ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

Αλγόριθμοι εφαρμογής των N-grams στην αναγνώριση συναισθηματικού λόγου και στην διόρθωση κειμένων

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Θεολόγος Δ. Αθανασέλης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός &
Μηχανικός Υπολογιστών Δ.Π.Θ (2000)

Αθήνα, Μάιος 2007

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

Αλγόριθμοι εφαρμογής των N-grams στην αναγνώριση συναισθηματικού λόγου και στην διόρθωση κειμένων

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Θεολόγος Δ. Αθανασέλης
Διπλωματούχου Ηλεκτρολόγου Μηχανικού &
Μηχανικού Υπολογιστών Δ.Π.Θ (2000)

Συμβουλευτική Επιτροπή : Γεώργιος Καραγιάννης
Στέφανος Κόλλιας
Ιωάννης Δολόγλου

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 4η Μαΐου 2007.

.....	.....	.....
Γεώργιος Καραγιάννης	Στέφανος Κόλλιας	Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π
.....	.....	.....
Ανδρέας Σταφυλοπάτης	Παναγιώτης Τσανάκας	Τιμολέων Σελλής
Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π

.....
Εμμανουήλ Σαρρής
Καθηγητής Δ.Π.Θ

.....
Θεολογος Δ. Αθανασέλης

Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Πρόλογος

Οι λέξεις αποτελούν τον θεμέλιο λίθο κάθε γλώσσας. Κάθε ανθρώπινη γλώσσα, είτε αυτή είναι προφορική, γραπτή ή νοηματική, αποτελείται από λέξεις. Κάθε περιοχή αυτού που ονομάζουμε γλωσσική τεχνολογία, από την αναγνώριση φωνής ως την μηχανική μετάφραση και από εκεί στην ανάκτηση πληροφοριών μέσω του διαδικτύου απαιτεί ενδελεχή γνώση γύρω από τις λέξεις. Αυτό όμως που δίνει μεγαλύτερη ώθηση στην επικοινωνία των ανθρώπων είναι η επιλογή και η διάταξη των λέξεων ώστε να παραχθούν φράσεις, προτάσεις που μεταφέρουν χρήσιμη σημασιολογική, συντακτική, και πραγματολογική πληροφορία. Μέχρι σήμερα έχουν γίνει πολλές προσπάθειες να αναλυθεί η γλώσσα, χωρίς να έχουν λυθεί όλα τα προβλήματα. Μόνο όμως τα τελευταία χρόνια όταν διαφορετικοί τομείς όπως η αναγνώριση και σύνθεση φωνής, η επεξεργασία φυσικής γλώσσας, και η υπολογιστική γλωσσολογία, άρχισαν να συνεργάζονται και να μοιράζονται τις γνώσεις κάθε επιστήμης ανοίχθηκε ένας νέος δρόμος σε αυτό που ονομάζουμε επεξεργασία φωνής και γλώσσας. Ένας παράγοντας που συντέλεσε στην γρήγορη εξέλιξη της επεξεργασίας φωνής και γλώσσας αποτελεί και η χρήση μεγάλων σωμάτων κειμένων είτε σε επίπεδο γραπτού είτε σε επίπεδο προφορικού λόγου. Η συλλογή εκτεταμένων σωμάτων κειμένων που είναι αναμφίβολα μια επίπονη εργασία πολλών ατόμων οδήγησε την επιστημονική κοινότητα με την χρήση στατιστικών μεθόδων να συνειδητοποιήσει πολλές πτυχές της γλώσσας γύρω από την δομή της και την εξάρτηση των λέξεων. Στόχος όλων αυτών των προσπαθειών είναι να παραχθούν πραγματικές εφαρμογές που θα μπορούν να αναγνωρίζουν τα λεγόμενα ενός χρήστη, θα μπορούν να μεταφράζουν από γλώσσα σε γλώσσα τις ερωτήσεις ενός επισκέπτη του διαδικτύου και θα μπορούν να διορθώνουν τα γραπτά και να μαθαίνουν σε ένα μαθητή πώς να χειρίζεται την γλώσσα σε περιβάλλοντα μάθησης εξ' αποστάσεως.

Η διατριβή αυτή επικεντρώνεται σε δυο νέους αλγόριθμους εφαρμογής των στατιστικών μοντέλων N-grams στην αναγνώριση συναισθηματικού λόγου και στην διόρθωση κειμένων. Οι αλγόριθμοι εφαρμογής των N-grams προτείνονται και υλοποιούνται για να βελτιωθεί η απόδοση συμβατικών συστημάτων αναγνώρισης φωνής όταν χρησιμοποιείται συναισθηματικός λόγος και να διορθωθεί η σειρά των λέξεων σε προτάσεις που εμφανίζουν συντακτικά λάθη. Τα αποτελέσματα και συμπεράσματα της μελέτης αυτής απέδειξαν την συμβολή που μπορεί να έχουν τα στατιστικά μοντέλα των N-grams στον τομέα της επεξεργασίας της φυσικής γλώσσας.

Στο πλαίσιο της διδακτορικής μου διατριβής γνώρισα ανθρώπους που με βοήθησαν πραγματικά να ανταποκριθώ στις προκλήσεις αυτής της έρευνας όπως τον επιβλέποντα μου, Καθ. Γ. Καραγιάννη, που με την αμέριστη συμπαράσταση του με οδήγησε στο να μπορέσω να

πραγματοποιήσω τους αρχικούς μου στόχους. Επίσης θέλω να τον ευχαριστήσω που με εμπιστεύτηκε σε πολλές ερευνητικές εργασίες στο Ινστιτούτο Επεξεργασίας του Λόγου (ΙΕΛ). Παράλληλα συνεργάστηκα με δυο πολύ αξιόλογους ερευνητές του ΙΕΛ, τον κ. Ιωάννη Δολόγλου, Ερευνητή Α΄ (μέλος της επιτροπής) και τον κ. Στέλιο Μπακαμίδη, Ερευνητή Α΄, που θέλω να τους ευχαριστήσω για την αδιάλειπτη βοήθεια που μου παρείχαν καθ' όλη την διάρκεια των σπουδών μου. Οι γνώσεις και η εμπειρία τους συνέτειναν στο να μπορέσω να ανταποκριθώ στις δυσκολίες της έρευνας αυτής, που πολλές φορές φαίνονταν ως ανυπέρβλητα εμπόδια. Επίσης θέλω να ευχαριστήσω τον Καθ. Σ. Κόλλια για την ουσιαστική συνεργασία που είχαμε σε θέματα επικοινωνίας ανθρώπου-μηχανής στο πλαίσιο ευρωπαϊκών ερευνητικών προγραμμάτων. Συνεργάστηκα άψογα με τον προπτυχιακό φοιτητή Κ. Μαμούρα και θέλω να τον ευχαριστήσω για την ειλικρινή του βοήθεια όπως και με τους συμφοιτητές μου, Γ. Γιαννόπουλο, Ε. Τσιλιγιάννη, και Α. Χαλαμανδάρη που συγκατοικήσαμε για πάνω από πέντε χρόνια στο ίδιο γραφείο. Στο σημείο αυτό, θέλω να αφιερώσω ότι έχω κάνει και ότι θα ευτυχίσω να κάνω στην υπόλοιπη μου ζωή σε δυο ανθρώπους, αρωγούς και συμπαραστάτες, τους γονείς μου Δημήτρη και Μαρία, που με στηρίζουν σε κάθε φάση της ζωής μου. Τέλος δεν πρέπει να ξεχάσω την φίλη μου Νατάσσα, που ανέχεται όλα αυτά τα χρόνια, τις παραξενιές και τις ιδιοτροπίες μου.

