισμένες

δυνατότητες. Οι υπάρχουσες λύσεις δεν είναι γενικού σκοπού και συνήθως δεν αίρουν το κύριο πρόβλημα της αμφισημίας. Για παράδειγμα, πολλά συστήματα ΜΚΟ χρησιμοποιούν μονοσήμαντη αντιστοιχία αριθμητικών μορφών με λεκτικές αναπαραστάσεις (π.χ. «14 σελίδες» → «δέκα τέσσερα σελίδες») ή αδυνατούν σε πολλές περιπτώσεις να παράγουν το πλήρες κείμενο μιας σύντμησης στον σωστό αριθμό και στο σωστό γένος.

Το πρόβλημα της κλίσης των αριθμητικών και του αναπτύγματος συντμήσεων σε ένα κείμενο είναι αρκετά σύνθετο, καθώς για την κανονικοποίηση πρέπει να γίνει σωστός εντοπισμός, αναγνώρισή τους, κατηγοριοποίηση, αντιμετώπιση δισημιών στις λέξεις του κειμένου, ομοιογραφημάτων και παραγωγή του μορφολογικά κατάλληλου αναπτύγματος.

Για την κανονικοποίηση κειμένων έχουν γίνει διάφορες εργασίες κυρίως για συγκεκριμένες γλώσσες. Για την Γερμανική γλώσσα έχουν αναπτυχθεί δύο διαφορετικά συστήματα, το ένα από τα Bell Labs [Mobius et al., 1997] ενώ το δεύτερο είναι το σύστημα FELIX [Fries and Wirth, 1997]. Το FELIX πραγματοποιεί συντακτική ανάλυση χρησιμοποιώντας τον αλγόριθμο του Zingle [Zingle, 1982] και η διατύπωση της πρότασης διευκρινίζεται από το μέρους του λόγου (POS). Στο σύστημα των Bell Labs, η αναπαράσταση της πληροφορίας στην βάση δεδομένων της εφαρμογής αυτής είναι παρόμοια με αυτήν στο FELIX. Ένα πλεονέκτημα είναι ότι γίνεται προσπάθεια να επιτευχθεί αποσαφήνιση αμφισημιών με χρήση κανόνων. Ο περιορισμός που τίθεται όμως είναι ότι δεν υπάρχει συντακτική ανάλυση.

Στην Ιαπωνική γλώσσα παρουσιάζονται παρόμοια προβλήματα με την ελληνική, καθώς υπάρχουν λέξεις όπου προφέρονται με διαφορετικό τρόπο ανάλογα με την σημασία που αποκτούν στην πρόταση [Ooyama et al., 1987]. Στην τεχνική της εργασίας αυτής χρησιμοποιείται βαθμολόγηση κάθε πιθανής σημασίας για μια λέξη της πρότασης, και μετά από ανάλυση, επιλέγεται η σημασία με την μεγαλύτερη βαθμολογία εξάρτησης. Έτσι, επιλέγεται η κατάλληλη προφορά της λέξης και το κείμενο κανονικοποιείται ανάλογα. Πιο συγκεκριμένα, για τα αριθμητικά, χρησιμοποιούνται πρότυπα με βάση τις γειτονικές λέξεις του αριθμητικού στην πρόταση.

Στην Αγγλική γλώσσα, μια τεχνική για κανονικοποίηση κειμένου που αναπτύχθηκε κάνει χρήση προ-επεξεργαστή, ο οποίος εκμεταλλεύεται την συντακτική ανάλυση της πρότασης και ανάλογα αποφασίζει για το είδος κάθε αριθμητικού και την

εξαγωγή του κειμένου από συντμήσεις [Coughlin, 1999]. Στην υλοποίηση αυτή καλύπτεται και αποσαφήνιση συντμήσεων όπου υπάρχουν παραπάνω του ενός πιθανά παράγωγα. Θα πρέπει να σημειωθεί ότι στην επεξεργασία αυτή δεν υπάρχουν κλίσεις των λέξεων, ενώ υποστηρίζεται ότι η τεχνική μπορεί να εφαρμοστεί και σε άλλες γλώσσες. Μια ολοκληρωμένη προσέγγιση του προβλήματος για την Αγγλική γλώσσα παρουσιάζεται στην [Sproat et al., 2001], ενώ στην [Toole, 2000] το πρόβλημα αντιμετωπίζεται μόνο για συντομογραφίες για εφαρμογές ανάκτησης πληροφοριών. Η [Mikheev, 2000] ασχολείται μόνο με το πρόβλημα της αμφισημίας κατά την κανονικοπίηση κειμένου.

