ΕΜΠΛΟΥΤΙΣΜΟΣ ΜΟΡΦΟΛΟΓΙΚΩΝ ΛΕΞΙΚΩΝ ΜΕ ΟΡΟΥΣ ΚΑΙ ΥΠΟΣΤΗΡΙΞΗ ΚΕΙΜΕΝΩΝ ΕΝΤΑΣΕΩΣ ΟΡΩΝ ΣΕ ΔΙΑΔΙΚΑΣΙΕΣ ΔΙΟΡΘΩΣΗΣ ΛΑΘΩΝ

Χ. Στάθης, Γ. Καραγιάννης

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή παρουσιάζεται το υποσύστημα εμπλουτισμού του περιβάλλοντος "Συμφωνία" του Ινστιτούτου Επεξεργασίας του Λόγου (ΙΕΛ). Η "Συμφωνία" είναι ένας διορθωτής τόσο ορθογραφικών λαθών όσο και λαθών συμφωνίας μεταξύ των λέξεων, που ανατρέχει και σε συντακτικούς κανόνες.

Χάρη στο υποσύστημα αυτό ο χρήστης θα επωφελείται από τον πλούτο της δομής του μορφολογικού λεξικού του περιβάλλοντος και θα μπορεί να δημιουργήσει ειδικευμένα ορολογικά λεξικά ειαάγοντας με ιδιαίτερη ευκολία οποιονδήποτε καινούριο όρο που απαντά σε ένα κείμενο εντάσεως όρων και δεν είναι γνωστός στο μορφολογικό λεξικό.

Μέσω μιας φιλικής διεπαφής ο χρήστης θα βοηθιέται στην ειαγωγή ενός νέου όρου. Το σύστημα προτείνει πιθανές κλίσεις του όρου είτε αυτός είναι ουσιαστικό είτε επίθετο είτε ρήμα κ.λ.π. και ο χρήστης επιλέγει την κατάλληλη. Έται ο όρος εισάγεται με όλες τις δυνατές μορφές του στο ειδικό ορολογικό - μορφολογικό λεξικό το οποίο θα μπορεί να συνεργασθεί αρμονικά με το μορφολογικό λεξικό της γενικής γλώσσας. Τα δύο μαζί θα μπορούν να καλύψουν ανάγκες διόρθωσης λαθών σε μελλοντικά κείμενα του χρήστη που θα είναι εντάσεως όρων στην ορολογική περιοχή του ενδιαφέροντος του.

SUMMARY

In this paper, we present the enrichment subsystem of the "Symphonia" application developed by the Institute for Language and Speech Processing. "Symphonia" is a spelling and agreement checker that uses simple syntax rules.

Taking advantage of the structure of the morphological dictionary, the user will be able to create specialized terminology dictionaries, by easily inserting each new term encountered in a document, which is unknown to the basic dictionary.

The interface is very ergonomic and user friendly. The system proposes a number of possible inflectional paradigms for the unknown word and the user chooses the correct one. Hence, every form of the new term will be known to the terminology dictionary, which will co-operate with the basic dictionary in order to be used for spelling and agreement checking in future documents that are rich in terms from the user 's area of interest.