Θεολόγος Αθανασέλης

Αθήνα, Μάιος 2007

Περίληψη

Το στατιστικό γλωσσικό μοντέλο, χρησιμοποιεί τεχνικές στατιστικής εκτίμησης γλωσσικών δεδομένων εκπαίδευσης, που εφαρμόζονται σε εκτεταμένα κείμενα, με σκοπό την μοντελοποίηση της γλώσσας. Ανάμεσα στις πιο δημοφιλείς τεχνικές στατιστικής εκτίμησης είναι και τα μοντέλα N-grams. Ο ρόλος τους είναι πολύ σημαντικός για μια σειρά από εφαρμογές της γλωσσικής τεχνολογίας, όπως η αναγνώριση φωνής, η οπτική αναγνώριση χαρακτήρων, η μηχανική μετάφραση και ακόμη η ορθογραφική διόρθωση. Με την παρούσα εργασία προτείνονται δυο νέοι αλγόριθμοι εφαρμογής των N-grams μοντέλων στην αναγνώριση φωνής συναισθηματικού λόγου και στην διόρθωση κειμένων.

Με αυτόν τον τρόπο η εργασία χωρίζεται σε δυο ενότητες. Στην πρώτη παρουσιάζεται ο αλγόριθμος εφαρμογής των N-grams μοντέλων στην αναγνώριση συναισθηματικού λόγου. Η αναγνώριση της γλωσσικής πληροφορίας του συναισθηματικού λόγου εκτός του ενδιαφέροντος που προκαλεί, παρουσιάζει και σημαντικά προβλήματα. Τα ποσοστά επιτυχίας των υπαρχόντων συστημάτων αναγνώρισης φωνής είναι αρκετά χαμηλά για εκφράσεις που έχουν έντονο συναισθηματικό χρώμα. Για αυτόν τον λόγο αναπτύχθηκε ένας αλγόριθμος που δημιουργεί ένα σώμα κειμένου με έντονο συναισθηματικό χαρακτήρα με την χρήση ενός συναισθηματικού λεξικού. Το επαυξημένο γλωσσικό μοντέλο υπολογίζεται από τον συνδυασμό ενός απλού σώματος κειμένου και του σώματος κειμένου με έντονο συναισθηματικό χαρακτήρα. Η ενσωμάτωση του επαυξημένου γλωσσικού μοντέλου σε ένα κλασσικό σύστημα αναγνώρισης φωνής έχει σαν αποτέλεσμα την βελτίωση της απόδοσης του κατά 20%.

Η δεύτερη ενότητα της εργασίας αυτής αφορά την χρήση των μοντέλων N-grams στην διόρθωση κείμενων που εμφανίζουν λάθη στην σειρά των λέξεων. Ο αλγόριθμος που αναπτύχθηκε έχει σαν στόχο την διόρθωση μιας πρότασης με λέξεις που βρίσκονται σε μη κατάλληλη θέση. Για αυτόν το λόγο λαμβάνονται υπόψη όλοι οι πιθανοί συνδυασμοί αντιμεταθέσεων των λέξεων της πρότασης εισόδου. Όμως για προτάσεις με N λέξεις έχουμε N! συνδυασμούς αντιμεταθέσεων και γίνεται κατανοητό ότι ο χώρος αναζήτησης είναι πολύ μεγάλος. Έτσι προτείνεται μια νέα μέθοδος γρήγορης αναζήτησης για τον περιορισμό των αντιμεταθέσεων που στηρίζεται στα έγκυρα διγράμματα. Οι παραγόμενες προτάσεις-αντιμεταθέσεις εξετάζονται και αξιολογούνται βάσει του αριθμού των έγκυρων τριγραμμάτων. Αποτέλεσμα αυτής της μεθόδου είναι η ανίχνευση και η διόρθωση προτάσεων με λάθη στην σειρά των λέξεων.

Λέξεις κλειδιά

Αναγνώριση Φωνής Συναισθηματικού Λόγου, Συναισθηματικά Εμπλουτισμένο Γλωσσικό Μοντέλο, Λεξικό Με Συναισθηματικούς Όρους, Διόρθωση Κειμένων, Μη Ορθή Σειρά Λέξεων, Μέθοδος Γρήγορης Αναζήτησης Βέλτιστης Λύσης, Φιλτράρισμα Αντιμεταθέσεων, Πίνακας Αντιστοίχισης.

Abstract

Statistical language model aims to estimate the probability distribution of various linguistic units such as words and sentences. Language models employ statistical estimation techniques using text. The most popular language models are N-grams models. These models are fundamental to a variety of language technologies, such as speech and optical recognition, statistical machine translation, and spelling correction. In the framework of this work, two new algorithms are introduced, for applying N-grams models in emotional speech recognition and sentence correction.

This work can be divided into two sections. The first one presents the algorithm for applying N-grams in emotional speech recognition. In spite of the remarkable recent progress in Large Vocabulary Recognition (LVR), it is still far behind the ultimate goal of recognising emotional speech. Read speech and non-read speech in a 'careful' style can be recognised with high accuracy using the state-of-the-art speech recognition technology. On the other hand, the classic Automatic Speech Recognition (ASR) faces problem on recovering the verbal content of the emotional speech. This work identifies a strategy, which hinges on the intuition that emotion affects language as well as speech variables. The issue is to identify corpora that reflect emotion-influenced language so that emotion-oriented language models can be trained from them. This work explains how an emotion-oriented language model (LM) can be generated from a standard corpus using an emotional dictionary. The emotional corpus is created by combining the standard corpus with the emotional corpus. This result corpus is subsequently used to design emotionally enriched language models that allow improved recognition performance with emotional utterances. Using a language model based on that technique improves recognition rate by about 20%.