Στην παρούσα εργασία γίνεται προσπάθεια μεθοδολογικής γενικής επίλυσης του προβλήματος της κανονικοποίησης αριθμητικών και συνεπτυγμένων ΓΕ που μπορεί να εφαρμοστεί σε οποιοδήποτε κείμενο, όχι συγκεκριμένου θεματικού πεδίου. Το προτεινόμενο μοντέλο στηρίζεται σε μια μη-εποπτευόμενη προσέγγιση που διαδοχικά: ανιχνεύει τις μη κανονικοποιημένες ΓΕ σε ένα κείμενο, τις κατηγοριοποιεί, αίρει τις αμφισημίες και παράγει δυναμικά τις γραμματικά και συντακτικά σωστές λεκτικές μορφές τους, σχηματίζοντας έτσι το κανονικοποιημένο κείμενο. Η μέθοδος βασίζεται σε συνδυασμό κανόνων, πινάκων αντιστοίχησης και εφαρμογής επεξεργασίας φυσικής γλώσσας (μορφολογική, συντακτική και σημασιολογική ανάλυση)**.**

2. Γλωσσικοί Πόροι

Το μοντέλο υποστηρίζεται από ένα σύνολο γλωσσικών πόρων για την γραμματική, μορφολογική και συντακτική ανάλυση των κειμένων.

Μορφολογικό λεξικό: Το λεξικό περιλαμβάνει όλες τις μορφολογικές πληροφορίες που αφορούν τις λέξεις, ένα κωδικό κλίσης των κλιτών λέξεων, που αντιστοιχεί σε ένα πρότυπο καταλήξεων, και το θέμα της λέξης. Με βάση αυτές τις πληροφορίες, είναι δυνατή η δυναμική γένεση όλων των κλίσεων μιάς λέξης, κάτι απαραίτητο για την ανάπτυξη συνεπτυγμένων ΓΕ.. Χρησιμοποιούνται συνολικά 53 διαφορετικά πρότυπα καταλήξεων για όλα τα ουσιαστικά της ελληνικής και 17 διαφορετικά για όλα τα επίθετα και τις μετοχές. Επίσης, το σύστημα χρησιμοποιεί όλα τα δυνατά πρότυπα καταλήξεων για όλους τους κωδικούς κλίσης των ουσιαστικών, των επιθέτων, των άρθρων και των αριθμητικών. Με αυτόν τον τρόπο, γνωρίζοντας για μια λέξη τον

κωδικό κλίσης της και το θέμα της, η λέξη αυτή μπορεί να κλιθεί σε όλα τα δυνατά πρόσωπα και σε όλες τις πτώσεις.

Συντακικός αναλυτής: Το σύστημα περιλαμβάνει δεδομένα για τα συντακτικά πρότυπα της ελληνικής, όπου περιέχονται όλες οι συντακτικές πληροφορίες που αφορούν τον προσδιορισμό του υποκειμένου και του αντικειμένου ή του κατηγορούμενου μιας πρότασης. Σε τρεις διαφορετικές περιοχές, μια για κάθε συντακτικό τύπο, είναι αποθηκευμένες οι ακολουθίες των χαρακτηρισμών που ελέγχονται για εντοπισμό στην πρόταση. Στον πίνακα Ι δίνονται ορισμένες από τις ακολουθίες χαρακτηρισμών για το υποκείμενο. Για το αντικείμενο και το κατηγορούμενο, οι ακολουθίες είναι κωδικοποιημένες με παρόμοιο τρόπο. Αρκετές ακολουθίες χαρακτηρισμών που αφορούν το αντικείμενο είναι κοινές και για το υποκείμενο, αλλά όχι όλες. Υπάρχουν ακολουθίες χαρακτηρισμών που μπορούν να αντιστοιχηθούν μόνο σε αντικείμενο και όχι σε υποκείμενο, καθώς και το ανάποδο.

Ipokeimeno
ΑΡ+ΟΥΣ+ΑΝΤ+ΣΥΝ+ΑΝΤ
ΑΡ+ΟΥΣ+ΣΥΝ+ΑΝΤ
ΑΝΤ+ΣΥΝ+ΑΡ+ΟΥΣ+ΑΝΤ
ΑΝΤ+ΣΥΝ+ΑΡ+ΟΥΣ
ΑΡ+ΕΠΙΘ
ΑΝΤ+ΣΥΝ+ΑΝΤ
ΑΡ+ΟΥΣ+ΑΝΤ
…

Πίνακας Ι Παραδείγματα από τις ακολουθίες χαρακτηρισμών για το υποκείμενο.