1 ΕΙΣΑΓΩΓΗ

Ένα σημαντικό πρόβλημα που αντιμετωπίζει ο χρήστης ενός ορθογραφικού διορθωτή ενσωματωμένου σε κάποιον επεξεργαστή κειμένου είναι το πεπερασμένο πλήθος λέξεων που γνωρίζει το εργαλείο και μπορεί να χρησιμοποιήσει για την διόρθωση. Έτσι, σε ένα κείμενο με ειδικό λεξιλόγιο, όπως π.χ. σε ένα λογοτεχνικό κείμενο ή ένα κείμενο με ειδικούς όρους, θα υπάρχει ένας σημαντικός αριθμός από λέξεις άγνωστες στον διορθωτή. Όλοι οι ορθογραφικοί διορθωτές που κυκλοφορούν σήμερα στο εμπόριο προσφέρουν την δυνατότητα να προσθέσει ο χρήστης μια άγνωστη λέξη στο λεξικό που χρησιμοποιεί το εργαλείο. Οι τρόποι με τους οποίους γίνεται αυτό είναι κυρίως δύο. Σύμφωνα με τον πρώτο, που είναι και ο πιο διαδεδομένος, η νέα λέξη προστίθεται σε έναν κατάλογο αναζήτησης οπότε, αν συναντηθεί ξανά στο κείμενο ή σε άλλα κείμενα, στην ίδια όμως πάντα μορφή, θα αναγνωριστεί κανονικά. Το σημαντικό μειονέκτημα της μεθόδου είναι ότι, αν η λέξη συναντηθεί σε άλλη μορφή (π.χ. σε άλλη πτώση), δεν θα αναγνωριστεί και μάλιστα σε κάποιες περιπτώσεις θα θεωρηθεί λάθος και θα προταθεί στη θέση της ο τύπος που είχε εισαχθεί αρχικά. Κατά την δεύτερη μέθοδο, η οποία είναι πολύ σπάνια, ζητείται από τον χρήστη να δώσει ολόκληρη την κλίση της άγνωστης λέξης, ώστε να γίνει εισαγωγή κάθε υπαρκτού τύπου της λέξης. Έτσι, αν η λέξη συναντηθεί αργότερα στον ίδιο ή σε διαφορετικό τύπο θα αναγνωριστεί κανονικά. Το σημαντικό μειονέκτημα της μεθόδου αυτής είναι ότι έχοντας να δώσει ολόκληρη την κλίση, ο χρήστης μπορεί εύκολα να κάνει λάθος, με αποτέλεσμα να εισαχθεί λανθασμένη πληροφορία στο λεξικό. Επί πλέον, η εισαγωγή όλων των τύπων είναι κουραστική με αποτέλεσμα οι χρήστες να την αποφεύγουν.

Στο πλαίσιο της εργασίας αυτής θα παρουσιάσουμε το υποσύστημα εμπλουτισμού του μορφολογικού λεξικού που είναι τμήμα της εφαρμογής "Συμφωνία" του Ινστιτούτου Επεξεργασίας του Λόγου. Το υποσύστημα αυτό επιτρέπει στον χρήστη να εισάγει άγνωστες λέξεις που συναντά ο διορθωτής στο μορφολογικό λεξικό με ολόκληρη την κλίση τους, χωρίς όμως να υποχρεώνει τον χρήστη να δώσει ο ίδιος την κλίση. Το σύστημα ζητά από τον χρήστη μικρό αριθμό πληροφοριών για την λέξη και με βάση αυτές παράγει και προτείνει κάποιες πιθανές κλίσεις για την λέξη. Ο χρήστης καλείται να επιλέξει την σωστή κλίση, η οποία και εισάγεται τελικά στο λεξικό. Έτσι ο διορθωτής θα γνωρίζει στο μέλλον κάθε δυνατό τύπο της λέξης αποφεύγοντας όμως το μειονέκτημα της εισαγωγής όλων των τύπων από τον χρήστη.

2 ΛΙΓΑ ΛΟΓΙΑ ΓΙΑ ΤΗΝ "ΣΥΜΦΩΝΙΑ"

Το πρόγραμμα "Συμφωνία" αναπτύχθηκε από το ΙΕΛ με σκοπό να αποτελέσει ένα πλήρες σύστημα διόρθωσης για τον επεξεργαστή κειμένου Microsoft Word. Στην παρούσα φάση