The second section concerns the use of N-grams in text correction, in order to identify word order errors and repair them. The proposed algorithm handles the word order errors using all the possible words permutations of the sentence. Note that, given a sentence with length N, the number of permutations is N!. This is a very large number and seems to be restrictive for further processing. For that reason, a new method is introduced, for repairing word order errors in sentences using the probabilities of most typical bigrams and trigrams, extracted from a large text corpus. This work presents an approach for repairing word order errors in text by reordering words in a sentence and choosing the version that maximizes the number of trigram hits according to a language model. The novelty of this method concerns the use of a fast algorithm for reordering the words. The fast algorithm's robustness relies on the use of the valid bigrams and not on every single pair of words. The correctness of each permuted sentence depends on the number of valid trigrams. Finally, this method detects and repairs sentences with wrong word order providing a list of N-best sentences.

KeyWords

Emotional Speech Recognition, Emotionally Oriented Language Model, Dictionary of Affect in Language, Text Correction, Fast Search Algorithm, Word Order Errors, Permutations Filtering, Confusion Matrix

Συντομογραφίες

ΕΒΣΚ	Εθνικό Βρετανικό Σώμα Κειμένου
ΚΜΜ	Κρυφά Μαρκοβιανά Μοντέλα
ΜΓΑ	Μέθοδος Γρήγορης Αναζήτησης
Μ.Ο	Μέσος Όρος
Ο.Φ	Ονοματική Φράση
Π.Α	Πίνακας Αντιστοίχισης
Ρ-Υ-Α	Ρήμα-Υποκείμενο-Αντικείμενο
Υ-Ρ-Α	Υποκείμενο-Ρήμα-Αντικείμενο
BNC	British National Corpus
ERMIS	Emotionally Rich Man Machine Interaction System
DAL	Dıctıonary of Affect language
HMM	Hidden Markov Model
LM	Language Model
MFCC	Mel Frequency Cepstral Coefficients
PCFG	Probabilistic Context Free Grammar
POS	Part of Speech
SAL	Sensitive Artificial Listener
SALAS	Sensitive Artificial Listeners Association
TOEFL	Test of English as a Foreign Language
WSJ	Wall Street Journal

Κεφάλαιο 1

1 Επεξεργασία γλώσσας

Αντικείμενο της γλωσσολογίας είναι ο χαρακτηρισμός και η εξήγηση του μεγάλου πλήθους από γλωσσικές παρατηρήσεις που μας περιβάλλουν στις συνομιλίες μας, στον γραπτό λόγο και σε οποιοδήποτε άλλο μέσο (Manning και Schutze, 1999). Ένα μέρος όλου αυτού έχει να κάνει με την γνωστική πλευρά δηλαδή με το πώς ο άνθρωπος μαθαίνει, παράγει και καταλαβαίνει την γλώσσα, ένα άλλο μέρος έχει σχέση με την κατανόηση των σχέσεων των γλωσσικών εκφράσεων και ένα άλλο μέρος αφορά την κατανόηση των γλωσσικών δομών με τις οποίες η γλώσσα «επικοινωνείται». Στην προσπάθεια οι άνθρωποι να προσεγγίσουν το τελευταίο πρόβλημα, υιοθετήθηκαν κανόνες που διέπουν την δομή των γλωσσικών εκφράσεων.

Φυσικά γίνεται κατανοητό ότι κάθε γραμματική έχει και τα προβλήματα της και κάπου υστερεί. Δεν είναι πιθανόν και εφικτό να παρέχουμε με ακρίβεια και λεπτομέρεια ένα χαρακτηρισμό για ποιες προτάσεις είναι ορθές και τι τις διαχωρίζει από όλες τις άλλες που λογίζονται ως μη ορθές. Αυτό εξηγείται εάν αναλογιστούμε ότι οι άνθρωποι αλλοιώνουν πολλές φορές συνειδητά αυτούς τους κανόνες που διέπουν την ανθρώπινη επικοινωνία. Συνεπώς μερικές φορές οι κανόνες "χαλαρώνουν" ώστε να γίνει πιο παραγωγική η χρήση της γλώσσας, χωρίς αυτό να σηματοδοτεί ότι όλοι οι κανόνες είναι λανθασμένοι (Manning και Schutze, 1999).

Άρα είναι προτιμότερο να ασχοληθούμε με το τι πρότυπα χρησιμοποιούνται κατά την χρήση της γλώσσας και όχι με το να προσπαθούμε να διακρίνουμε και να χωρίσουμε τις προτάσεις σε γραμματικά ορθές και μη. Στην προσπάθεια μας αυτή χρησιμοποιούμε την στατιστική που βασίζεται στην θεωρία των πιθανοτήτων. Ο σκοπός αυτής της εργασίας είναι να δείξουμε πώς μπορεί ένα στατιστικό μοντέλο της γλώσσας να χρησιμοποιηθεί σε διαφορετικές εκφάνσεις του αντικειμένου που καλείται επεξεργασία φυσικής γλώσσας. Άλλωστε η πρακτική εφαρμογή της στατιστικής μοντελοποίησης της γλώσσας σε μια πλατιά γκάμα εφαρμογών με επιτυχή αποτελέσματα αποδεικνύει την καταλληλότητα αυτής της μεθόδου.

Στην στατιστική επεξεργασία φυσικής γλώσσας, οι άνθρωποι δεν παρατηρούν απλά τον τρόπο χρήσης της γλώσσας μέσα από ηχητικά σήματα αλλά βασίζονται σε κείμενα και λαμβάνουν σοβαρά υπόψη τους το γλωσσικό περιεχόμενο των κειμένων, όπου και αντανακλά κατά κάποιο τρόπο την δομή και τα φαινόμενα της γλώσσας. Υιοθετώντας λοιπόν προσεγγίσεις όπως αυτές που βασίζονται στην επεξεργασία σωμάτων κειμένων οι ερευνητές συνηγορούν στην άποψη ενός Βρετανού γλωσσολόγου, του J.R. Firth (1957) και είχε ως σλόγκαν ότι «*μπορείς να μάθεις μια λέξη αν γνωρίζεις την παρέα της*», που προσπαθούσε να βρει αυτόματες μεθόδους μοντελοποίησης της γλώσσας ώστε να εξερευνήσει την δομή της. Ας δούμε λοιπόν τα βασικά σημεία της γλώσσας και τον ρόλο της, στην επικοινωνία των ανθρώπων πριν περιγράψουμε τις αρχές της μοντελοποίησης της γλώσσας.

1.1 Γλώσσα

Η επιστημονική εξέλιξη της γλώσσας επιβάλλει την διάκριση δύο βασικών της πλευρών, της γλώσσας ως εσωτερικού συστήματος («λόγου») και της γλώσσας ως εφαρμογής αυτού του συστήματος κατά άτομα («ομιλίας») (Μπαμπινιώτης, 1998).