Λεξικό ΓΕ: Περιέχεται το πλήρες ανάπτυγμα 838 κοινών συντμήσεων, ακρωνυμιών και αρκτικόλεξων που απαντάμε στην ελληνική. Έτσι, το σύστημα μπορεί να αναπτύξει δυναμικά κάθε ΓΕ με βάση την γραμματική και την σύνταξη της πρότασης σε κάθε περίπτωση, ενώ ξεχωρίζονται οι περιπτώσεις που οι ΓΕ εμφανίζονται με κόμματα ή χωρίς.

Γεννήτρια Προφοράς Αριθμητικών: Για την σωστή προφορά των αριθμητικών μορφών, κρατούνται κανονικές εκφράσεις που χρησιμοποιούνται για τον εννοιολογικό προσδιορισμό τους σε μια πρόταση και οι οποίες μπορούν να ορίζονται από τον χρήστη. Για κάθε περίπτωση αριθμητικού, έχει αποθηκευθεί μια εγγραφή με 3 πεδία:

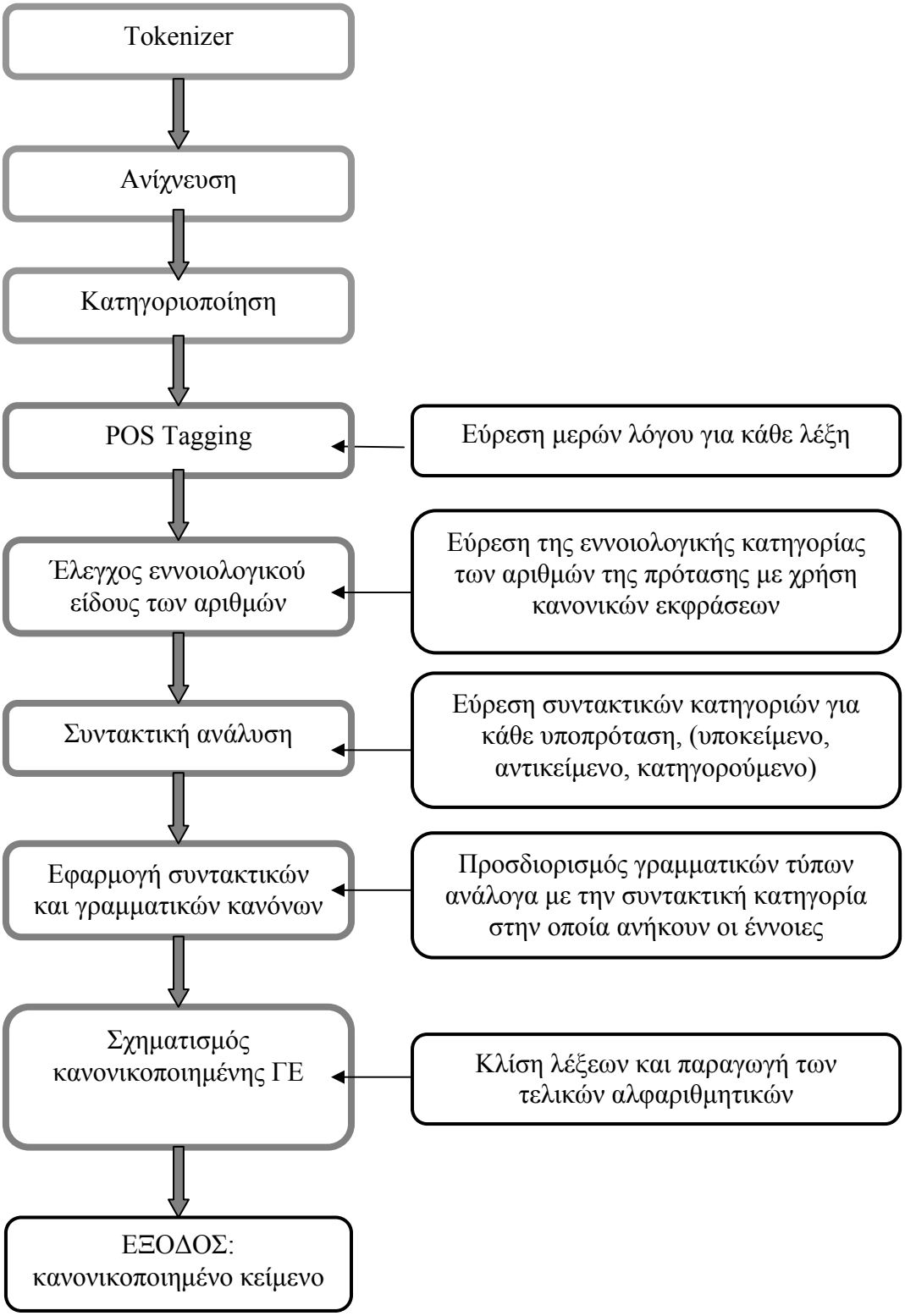
Την κανονική έκφραση για το αριθμητικό, το εννοιολογικό είδος του αριθμητικού και τον τρόπο προφοράς του αριθμητικού. Ο τρόπος προφοράς είναι κωδικοποιημένος με σύμβολα, που ερμηνεύονται από το σύστημα κατά την παραγωγή του κανονικοποιημένου κειμένου (Πίνακας ΙΙ). Όσα σύμβολα αντιστοιχούν σε ψηφία ("#" = 0-9) και είναι συνεχόμενα στο πεδίο της προφοράς κλίνονται σαν αριθμοί από το σύστημα με βάση το πεδίο [Προφορά], όπου για κάθε ψηφίο ισχύει: X= {ονομαστική, θηλυκό, ενικός}, Y= {γενική, αρσενικό, ενικός}, Z= {ονομαστική, ουδέτερο, ενικός}.