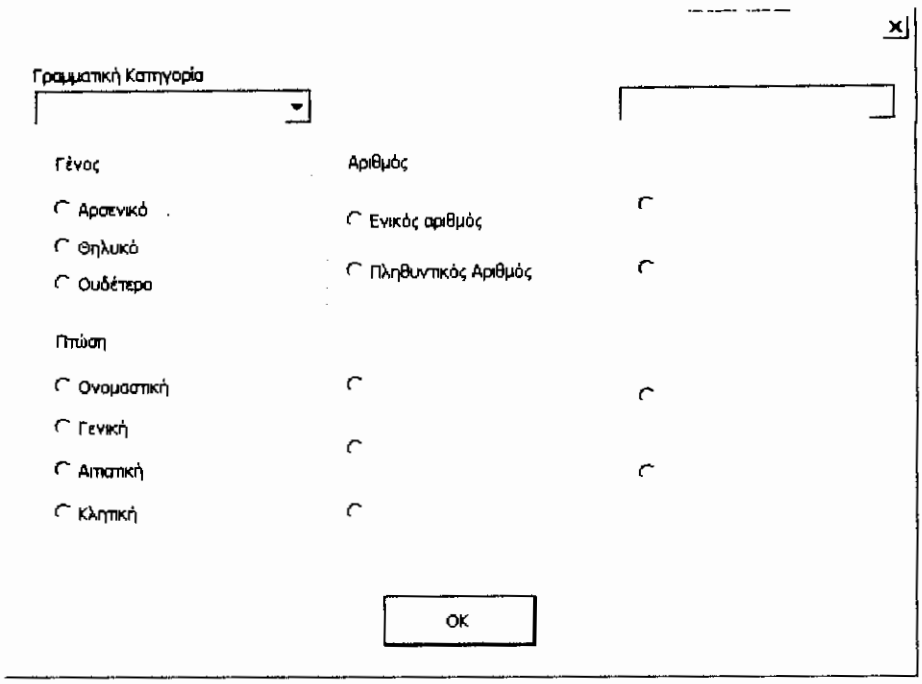
λειτουργεί για τις εκδόσεις Word 97 και Word 2000. Το πρόγραμμα αποτελείται από τρία υποσυστήματα: τον ορθογραφικό διορθωτή, τον έλεγχο συμφωνίας και τον εμπλουτισμό του λεξικού. Ο ορθογραφικός διορθωτής είναι συμβατός με την διεπαφή ορθογραφικής διόρθωσης του Word και λειτουργεί πλήρως ενσωματωμένος στο περιβάλλον του επεξεργαστή κειμένου. Ελέγχει την ορθογραφία καθώς πληκτρολογεί ο χρήστης και επισημαίνει τα λάθη με κόκκινη υπογράμμιση. Σε περίπτωση λάθους, ο χρήστης με δεξί κλίκ πάνω στο λάθος μπορεί να δει τις προτεινόμενες εναλλακτικές επιλογές, αν υπάρχουν, και να αντικαταστήσει την υπογραμμισμένη λέξη με κάποια από αυτές. Επίσης, για λόγους συμβατότητας, προσφέρεται και η δυνατότητα να γίνει εισαγωγή μιας άγνωστης λέξης χρησιμοποιώντας την απλή μέθοδο που αναφέρθηκε αρχικά, δηλαδή να εισάγεται σε ένα χωριστό κατάλογο ο συγκεκριμένος άγνωστος τύπος. Σε κάθε έλεγχο που πραγματοποιεί, ο διορθωτής συμβουλεύεται τον κατάλογο αυτόν και αν ο εξεταζόμενος τύπος υπάρχει στην λίστα θεωρείται σωστός. Η δυνατότητα αυτή εξυπηρετεί την περίπτωση των μη κλιτών λέξεων και ειδικών τύπων, όπως θα δούμε και στην συνέχεια.