Γλώσσα είναι ο λόγος και η ομιλία το εσωτερικό, γενικό σύστημα που χαρακτηρίζει την δομή μιας φυσικής γλώσσας (ο λόγος) και η συγκεκριμένη από τα άτομα μιας γλωσσικής κοινότητας πραγμάτωση του (η ομιλία). Γλώσσα είναι εξάλλου και οι φθόγγοι που αποτελούν τη μορφή μιας γλώσσας είτε ως εσωτερικές οντότητες (ακουστικές εικόνες των λέξεων) είτε ως υλικές πραγματώσεις των ακουστικών εικόνων ως φυσικοί φθόγγοι (κινητικο-ακουστικής υφής). Η γλώσσα θα μπορούσε να ορισθεί χονδρικά ως το κύριο μέσο επικοινωνίας μεταξύ των μελών μιας κοινωνίας. Γλώσσα είναι βεβαίως και το περιεχόμενο, οι σημασίες που δηλώνονται από τις διάφορες φωνολογικές μορφές. Αλλά γλώσσα είναι και η σύζευξη της σημασίας και της φωνολογικής αντιπροσώπευσης περιεχομένου και μορφής. Γλώσσα είναι ακόμη και η επικοινωνία των ανθρώπων σε μια κοινωνία (γλωσσική κοινότητα). Τέλος γλώσσα είναι και η κατάσταση και εξέλιξη της επικοινωνίας δηλαδή το σύστημα του λόγου και τα διαχρονικά στάδια της εξέλιξης του (Μπαμπινιώτης, 1998).

1.2 Γλωσσική επικοινωνία

1.2.1 Ο ρόλος της γλώσσας

Χονδρικά θα μπορούσε κάποιος να υποστηρίξει ότι η γλώσσα είναι το κύριο μέσο της επικοινωνίας μεταξύ των ανθρώπων. Προτού εξετάσουμε τα βασικά στοιχεία ενός τέτοιου ορισμού ας μιλήσουμε για την υφή της γλώσσας και κυρίως ποια είναι τα συστατικά της. Τα

απαραίτητα συστατικά κάθε γλώσσας είναι δύο, οι φθόγγοι και οι σημασίες, ή αλλιώς η μορφή και το περιεχόμενο. Στην γλώσσα δεν υπάρχουν φθόγγοι χωρίς σημασίες και σημασίες που να μην εκφράζονται από φθόγγους. Περαιτέρω τόσο οι φθόγγοι όσο και οι σημασίες έχουν σε κάθε γλώσσα μια ιδιαίτερη δομή. Έτσι η φωνολογική δομή τής Νεοελληνικής γλώσσας επιτρέπει μεταξύ άλλων στην αρχή των λέξεων τους συνδυασμούς (συμφωνικά συμπλέγματα) των /χθ-/,/tr-/,/kl/, όχι όμως και την αντίστροφη διάταξη των ίδιων συμπλεγμάτων */θχ-/,*/rt-/,*/lk/. Ομοίως στο επίπεδο των σημασιών, ενώ σημασιολογικοί συνδυασμοί του τύπου «ο τοίχος βάφτηκε άσπρος» είναι απολύτως αποδεκτοί σημασιολογικές συνάψεις όπως «ο τοίχος θρηνεί» δεν είναι αποδεκτές στην κυριολεκτική χρήση της γλώσσας.

Από τις πιο απλές μονάδες βαίνουμε στις πιο σύνθετες και στο τέλος καταλήγουμε στις συνθετότερες. Από τους φθόγγους στα φωνήματα (τους φθόγγους που έχουν διαφορετική αξία για την σημασία των λέξεων) από τα φωνήματα στα μορφήματα (στις ελάχιστες σημασιολογικές μονάδες) από τα μορφήματα στις λέξεις (που μπορούν και αποτελούνται από ένα ή περισσότερα μορφήματα) από τις λέξεις στις φράσεις (τα συντάγματα, τους μικρότερους δυνατούς συντακτικούς συνδυασμούς λέξεων) και από τις φράσεις στις προτάσεις (διάφορα δομικά σχήματα) τέλος από τις προτάσεις σε μεγαλύτερα σύνολα προτάσεων (παραγράφους, κείμενα).

1.2.2 Λειτουργία της γλωσσικής επικοινωνίας

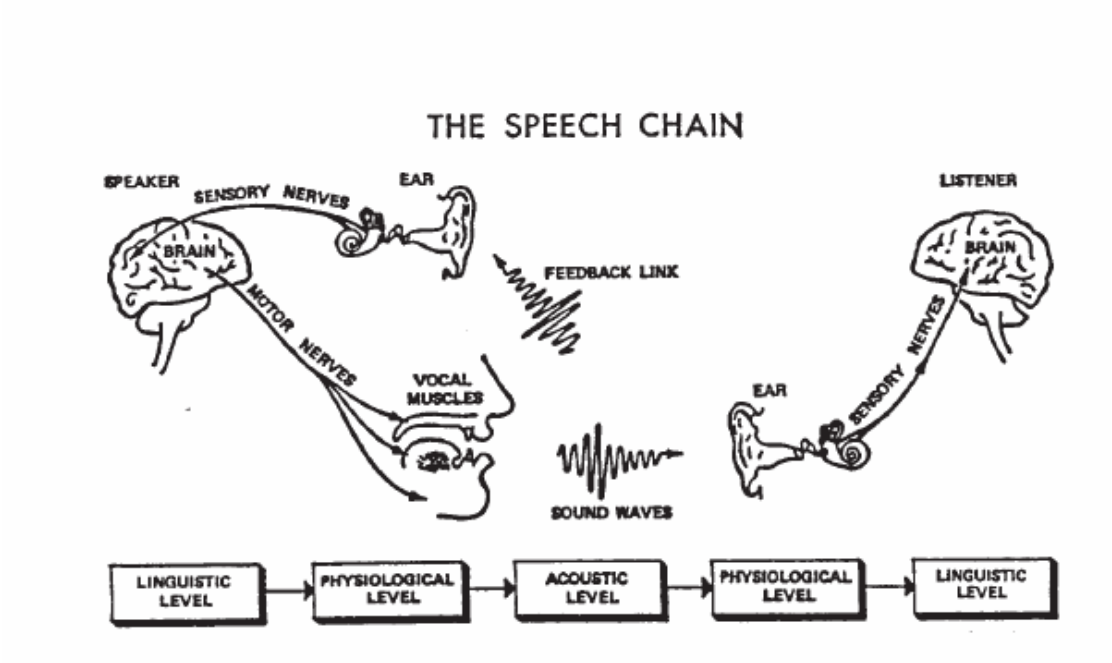
Άρα θέλοντας να δώσουμε έναν ορισμό της γλώσσας στο επίπεδο της γλωσσικής επικοινωνίας θα λέγαμε ότι