Κανονική έκφραση	**Τύπος**	**Προφορά**
##/#/####	Ημερομηνία	XX Y ZZZZ
#/#/####	Ημερομηνία	X Y ZZZZ
##/#	Ημερομηνία	XX Y
##.## μμ	Ώρα	XX ZZ μετά μεσημβρίας
#.# πμ	Ώρα	X Z προ μεσημβρίας
##.# πμ	Ώρα	XX Z προ μεσημβρίας
#.## πμ	Ώρα	X ZZ προ μεσημβρίας

Πίνακας ΙΙ Παραδείγματα ακολουθιών κανονικών εκφράσεων για τη προφορά αριθμητικών.

3. Το Υβριδικό Μοντέλο Κανονικοποίησης

Μια σχηματική αναπαράσταση της αρχιτεκτονικής του συστήματος παρουσιάζεται στο Σχήμα 1. Αποτελείται από τις παρακάτω διεργασίες, οι οποίες εκτελούνται διαδοχικά:



Σχήμα 1. Σχηματική αναπαράσταση της αρχιτεκτονικής του συστήματος

Εντοπισμός ορίων ΓΕ (tokenization): Τοποθετεί τα όρια μιας γραπτής έκφρασης με οποιαδήποτε σύμβολα που ορίζεται ως μια αλληλουχία χαρακτήρων χωρισμένων με διαστήματα. Αναγνωρίζει αν μια τελεία ή κόμμα αποτελούν μέρος ενός αριθμητικού ή μιας συντομογραφίας ή οριοθετούν το τέλος μιας φράσης ή πρότασης.

Ανίχνευση συντομογραφίας ή αριθμητικού: Εντοπίζεται στη ΓΕ αν υπάρχει συντομογραφία ή αριθμητικό οποιασδήποτε μορφής.

Κατηγοριοποίηση συντομογραφίας ή αριθμητικού: Προσδιορίζεται η κατηγορία: α) του αριθμητικού: κανονικός αριθμός, τακτικός αριθμός, τηλέφωνο, ώρα, ημερομηνία, έτος, χρήματα, επί τοις εκατό, ταχυδρομικός κωδικός, αριθμός δρόμου, ψηφία (π.χ. δωμάτιο 101) ή β) της συντομογραφίας: πλήρους ανάπτυξης, ακολουθία γραμμάτων (π.χ. ΦΠΑ), ως λέξη (π.χ. ΝΑΤΟ).

POS Tagging: Ανακτάται το μέρος του λόγου (POS) για κάθε λέξη από το μορφολογικό λεξικό του συστήματος.

Έλεγχος Εννοιολογικού Είδους Αριθμών: Ελέγχεται αν οι λέξεις της πρότασης που περιέχουν αριθμούς ανήκουν σε κάποια από τις εννοιολογικές κατηγορίες όπως έχουν οριστεί από τον χρήστη. Αν ναι, προσδιορίζεται και ο τρόπος προφοράς τους.