Το υποσύστημα ελέγχου της "Συμφωνίας" εξετάζει μία προς μία τις λέξεις σε συνδυασμό με τις γειτονικές λέξεις και αποφασίζει για τα γραμματικά χαρακτηριστικά που θα έπρεπε να έχει η εξεταζόμενη λέξη σε συμφωνία με τα χαρακτηριστικά των λέξεων του περιβάλλοντος και με βάση κάποιους κανόνες που εξετάζονται σε κάθε περίπτωση. Αν η λέξη έχει πράγματι τα χαρακτηριστικά που επιβάλλουν οι κανόνες, τότε θεωρείται σωστή. Στην αντίθετη περίπτωση είτε προτείνονται τα σωστά χαρακτηριστικά (π.χ. η ίδια λέξη σε άλλη πτώση) είτε προτείνεται μια άλλη λέξη με τα σωστά χαρακτηριστικά (π.χ. δόση - δώσει). Ένα σημαντικό λάθος που γίνεται πολλές φορές αφορά τα ρήματα στο τρίτο πρόσωπο του ενικού της μεσοπαθητικής φωνής και το δεύτερο πρόσωπο του πληθυντικού της ενεργητικής φωνής (π.χ. συμβουλεύεται - συμβουλεύετε). Ο έλεγχος συμφωνίας, εξετάζοντας τα γραμματικά χαρακτηριστικά των γειτονικών λέξεων, μπορεί να αποφασίσει για τον σωστό αριθμό και να προτείνει τον σωστό τύπο. Επίσης συχνά γίνεται σύγχυση μεταξύ της γενικής ενικού και αιτιατικής πληθυντικού θηλυκών ουσιαστικών (π.χ. απόφασης - αποφάσεις) ή μεταξύ ρημάτων και παράγωγων θηλυκών ουσιαστικών (π.χ. διάταξης - διατάξεις). Σε όλες τις παραπάνω περιπτώσεις, η "Συμφωνία" μπορεί να αποφασίσει για τα σωστά γραμματικά χαρακτηριστικά της εξεταζόμενης λέξης και να βρει έτσι λάθη που δεν είναι ορθογραφικά και δεν θα εντοπίζονταν από τον απλό ορθογραφικό διορθωτή.

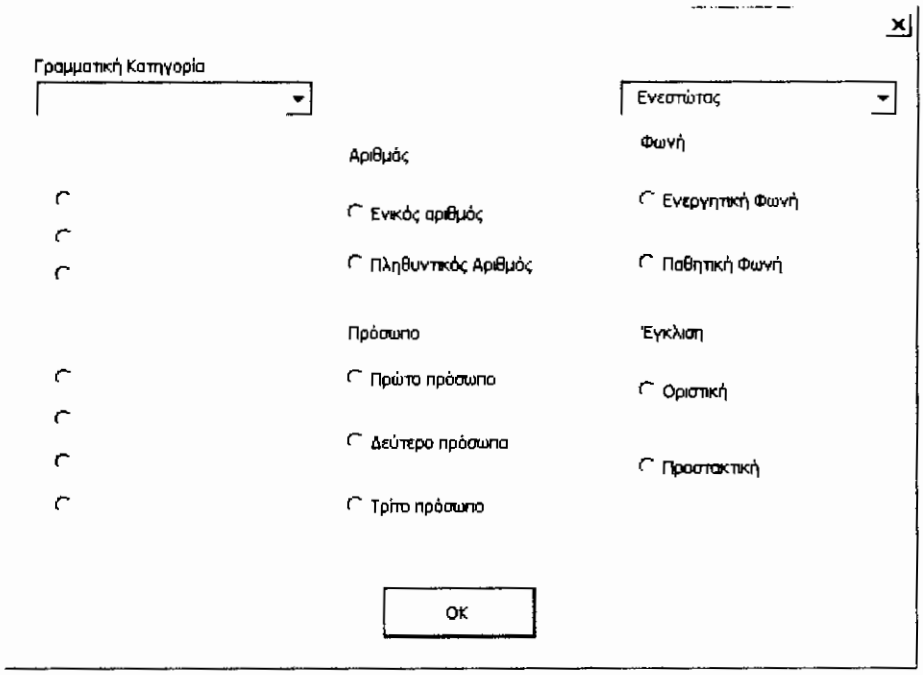
3 ΤΟ ΥΠΟΣΥΣΤΗΜΑ ΕΜΠΛΟΥΤΙΣΜΟΥ

Το υποσύστημα εμπλουτισμού έχει σκοπό να βοηθήσει τον χρήστη να εισάγει νέες λέξεις στο λεξικό με ολόκληρη την κλίση τους, ώστε να μην εμφανίζονται ως άγνωστες