> *«η γλώσσα είναι ένας κώδικας σημείων ορισμένης μορφής (γλωσσικής), με την οποία επιτυγχάνεται η επικοινωνία μεταξύ των μελών μιας γλωσσικής κοινότητας» (Μπαμπινιώτης, 1998)*

Με τον όρο γλωσσικό κώδικα νοούμε ένα πεπερασμένο σε αριθμό σύστημα συστατικών στοιχείων με απεριόριστη δυνατότητα συνδυασμών. Τα συστατικά του γλωσσικού κώδικα απαρτίζονται από γλωσσικά σημεία (λέξεις) δηλαδή από συνδυασμούς σημασίας και φθόγγων που χρησιμεύουν ως μονάδες και διαρθρώνονται δομικά (συντακτικά) σε σύνολα (προτάσεις) βάσει κανόνων.

Γλωσσικά σημεία είναι γενικώς τα στοιχεία που απαρτίζουν ένα μήνυμα, μια μορφή έκφρασης, μια οποιαδήποτε μορφή επικοινωνίας. Έτσι ένα γλωσσικό μήνυμα αποτελείται

χονδρικά από λέξεις, ενώ ένα μουσικό μήνυμα αποτελείται από μουσικούς φθόγγους. Τα διάφορα είδη σημείων (τα σήματα Morse, τα συμπτώματα των ασθενειών) συγκροτούν αντίστοιχα σημειακά συστήματα. Τέτοια συστήματα είναι η γλώσσα, η μουσική, η ζωγραφική, η αρχιτεκτονική, η αριθμητική, ο κώδικας κυκλοφορίας (σήματα τροχαίας). Τα γλωσσικά σημεία, οι λέξεις, αποτελούν ένα σύνολο στοιχείων με τα οποία κυρίως επικοινωνούμε. Παράλληλα σε διαφορετικό βαθμό υπάρχει και το σύνολο των χειρονομιών και λοιπών εκφραστικών κινήσεων (κινήσεις κεφαλιού, νεύματα κτλ) που συμπληρώνουν και χρωματίζουν την γλωσσική επικοινωνία.



Σχήμα 1.1 Η αλυσίδα παραγωγής και κατανόησης του προφορικού λόγου.

Αντικειμενικός στόχος της γλώσσας είναι φυσικά η επικοινωνία. Λέγοντας επικοινωνία εννοούμε την ανταλλαγή μηνυμάτων, δηλαδή την εκπομπή μηνυμάτων από τον πομπό (ομιλητής) και την λήψη αυτών εκ μέρους του δέκτη (ακροατής) με το οποίο πραγματοποιείται η συνεννόηση μεταξύ των μελών μίας γλωσσικής κοινότητας (Σχήμα 1.1).

Η λειτουργία της γλωσσικής επικοινωνίας και τα στοιχεία που την αποτελούν μπορούν να παρασταθούν στο παραπάνω διάγραμμα (Σχήμα 1.1). Θέλοντας να αναλύσουμε τα κυριότερα μέρη κάθε επικοινωνίας μπορούμε πούμε ότι σε κάθε επικοινωνία δεσπόζουν τα εξής:

1. οι επικοινωνούντες, δηλαδή ο πομπός και ο δέκτης.
2. ο κώδικας
3. το μήνυμα, δηλαδή το ερώτημα, η απάντηση, το σχόλιο, η έκφραση σκέψης.
4. ο δίαυλος, δηλαδή ο τρόπος (φυσικός ή τεχνητός) μετάδοσης τού μηνύματος.

5. οι λειτουργίες εγγραφής (κωδικοποίησης) και ανάγνωσης (αποκωδικοποίησης) του μηνύματος και
6. οι συνθήκες επικοινωνίας.

Έτσι λοιπόν θέλοντας να περιγράψουμε τα βήματα μιας συνομιλίας θεωρούμε ότι ο ομιλητής στέλνει ένα μήνυμα στον ακροατή. Το μήνυμα κωδικοποιείται, εκφράζεται δηλαδή με σημεία (λέξεις) σε ορισμένη προκαθορισμένη κατά περίπτωση δομή, δηλαδή με στοιχεία που αντλούνται από τον κώδικα (το σύστημα γλώσσας) του ομιλητή. Ακολουθεί η μετάδοση του μηνύματος από τον ακροατή στον δέκτη διά των φυσικών ή τεχνητών διαύλων που επιτελείται με την χρήση ηχητικών και/ή οπτικών σημάτων. Στη συνέχεια αρχίζει η αντίστροφη διαδικασία. Το μήνυμα προσλαμβάνεται από τα ακουστικά ή οπτικά όργανα του ακροατή και αποκωδικοποιείται βάσει των στοιχείων του κώδικα (του εσωτερικού συστήματος γλώσσας του ακροατή) για να γίνει τελικά αντιληπτό από τον ακροατή.

Λέγοντας ότι η γλώσσα είναι ο κύριος παράγοντας και χρησιμεύει ως όργανο επικοινωνίας ξεχνάμε ότι η γλώσσα είναι ο κύριος φορέας της σκέψης του ανθρώπου. Νοήματα και έννοιες, αισθήματα, και συναισθήματα, ψυχοσύνθεση, νοοτροπία, και στάση ατόμων και ομάδων έναντι του κόσμου ενσωματώνονται, μορφοποιούνται και εκφράζονται μέσω της γλώσσας.

1.3 Τεχνικές μοντελοποίησης της γλώσσας

Ένα μοντέλο γλώσσας είναι απλά μια περιγραφή της γλώσσας. Σε πιο απλή μορφή αποτελεί την αναπαράσταση μιας λίστας από προτάσεις που ανήκουν σε μια γλώσσα. Τα πιο σύνθετα μοντέλα γλώσσας ενδεχομένως να επιδιώκουν να περιγράψουν την δομή και το νόημα των προτάσεων σε μια γλώσσα. Ιστορικά, όλες οι προσπάθειες για την μοντελοποίηση της γλώσσας εμπίπτουν σε δύο κατηγορίες. Τα πιο παλιά και γνωστά μοντέλα είναι οι γραμματικές που έχουν αναπτυχθεί στο πεδίο της γλωσσολογίας. Τα τελευταία χρόνια, τα στατιστικά μοντέλα που χρησιμοποιήθηκαν στον τομέα της αναγνώρισης φωνής έχουν κερδίσει το ενδιαφέρον των ερευνητών. Σε αυτή την εργασία γίνεται εκτεταμένη αναφορά σε αυτά τα στατιστικά γλωσσικά μοντέλα και στις εφαρμογές που μπορεί να έχουν σε θέματα αναγνώρισης φωνής και διόρθωσης κειμένων. Στην ενότητα αυτή θα παρουσιάσουμε τις δύο κατηγορίες μοντελοποίησης της γλώσσας.