Συντακτική Ανάλυση της πρότασης: Η διεργασία αυτή εκτελείται μόνο όταν υπάρχουν αριθμητικά που δεν έχει προσδιοριστεί η εννοιολογία τους στην πρόταση. Προσδιορίζεται υποκείμενο, αντικείμενο και κατηγορούμενο στην πρόταση.

Εφαρμογή Γραμματικών και Συντακτικών κανόνων: Προσδιορίζονται οι γραμματικοί τύποι για κάθε λέξη που περιέχει αριθμούς και δεν έχει προσδιοριστεί ήδη η προφορά της κατά τον Εννοιολογικό Έλέγχο του αριθμού. Επίσης, προσδιορίζονται οι γραμματικοί τύποι και για τις ΓΕ της πρότασης, με βάση την σύνταξη της πρότασης. Χρησιμοποιείται μοντέλο κανόνων που μπορεί να ορίζεται είτε από τον χρήστη, (κάνοντας χρήση κανονικών εκφράσεων) είτε από την συντακτική ανάλυση της πρότασης. Σε όλες τις περιπτώσεις οι κανονικές εκφράσεις υπερισχύουν όλων των

κανόνων σύνταξης, οι οποίοι εφαρμόζονται μόνο σε περίπτωση που βρεθούν αριθμητικά που δεν έχει προσδιοριστεί η προφορά τους. Το τμήμα αυτό χρησιμοποιεί και σημασιολογία στο υποκείμενο της πρότασης, για την ανίχνευση αριθμητικού που περιέχει (π.χ. τηλεφωνικό αριθμό) και δεν έχει δηλωθεί η κανονική του έκφραση.

Σχηματισμός της Κανονικοποιημένης Ελληνικής Πρότασης: Στη τελευταία διεργασία, το σύστημα κλίνει τις λέξεις και παράγει το κανονικοποιημένο κείμενο για τα αριθμητικά, αλλά και το πλήρες ανάπτυγμα των ΓΕ, με βάση τους γραμματικούς τύπους για το καθένα, όπως αυτοί έχουν προσδιοριστεί προηγουμένως [Karberis and Kouroupetroglou, 2002].

Κατά την παραγωγή του κανονικοποιημένου αριθμητικού παράγονται αρχικά οι αντίστοιχες λέξεις του στο σωστό γένος, σε πτώση ονομαστική. Ο σχηματισμός αυτός επιτυγχάνεται με δυναμικό χωρισμό του αριθμού στα ψηφία του. Η διαδικασία έχει ως εξής: ξεκινώντας από δεξιά προς τα αριστερά, τα ψηφία του αριθμού χωρίζονται σε τριάδες:

- Η πρώτη τριάδα από δεξιά δεν αλλάζει, γιατί δεν απαιτείται επιπλέον κείμενο από το τη λεκτική απόδοση των ψηφίων της.
- Στην δεύτερη, προστίθεται δεξιά από το κείμενο των ψηφίων της το κείμενο "χιλιάδες"
- Στην τρίτη προστίθεται το κείμενο "εκατομμύρια".
- Στην τέταρτη προστίθεται το κείμενο "δισεκατομμύρια".
- Στην πέμπτη προστίθεται το κείμενο "τρισεκατομμύρια".
- Από την έκτη τριάδα και πέρα, ανάλογα με τον αριθμό της τριάδας, προστίθεται το κείμενο που παράγεται από την παρακάτω διαδικασία:
    - κείμενο του {αριθμού της τριάδας - 2}
    - αφαίρεση από αυτό του τελευταίου γράμματος
    - πρόσθεση του αλφαριθμητικού "άκις εκατομμύρια"

Για παράδειγμα, το επιπλέον κείμενο για την $7^{η}$ τριάδα από αριστερά, έχει ενδιάμεσο κανονικοποημένο κείμενο τα ψηφία της και επιπλέον:

Πράξη Διαδικασίας	**Ενδιάμεσο Κανονικοποιημένο Κείμενο**
κείμενο του {αριθμού της τριάδας - 2}	*Πέντε*
αφαίρεση από αυτό του τελευταίου γράμματος	*Πέντ*
πρόσθεση του αλφαριθμητικού "άκις εκατομμύρια"	*Πεντάκις εκατομμύρια*

Σε κάθε τριάδα, τα ψηφία μετατρέπονται σε κείμενο με βάση την θέση τους στην τριάδα, σύμφωνα με τον παρακάτω πίνακα:

Θέση ψηφίου στην τριάδα	**Ενδιάμεσο Κανονικοποιημένο Κείμενο**
Δεξιό	*Ένα, ... , Δέκα*
Μεσαίο	*Δέκα, ..., Ενενήντα*
Αριστερό	*Εκατό, ...Εννιακόσια*

Στη συνέχεια, η συνάρτηση παραγωγής της τελικής κλίσης καλείται για τις λέξεις του κειμένου αυτού που βρίσκονται στην πρώτη τριάδα από δεξιά, καθώς επίσης και για τις λέξεις
που αποτελούν επιπλέον κείμενο για τις τριάδες, (όπως χιλιάδες, εκατομμύρια, δισεκατομμύρια, κοκ). Τα ορίσματα της συνάρτησης σε κάθε κλήση της, είναι το γένος, η πτώση και ο αριθμός.

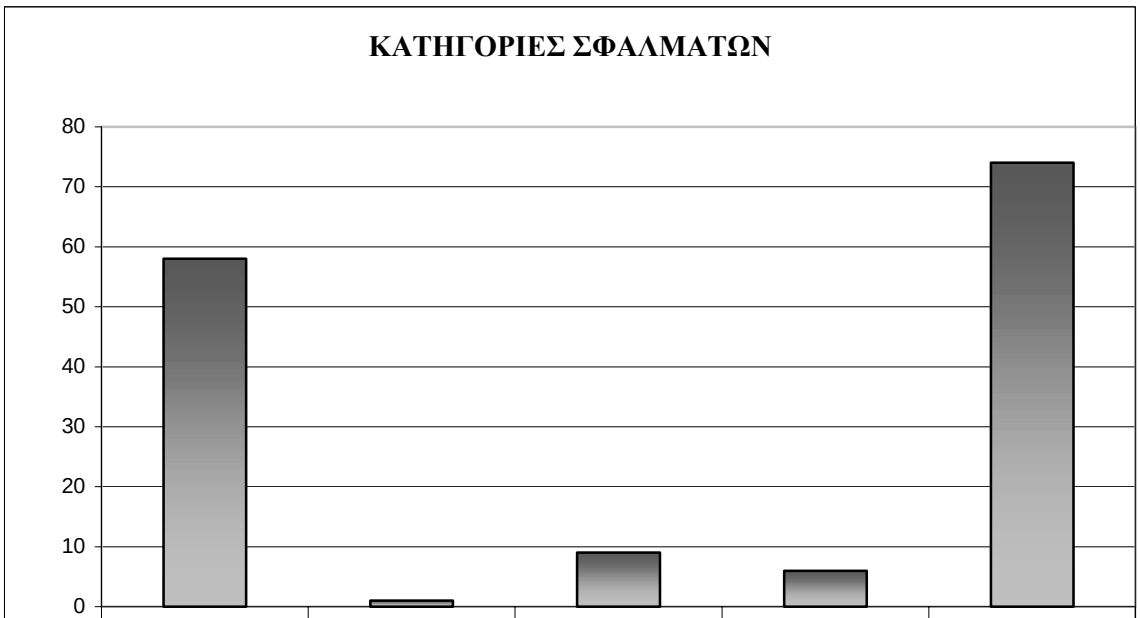
4. Αποτελέσματα - Αξιολόγηση

Για την εξαγωγή ποιοτικών και ποσοτικών συμπερασμάτων της μεθοδολογίας χρησιμοποιήθηκε ένα σώμα κειμένων (corpus) που έχει δημιουργηθεί στο εργαστήριο της "Ομάδας Ομιλίας" του Τμήματος Πληροφορικής και Τηλεπικοινωνιών του Πανεπιστημίου Αθηνών. Τα κείμενα αυτά έχουν συλλεχθεί από ιστοσελίδες του διαδικτύου (ειδήσεις, πολιτιστικά, λογοτεχνία) καθώς και από ερευνητικές εργασίες.

Στην βάση δεδομένων του συστήματος κωδικοποιήθηκαν αρχικώς 176 ειδικές περιπτώσεις αριθμητικών, ενώ ο αριθμός των ΓΕ ανέρχεται σε 838. Το μέγεθος του

λεξικού που χρησιμοποιήθηκε ήταν 1.077.458 λέξεις. Σε αυτή την υλοποίηση, όσες λέξεις του κειμένου εισόδου δεν περιλαμβάνονται στο λεξικό θεωρούνται σωστές εκφράσεις και δεν λαμβάνονται υπόψη κατά την επεξεργασία. Με τον τρόπο αυτό μπορεί να υπάρξει πρόβλημα στην περίπτωση που η εφαρμογή χρειαστεί την μορφολογία της άγνωστης λέξης.