όταν συναντηθούν σε οποιαδήποτε μορφή τους σε μελλοντικά κείμενα. Το πρόβλημα που αντιμετωπίζει το σύστημα είναι η ανεύρεση της ορθής κλίσης για μια λέξη η οποία είναι εντελώς άγνωστη στο λεξικό. Για τον σκοπό αυτό, ο χρήστης καλείται να δώσει κάποια χαρακτηριστικά για την λέξη, τα οποία σε συνδυασμό με την κατάληξη της λέξης χρησιμοποιούνται από το σύστημα προκειμένου αυτό να βρει κάποια κλιτικά παραδείγματα τα οποία να ταιριάζουν στα χαρακτηριστικά που έδωσε ο χρήστης και στην κατάληξη της λέξης. Πιο συγκεκριμένα ο αλγόριθμος έχει ως εξής: Το σύστημα αποκόπτει έναν χαρακτήρα από το τέλος της λέξης θεωρώντας τον ως κατάληξη και αναζητά τα κλιτικά παραδείγματα που περιέχουν αυτήν την κατάληξη με τα χαρακτηριστικά (π.χ. πτώση, αριθμό κλπ) που έδωσε ο χρήστης. Είτε βρεθούν τέτοια κλιτικά είτε όχι, το σύστημα επαναλαμβάνει την αναζήτηση έχοντας αποκόψει έναν ακόμη χαρακτήρα και θεωρώντας δύο χαρακτήρες ως κατάληξη. Η διαδικασία επαναλαμβάνεται αποκόπτοντας συνεχώς χαρακτήρες από το τέλος και θεωρώντας ολοένα και μεγαλύτερες καταλήξεις μέχρι το μέγιστο μήκος κατάληξης που υπάρχει στο μορφολογικά λεξικό. Στο τέλος αυτής της διαδικασίας έχει συγκεντρωθεί ένας αριθμός από κλιτικά παραδείγματα τα οποία παράγουν τον συγκεκριμένο άγνωστο τύπο με τα χαρακτηριστικά που έδωσε ο χρήστης. Με βάση τα κλιτικά αυτά παράγονται οι κλίσεις και παρουσιάζονται στον χρήστη, ο οποίος πρέπει να επιλέξει την σωστή. Ο παραπάνω αλγόριθμος εφαρμόζεται με μικρές παραλλαγές ανάλογα με την γραμματική κατηγορία που θα δώσει ο χρήστης για την λέξη. Η πρώτη οθόνη προσαρμόζεται ανάλογα με την γραμματική κατηγορία που θα δοθεί. Στο Σχήμα 1 φαίνεται η οθόνη επιλογής χαρακτηριστικών, όπου ο χρήστης έχει επιλέξει την γραμματική κατηγορία Ουσιαστικό. Στην περίπτωση αυτή είναι ενεργές οι επιλογές για το Γένος, Πτώση και Αριθμό. Το ίδιο θα συμβεί εάν επιλέξει Επίθετο. Στην περίπτωση που επιλέξει Ρήμα, θα ενεργοποιηθούν οι επιλογές για Πρόσωπο, Χρόνο, Φωνή και Έγκλιση, με ταυτόχρονη απενεργοποίηση των επιλογών για Γένος και Πτώση όπως φαίνεται στο Σχήμα 2. Πρέπει να τονιστεί ακόμα ότι ο χρήστης δεν είναι υποχρεωμένος να συμπληρώσει όλα τα χαρακτηριστικά. Απλώς όσο λιγότερα δώσει τόσο περισσότερα κλιτικά παραδείγματα θα ταιριάζουν και έτσι θα πρέπει στο τέλος να διαλέξει ανάμεσα σε περισσότερες κλίσεις.