Παραδοσιακά, η κάθε γλώσσα μοντελοποιείται με την χρήση της γραμματικής. Στην γλωσσολογία έχει παρατηρηθεί ότι η γλώσσα δομείται κατά κάποιον τρόπο ιεραρχικά και επιπλέον είναι διαθέσιμος ένας μικρός αριθμός πρωταρχικών δομικών μονάδων που μπορούν να συνδυαστούν με συγκεκριμένους τρόπος ώστε να δημιουργήσουν διάφορους τύπους που

εμφανίζονται στην γλώσσα. Σε πολύ χαμηλό επίπεδο της γραπτής γλώσσας υπάρχουν τα γράμματα. Τα γράμματα συνδυάζονται για να φτιάξουν λέξεις. Οι λέξεις με την σειρά τους ενώνονται για να συνθέσουν φράσεις όπως είναι μια ονοματική φράση π.χ «Θεολόγος Αθανασέλης» ή «ένα πλοίο», ή μια προθετική φράση π.χ «κάτω από το τραπέζι». Στην συνέχεια οι φράσεις μπορούν να δημιουργήσουν προτάσεις, που με την σειρά τους μπορούν να φτιάξουν παραγράφους και ούτω κάθε εξής.

Οι γραμματικές μπορούν να χρησιμοποιηθούν για να περιγράψουν την ιεραρχική δομή με ένα συνοπτικό τρόπο (Chomsky, 1964). Μια γραμματική περιλαμβάνει κανόνες που περιγράφουν τους επιτρεπτούς τρόπους συνδυασμού δομών σε ένα επίπεδο για να φτιάξουν δομές σε ένα ανώτερο επίπεδο. Για παράδειγμα ένα κανόνας γραμματικής μπορεί να έχει την εξής μορφή:

ονοματική φράση→άρθρο ουσιαστικό

που για χάριν συντομίας μπορεί να γραφτεί σαν ονοματική φράση (Ο.Φ)

(Ο.Φ)→(Άρθρο)Α (Ουσιαστικό) Ο

αυτό σημαίνει ότι κάθε άρθρο που προηγείται από ένα ουσιαστικό μπορεί να συνθέσει μια ονοματική φράση. Συνδυάζοντας τον προηγούμενο κανόνα με τους επόμενους

άρθρο→ ο | το

ουσιαστικό→πλοίο | γάτα | δένδρο

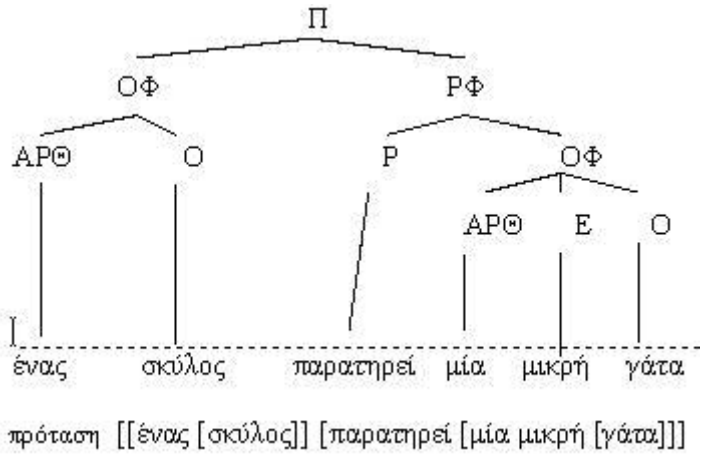
μπορούμε να φτιάξουμε προτάσεις του τύπου όπως το πλοίο, το δένδρο που αποκαλούνται ονοματικές φράσεις (Ο.Φ).

Μια γραμματική είναι μια συλλογή από κανόνες που περιγράφουν πώς θα φτιαχτούν υψηλού επιπέδου δομές, σαν τις προτάσεις, από χαμηλού επιπέδου δομές, σαν τις λέξεις. Χρησιμοποιώντας αυτού του είδους την αναπαράσταση κάποιος μπορεί να περιγράψει ένα σύνολο προτάσεων σε μια γλώσσα. Σε αυτήν την κατεύθυνση έχει γίνει πολύ δουλειά από τους γλωσσολόγους αν και χρησιμοποιούν πιο πλούσιες γραμματικές αναπαραστάσεις από αυτές που μόλις περιγράψαμε.

Αυτού του είδους οι γραμματικές έχουν ευρεία εφαρμογή στο πεδίο της επεξεργασίας φυσικής γλώσσας. Η επεξεργασία φυσικής γλώσσας έχει σαν στόχο να φτιάξει αυτόματα συστήματα για την καλύτερη επεξεργασία της. Για παράδειγμα ένας από τους στόχους της κατανόησης φυσικής γλώσσας είναι να δημιουργηθεί ένα σύστημα που να αντιλαμβάνεται το ερώτημα ενός χρήστη σχετικά με το «Ποια είναι η πρωτεύουσα του νομού Λέσβου» αντί για την ερώτηση που μοιάζει σαν το «Πρωτεύουσα/Λέσβου».

Οι γραμματικές είναι πολύ σημαντικά μοντέλα της γλώσσας στον τομέα της επεξεργασίας φυσικής γλώσσας επειδή παρέχουν πληροφορία για την δομή μιας πρότασης με

σκοπό τον καθορισμό των νοημάτων. Σε ένα σύστημα το πρώτο βήμα για την επεξεργασία μιας πρότασης είναι η συντακτική ανάλυση μιας πρότασης ώστε να παραχθεί το συντακτικό δένδρο. Το συντακτικό δένδρο δείχνει ποιους κανόνες πρέπει να χρησιμοποιήσουμε σε μια γραμματική για να συνθέσουμε υψηλού επιπέδου δομή, δηλ. μια πρόταση, από χαμηλότερου επιπέδου δομές, όπως οι λέξεις. Στο Σχήμα 1.2 φαίνεται το συντακτικό δένδρο μιας πρότασης, «*Ένας σκύλος παρατηρεί μία μικρή γάτα*». Στο συντακτικό δένδρο οι ανώτεροι 3 κόμβοι αναπαριστούν τις εφαρμογές των κανόνων.



Σχήμα 1.2 Η συντακτική ανάλυση μιας πρότασης.