Συνολικά 374 προτάσεις των κειμένων της βάσης περιείχαν αριθμούς ή ΓΕ και κανονικοποιήθηκαν. Από τις 374 προτάσεις, οι 285 κανονικοποιήθηκαν ορθά ενώ στις 74 παρατηρήθηκαν λάθη στην κανονικοποίηση. Ακολουθεί ένα σχετικό διάγραμμα όπου τα λάθη έχουν κατηγοριοποιηθεί (Σχήμα 2). Τα περισσότερα λάθη παρατηρήθηκαν στην κανονικοποίηση αριθμητικών και εμφανίστηκαν λόγω απουσίας συγκεκριμένης κανονικής έκφρασης. Ενδεικτικά μπορεί να αναφερθεί ότι εξαιτίας της απουσίας μιας κανονικής έκφρασης για την προφορά αριθμητικών που περιέχονται σε νόμους, τα λάθη κανονικοποίησης σε ένα νομικό κείμενο που περιείχε 88 προτάσεις, ανήλθαν σε 32.



	Δεν υπάρχει η καν. εκφ.	Δεν υπάρχει η ΓΕ στην ΒΔ	Αναγνώριση αριθμητικού	Άγνωστες λέξεις	ΣΥΝΟΛΟ ΛΑΘΩΝ
ΛΑΘΗ	58	1	9	6	74

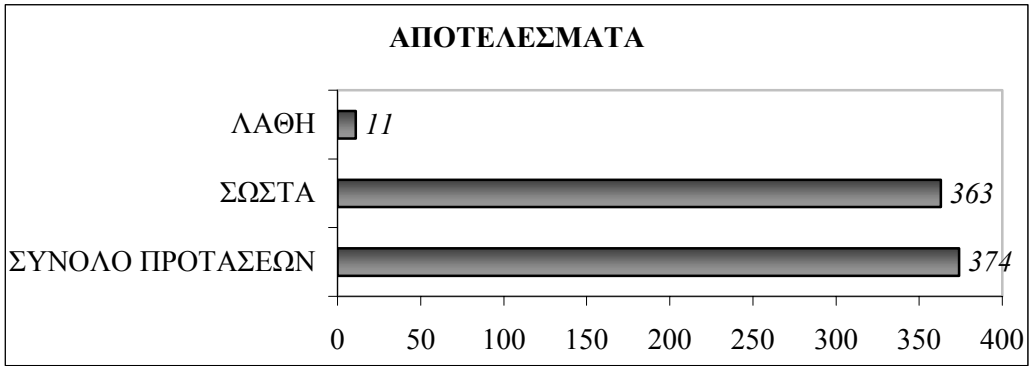
Σχήμα 2 Κατηγοριοποίηση σφαλμάτων.

Στη συνέχεια, έγιναν οι παρακάτω επεκτάσεις:

- προστέθηκαν οι κανονικές εκφράσεις που είχαν παραληφθεί και αφορούσαν αριθμητικά που περιέχονται κυρίως σε διατυπώσεις νομικών κειμένων.
- προστέθηκαν στο λεξικό ΓΕ, όσες ΓΕ δεν υπήρχαν αλλά εμφανίζονταν στα κείμενα.

και η διαδικασία ελέγχου ξαναεκτελέστηκε από την αρχή. Τα αποτελέσματα μετά από αυτές τις μικρές διορθώσεις ήταν αρκετά ικανοποιητικά, καθώς το ποσοστό σωστής κανονικοποίησης ήταν 97,06% (Σχήμα 3). Τα εναπομείνοντα λάθη οφείλονται σε:

1. άγνωστες λέξεις στο κείμενο εισόδου, (που δεν περιέχονται δηλαδή στο λεξικό της εφαρμογής),
2. αποτυχία της συντακτικής ανάλυση της πρότασης,
3. σε αριθμητικά που περιέχονται σε πίνακες με δεδομένα. π.χ. σε έναν πίνακα βαθμολογίας αγώνων ποδοσφαιρικού πρωταθλήματος, η στήλη που περιέχει τους βαθμούς πρέπει να είναι σε γένος αρσενικό. Η παρούσα υλοποίηση αδυνατεί να το εντοπίσει αλλά μπορεί να αντιμετωπιστεί σε νεότερη έκδοση.



Σχήμα 3 Τελικά αποτελέσματα αξιολόγησης της εφαρμογής.