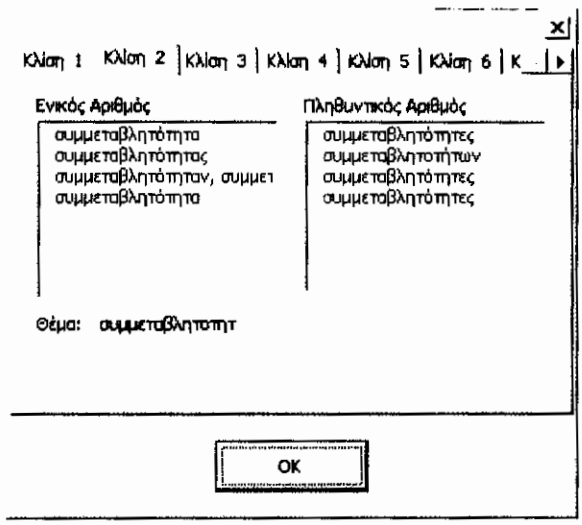
Σχήμα 1: Οθόνη επιλογής χαρακτηριστικών



Σχήμα 2: Οθόνη επιλογής χαρακτηριστικών στην περίπτωση ρήματος

3.1 ΟΥΣΙΑΣΤΙΚΑ

Στην περίπτωση των Ουσιαστικών ο χρήστης πρέπει να δώσει γένος, αριθμό και πτώση για την άγνωστη λέξη. Το σύστημα θα αναζητήσει τα κλιτικά που περιέχουν την κατάληξη στο συγκεκριμένο γένος, αριθμό και πτώση και θα παρουσιάσει τις κλίσεις. Η οθόνη παρουσίασης των κλίσεων φαίνεται στο Σχήμα 3.



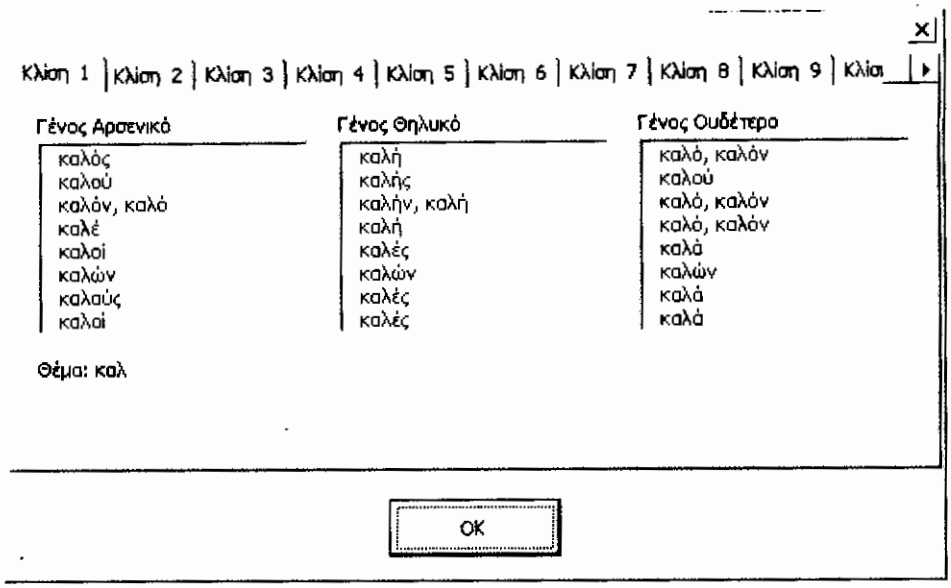
Σχήμα 3: Οθόνη παρουσίασης πιθανών κλίσεων ουσιαστικών

3.2 ΕΠΙΘΕΤΑ

Για τα επίθετα ισχύουν ακριβώς τα ίδια, μόνο που αλλάζει η οθόνη παρουσίασης των κλίσεων, η οποία φαίνεται στο Σχήμα 4.