Οι λέξεις μιας πρότασης που ανήκουν στον ίδιον κόμβο σε ένα συντακτικό δένδρο πρόκειται να αντιστοιχούν σε μονάδες που είναι σχετικές στο καθορισμό του νοήματος της πρότασης και καλούνται σαν συστατικά. Για παράδειγμα οι φράσεις «*Ένας σκύλος*», «*μία μικρή γάτα*» και «*Ένας σκύλος κοιτάζει μία μικρή γάτα*» αποτελούν συστατικά ενώ «*παρατηρεί μία*» δεν είναι. Ένα ακόμη παράδειγμα είναι η φράση «ο Νίκος πέταξε την μπάλα πίσω από τον φράχτη» που ενώ η φράση «πίσω από τον φράχτη» περιέχει νόημα δεν είναι συστατικό επειδή δεν περιγράφει την μπάλα αλλά την ενέργεια «πέταξε». Μ' αυτόν τον τρόπο γίνεται κατανοητό πώς τα γραμματικά μοντέλα καθορίζουν ποιες προτάσεις ανήκουν σε μια γλώσσα αλλά επίσης και το νόημα που κρύβεται μέσα από μια ακολουθία λέξεων μιας γλώσσας.

Ενώ οι γραμματικές αποτελούν ένα βασικό εργαλείο μοντελοποίησης της γλώσσας, είναι γενικά αποδεκτό ότι για να φτιάξουμε γραμματικές που θα μοντελοποιούν μη ελεγχόμενες γλώσσες είναι πολύ μακρινό. Αντιθέτως το ενδιαφέρον έχει μεταφερθεί σε εφαρμογές όπου δεν χρειάζεται τόσο λεπτομερειακή μοντελοποίηση γλώσσας. Μια από αυτές τις εφαρμογές που έχουν προσελκύσει το ενδιαφέρον των ερευνητών είναι η αναγνώριση φωνής, δηλ. εφαρμογές που μεταγράφουν την ομιλία του χρήστη σε γραπτό κείμενο.

Στην αναγνώριση φωνής το γλωσσικό μοντέλο χρησιμοποιείται για να βοηθήσει στην αποσαφήνιση αμφίσημων λέξεων. Έχει αποδειχθεί ότι τα στατιστικά μοντέλα που φέρουν πληροφορία για την συχνότητα των λέξεων ή συνδυασμών τους σε μια γλώσσα παίζουν σημαντικό ρόλο στην απόδοση ενός συστήματος αναγνώρισης φωνής. Τα στατιστικά γλωσσικά μοντέλα λειτουργούν αποδοτικά και είναι πολύ διαφορετικά από τις καθιερωμένες γραμματικές. Ένα παράδειγμα στατιστικού μοντέλου της γλώσσας είναι το διγραμμικό μοντέλο. Κάθε πιθανότητα $p(w_i \mid w_{i-1})$ έχει σαν στόχο να προβάλει πόσο συχνά εμφανίζεται η λέξη w_i μετά την λέξη w_{i-1} και αυτές οι πιθανότητες υπολογίζονται από την στατιστική επεξεργασία μεγάλου όγκου κειμένων. Υπάρχουν μεγάλες διαφορές μεταξύ των στατιστικών μοντέλων που χρησιμοποιήθηκαν στην αναγνώριση φωνής και στα γραμματικά μοντέλα που έχουν χρησιμοποιηθεί από γλωσσολόγους στην επεξεργασία φυσικής γλώσσας, πέραν από τις ανόμοιες αναπαραστάσεις τους. Στην γλωσσολογία αυτό που επιχειρείται είναι να φτιαχτούν γραμματικές που να ανταποκρίνονται επακριβώς σε ένα σύνολο από γραμματικές προτάσεις. Στην αναγνώριση φωνής αυτό που επιχειρείται είναι να μοντελοποιηθούν πόσο συχνά εμφανίζονται συγκεκριμένες ακολουθίες λέξεων ανεξάρτητα από το αν είναι σωστές από γραμματικής άποψης. Στην γλωσσολογία και στην επεξεργασία φυσικής γλώσσας αυτό που ενδιαφέρει είναι η δημιουργία ενός συντακτικού δένδρου που να φανερώνει το νόημα των προτάσεων. Στην αναγνώριση φωνής δεν απαιτείται καθόλου δομική ανάλυση μιας πρότασης ή κάποια επιπλέον επεξεργασία της γλώσσας. Στην γλωσσολογία, τα μοντέλα που αναπτύσσονται δεν βασίζονται στην συχνότητα συνεμφάνισης των λέξεων (μονογραμμάτων, διγραμμάτων και τριγραμμάτων) όπως συμβαίνει στην αναγνώριση φωνής.

Τέλος οι γραμματικές που χρησιμοποιούνται στην γλωσσολογία έχουν υλοποιηθεί χειρονακτικά. Ένας γλωσσολόγος συνήθως σχεδιάζει γραμματικές χωρίς αυτοματοποιήσεις. Στην αντίθετη περίπτωση τα μοντέλα στην αναγνώριση φωνής δημιουργούνται λαμβάνοντας υπόψη τα στατιστικά στοιχεία μεγάλου όγκου κειμένων. Τέτοιου είδους μοντέλα περιλαμβάνουν διάφορες πιθανότητες που πρέπει να εκτιμηθούν και συνεπώς η εκτίμηση είναι εφικτή μόνο με πρακτικό τρόπο μέσω της αυτοματοποιημένης ανάλυσης κειμένων που διατίθενται στο διαδίκτυο ή σε ηλεκτρονική μορφή.

1.4 Οι άξονες της διατριβής

Ο σκοπός αυτής της διατριβής είναι να εφαρμόσει τα μοντέλα των N-grams στην αναγνώριση συναισθηματικού λόγου και στην διόρθωση κειμένων με συντακτικά λάθη. Με την μέχρι τώρα βιβλιογραφία αποδεικνύεται η επιτυχία των N-grams σε πολλές εκφάνσεις της γλωσσικής τεχνολογίας. Με προεξέχουσα την χρήση των N-grams στην αναγνώριση συνεχούς λόγου

μεγάλου λεξιλογίου. Έτσι λοιπόν αναλογιζόμενοι την δυναμική των N-grams σε αντίστοιχους τομείς επιδιώκουμε να εμπλουτίσουμε τα N-grams με συναισθηματικά δεδομένα ώστε να μοντελοποιήσουμε την γλωσσική δομή του συναισθηματικού λόγου και να βελτιώσουμε την απόδοση ενός συμβατικού συστήματος αναγνώρισης φωνής όταν χρησιμοποιείται αντίστοιχος λόγος.