5. Συμπεράσματα

Στην εργασία αυτή παρουσιάσαμε ένα νέο υβριδικό σύστημα κανονικοποίησης κειμένου, που μπορεί να χειρίζεται αποτελεσματικά αριθμητικές και συνεπτυγμένες γραπτές εκφράσεις (ΓΕ) σε κείμενα της ελληνικής γλώσσας. Το μοντέλο ανιχνεύει τις μη κανονικοποιημένες ΓΕ, τις ταξινομεί, αίρει τις αμφισημίες και παράγει τις γραμματικά και συντακτικά αντίστοιχες σωστές λεκτικές μορφές. Η μέθοδος βασίζεται

σε συνδυασμό κανόνων, πινάκων αντιστοίχησης και εφαρμογής επεξεργασίας φυσικής γλώσσας (μορφολογική, συντακτική και εννοιολογική ανάλυση). Η εργασία αποτελεί μια προσπάθεια μεθοδολογικής γενικής επίλυσης του προβλήματος που μπορεί να εφαρμοστεί σε οποιοδήποτε κείμενο, όχι συγκεκριμένου τομέα, στηρίζεται σε μια μη-εποπτευόμενη προσέγγιση και μπορεί να επεκταθεί και σε άλλες περιπτώσεις μη-λεκτικών μορφών. Στο εργαστήριο Ομιλίας του Πανεπιστημίου Αθηνών, σχεδιάστηκε μια εφαρμογή που υλοποιεί το μοντέλο αυτό σε λειτουργία πραγματικού χρόνου η οποία ενσωματώθηκε με επιτυχία στο σύστημα μετατροπής κειμένου σε ομιλία ΔΗΜΟΣΘέΝΗΣ [Xydas and Kouroupetroglou, 2001]. Η αξιολόγηση του συστήματος με μια συλλογή κειμένων από το διαδίκτυο καθώς και επιστημονικών εργασιών, έδειξε την αποτελεσματικότητα του μοντέλου, καθώς το ποσοστό ορθής κανονικοποίησης που επιτεύχθηκε ήταν 97,1%.

Αναφορές

Black Alan W., Chen Stanley, Kumar Shankar, Ostendorf Mari, Richards Christopher, Sproat Richard and Yarowsky David. 1999. “Normalization of non-standard words”. In www.clsp.jhu.edu/ws99/

Coughlin Deborah. 1999. “Leveraging Syntactic Information for Text Normalization”. Lecture Notes in Artificial Intelligence (LNAI), Vol. 1692, p.95-100.

Fries Georg and Wirth Antje. 1997. "FELIX – A TTS System with Improved pre-processing and source signal generation". In Proceedings of . EUROSPEECH 97, Vol. II, p. 589-592.

Karberis Georgios and Kouroupetroglou Georgios. 2002. “Transforming Spontaneous Telegraphic Language to Well-Formed Greek Sentences for Alternative and Augmentative Communication”. Lecture Notes in Artificial Intelligence (LNAI), Vol. 2308, p. 155-166.

Mikheev Andrei. 2000. “Document Centered Approach to Text Normalization”. In Proceedings of SIGIR'2000 (Athens) ACM, p. 136-143.

Mobius Bernd, Sproat Richard, Van Santen Jan P. H. and Olive Joseph P. 1997. “The Bell Labs German Text-To-Speech system: An overview”. In Proceedings of EUROSPEECH 97, Volume IV, p. 2443-2446.

Ooyama Yoshifumi, Miyazaki Masahiro and Ikehara Satoru. 1987. “Natural Language Processing in a Japanese Text-To-Speech System”. In Proceedings of the 15th Annual Computer Science Conference, p. 40-47.

Sproat Richard, Black Alan W., Chen Stanley, Kumar Shankar, Ostendorf Mari and Richards Christopher. 2001. "Normalization of non-standard words". Computer Speech and Language, 15(3), p. 287-333.

Toole Janine. 2000. “A Hybrid Approach to the Identification and Expansion of Abbreviation”. In Proceedings of RIAO 2000, Vol.1, p.725-736.

Xydas Gerasimos and Kouroupetroglou Georgios. 2001. "The DEMOSTHeNES Speech Composer". In Proceedings of the 4th ISCA Tutorial and Workshop on Speech Synthesis (SSW4), Perthshire, Scotland, p. 167-172.

Zingle Henri. 1982. “Traitement de la prosodie allemande dans un systeme de synthese de la parole”. These pour le ‘Doctorat d’Etat, Universite de Strasbourg II.