3.3 ΜΕΤΟΧΕΣ

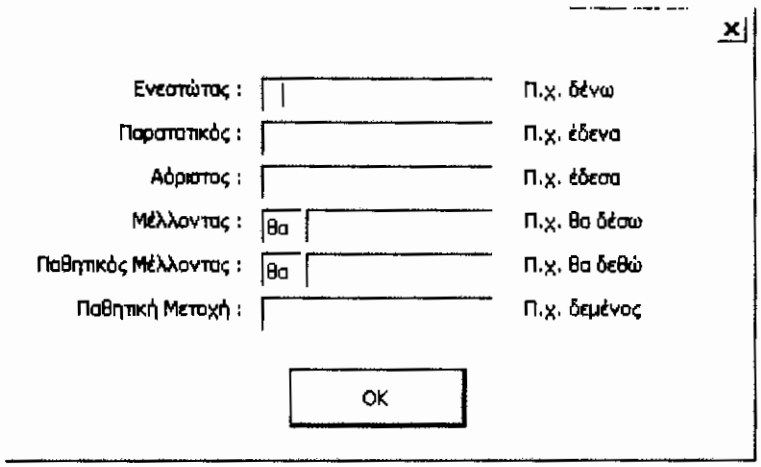
Η περίπτωση των μετοχών παθητικού παρακειμένου είναι ιδιόμορφη γιατί δεν αρκεί να βρεθεί μόνο η σωστή κλίση αλλά πρέπει, στην συνέχεια, να συνδεθεί η κλίση αυτή με το αντίστοιχο ρήμα, αν αυτό υπάρχει στο λεξικό. Η αναζήτηση κλίσης για τις παθητικές μετοχές δεν διαφέρει σε τίποτα από την περίπτωση των επιθέτων. Στην συνέχεια, αφού ο χρήστης διαλέξει κλίση, το σύστημα του ζητά να γράψει σε ποιο ρήμα ανήκει η μετοχή. Το ρήμα αναζητείται στο λεξικό και αν βρεθεί, τότε η μετοχή συνδέεται με το ρήμα. Αν το ρήμα δεν βρεθεί τότε αναζητείται και η κλίση του ρήματος σύμφωνα με την επόμενη παράγραφο.



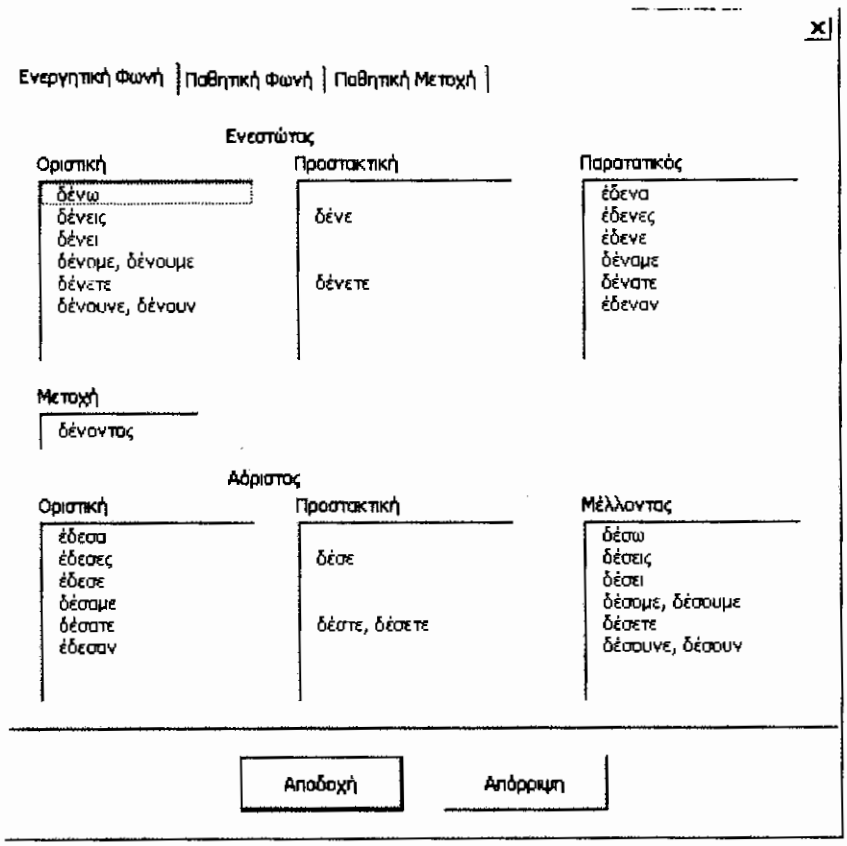
Σχήμα 4: Οθόνη παρουσίασης κλίσεων επιθέτων

3.4 ΡΗΜΑΤΑ

Επειδή η κλίση του ρήματος συνήθως παράγεται από πολλά θέματα και κλιτικά παραδείγματα, δεν αρκεί μόνο ο άγνωστος τύπος και τα χαρακτηριστικά του για να βρεθεί η πλήρης κλίση. Έτσι ζητούνται από τον χρήστη όλοι οι χρόνοι του ρήματος που πιθανά παράγονται από άλλα θέματα, ώστε να μπορέσει το σύστημα να προσδιορίσει όλα τα θέματα και από τις καταλήξεις να βρει και τα άλλα πιθανά κλιτικά παραδείγματα. Στο Σχήμα 5 φαίνεται η οθόνη στην οποία ο χρήστης δίνει τους διάφορους χρόνους για το ρήμα. Για κάθε χρόνο χωριστά γίνεται αναζήτηση της κλίσης και παρουσιάζονται διαφορετικές οθόνες σε κάθε περίπτωση. Αφού ο χρήστης επιλέξει κλίσεις για όλους τους χρόνους που έδωσε, παρουσιάζεται συγκεντρωμένη η κλίση του ρήματος για την τελική επισκόπηση (Σχήμα 5).



Σχήμα 5: Οθόνη εισαγωγής χρόνων ρήματος



Σχήμα 6: Οθόνη παρουσίασης συνολικής κλίσης ρήματος

3.5 ΑΚΛΙΤΑ

Η περίπτωση των ακλίτων περιλαμβάνει μόνο επιρρήματα δεδομένου ότι όλες οι άλλες κατηγορίες υπάρχουν πλήρως στο μορφολογικό λεξικό. Η εισαγωγή επιρρημάτων είναι πολύ απλή διότι δεν απαιτείται από τον χρήστη να δηλώσει παρά μόνο την γραμματική κατηγορία.

4 ΣΥΜΠΕΡΑΣΜΑΤΑ - ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

Στην εργασία αυτή παρουσιάστηκε το υποσύστημα εμπλουτισμού του μορφολογικού λεξικού που χρησιμοποιεί η εφαρμογή διόρθωσης "Συμφωνία". Με το υποσύστημα αυτό ο χρήστης του διορθωτή μπορεί να εισάγει στο λεξικό λέξεις που είναι άγνωστες, με ολόκληρη την κλίση τους. Έτσι σε περίπτωση που η λέξη συναντηθεί αργότερα σε άλλη μορφή θα είναι γνωστή στον διορθωτή. Ο χρήστης μέσω μιας φιλικής διεπαφής, δίνει κάποια στοιχεία για την άγνωστη λέξη και το σύστημα του παρουσιάζει κάποιες πιθανές κλίσεις από τις οποίες επιλέγει την σωστή.

Το υποσύστημα εμπλουτισμού μπορεί να χρησιμοποιηθεί σε κείμενα εντάσεως όρων για την δημιουργία μορφολογικού λεξικού με ειδική ορολογία. Σε μελλοντική μορφή οι νέοι όροι θα αποθηκεύονται σε ξεχωριστό λεξικό, το οποίο θα χρησιμοποιείται από τον διορθωτή μαζί με το βασικό για την διόρθωση ειδικευμένων κειμένων. Θα υπάρχει η δυνατότητα κατασκευής πολλών ειδικευμένων λεξικών και, μέσω ειδικής διεπαφής, ο χρήστης θα μπορεί να επιλέγει ποια ειδικά λεξικά θα χρησιμοποιούνται κάθε φορά μαζί με το βασικό, ανάλογα με το είδος του κειμένου που θέλει κάθε φορά να επεξεργαστεί.

Χρήστος Β. Στάθης, Ηλεκτρολόγος Μηχανικός Ε.Μ.Π.
Γιώργος Καραγιάννης, Καθηγητής Ε.Μ.Π.

Ινστιτούτο Επεξεργασίας του Λόγου
Αρτέμιδας 6 & Επιδαύρου 151 25 Μαρούσι