Οι φωνητικές διεπαφές με τους υπολογιστές είναι ένας τομέας που διεγείρει και εντυπωσιάζει όλους όσους ασχολούνται με την επεξεργασία της φυσικής γλώσσας, εδώ και δεκαετίες. Για πολλούς, η δυνατότητα ενός χρήστη να συνομιλεί ελεύθερα με μια μηχανή αποτελεί την κορυφή της πυραμίδας που στην βάση της βρίσκεται η κατανόηση της διαδικασίας παραγωγής και αντίληψης του λόγου. Μην ξεχνάμε ότι οι διαδικασίες αυτές αποτελούν βασικές δομές της ανθρώπινης επικοινωνίας. Στην σημερινή εποχή τέτοιου είδους διεπαφές που λαμβάνουν υπόψη τους το συναίσθημα του ομιλητή αποτελούν αναγκαιότητα και όχι πολυτέλεια. Το συναίσθημα στον λόγο είναι ένας τομέας που λαμβάνει όλο και μεγαλύτερο ενδιαφέρον τα τελευταία χρόνια, τόσο σε επίπεδο σύνθεσης φωνής όσο και σε επίπεδο της αυτόματης αναγνώρισης φωνής. Ο στόχος είναι να σχεδιαστούν συστήματα που αντιλαμβάνονται τα ανθρώπινα συναισθήματα. Για να επιτευχθεί αυτό είναι απαραίτητο να δημιουργηθούν συστήματα αναγνώρισης φωνής που θα μπορούν να χειριστούν το συναισθηματικό λόγο. Στην διατριβή αυτή εξετάζεται μια εναλλακτική στρατηγική αντιμετώπισης του προβλήματος της αναγνώρισης συναισθηματικού λόγου. Η όλη ιδέα αυτής της στρατηγικής είναι ότι το συναίσθημα ενός ομιλητή δεν εκφράζεται μόνο από την χροιά του λόγου του αλλά από τις λέξεις που χρησιμοποιεί και από την σειρά με την οποία τις διατυπώνει. Για αυτό τον λόγο τίθεται προς διερεύνηση, η βελτίωση της απόδοσης συμβατικών συστημάτων αναγνώρισης φωνής με την ενσωμάτωση ενός ειδικού γλωσσικού μοντέλου προσανατολισμένο σε δεδομένα με συναισθηματικό βάρος. Η κατασκευή του γλωσσικού μοντέλου βασίζεται, στα μοντέλα N-grams και τα αποτελέσματα χρήσης αυτής της μεθόδου φαίνονται να είναι ενθαρρυντικά.

Στο υπόλοιπο μισό της εργασίας αυτής και παίρνοντας αφορμή από τα επιτυχή αποτελέσματα της χρήσης των N-grams στην μοντελοποίηση της γλώσσας θα παρουσιάσουμε ένα νέο αλγόριθμο εφαρμογής των N-grams στην διόρθωση προτάσεων αναδιατάσσοντας τις λέξεις. Το κίνητρο της παρούσας διατριβής είναι να διορθωθούν προτάσεις με λάθη στην σειρά των λέξεων με την χρήση αποκλειστικά και μόνο στατιστικών μεθόδων. Ο προτεινόμενος αλγόριθμος έχει στον πυρήνα την μέθοδο της γρήγορης αναζήτησης της βέλτιστης λύσης. Με την μέθοδο αυτή περιορίζεται ο χώρος αναζήτησης της βέλτιστης λύσης αφού χρησιμοποιούνται μόνο τα έγκυρα ζεύγη λέξεων. Μέχρι τώρα η ανίχνευση προτάσεων με λάθη στην σειρά των λέξεων γίνεται με την χρήση συντακτικών αναλυτών ή με την χρήση αρνητικών

προτύπων (εκμάθηση από κείμενα με λάθη). Τα αποτελέσματα της χρήσης τέτοιων μεθόδων είναι θετικά και παρουσιάζουν υψηλό δείκτη αξιοπιστίας χωρίς όμως να αποδίδουν το ίδιο καλά και στον τομέα της διόρθωσης αυτών των λαθών. Όπως γίνεται κατανοητό η γραφή μιας πρότασης που δεν ακολουθεί τους επιτρεπτούς κανόνες διάταξης των λέξεων δημιουργεί σημαντικά προβλήματα στην κατανόηση της. Άρα η διαδικασία της διόρθωσης αυτών των προτάσεων είναι επιβεβλημένη αρκεί να μπορεί να δώσει τις λέξεις της συγκεκριμένης πρότασης στην σωστή σειρά, με το σωστό νόημα και σε άμεσο χρόνο.

Το μεγαλύτερο πλεονέκτημα αυτής της μεθόδου έγκειται στην διόρθωση κειμένων χωρίς την χρήση συντακτικού αναλυτή και αυτό γιατί ένας τέτοιος αναλυτής δεν είναι υπαρκτός σε κάθε γλώσσα και η δημιουργία του απαιτεί χρόνο και κόπο. Αντίθετα η μέθοδος αυτή λύνει το πρόβλημα της έλλειψης συντακτικών αναλυτών με την χρήση υφιστάμενων γλωσσικών μοντέλων. Η κατασκευή γλωσσικού μοντέλου είναι εφικτή για κάθε γλώσσα από την στιγμή που υπάρχουν τα απαραίτητα λογισμικά και οι μεγάλες συλλογές γραπτών κειμένων.

1.5 Οργάνωση της διατριβής

Το υπόλοιπο της εργασίας οργανώνεται ως εξής: στο δεύτερο μέρος της εργασίας αυτής περιλαμβάνεται μια σύντομη επισκόπηση σε μεθόδους στατιστικής μοντελοποίησης με κυριότερο εκφραστή τα μοντέλα N-grams. Τα προβλήματα που δημιουργούνται κατά την δημιουργία των N-grams μοντέλων αφορούν την αδυναμία αυτών να μοντελοποιήσουν φαινόμενα του γραπτού λόγου με απόσταση μεγαλύτερη από 3 λέξεις (λόγω χρήσης τριγραμμάτων) και η σπανιότητα μερικών δεδομένων εκπαίδευσης. Έτσι το υπόλοιπο του δευτέρου μέρους επικεντρώνεται σε μεθόδους εξομάλυνσης του γλωσσικού μοντέλου.

Γνωρίζοντας την ευρύτητα της χρήσης του γλωσσικού μοντέλου γίνεται προσπάθεια ώστε αυτά να προσαρμοστούν και να χρησιμοποιηθούν με το κατάλληλο τρόπο σε δυο νέες εφαρμογές όπως η αναγνώριση του συναισθηματικού λόγου και η διόρθωση προτάσεων με λανθασμένη σειρά των στοιχείων-λέξεων τους. Τα επόμενα κεφάλαια, 3 και 4, περιγράφουν τους αλγορίθμους εφαρμογής του γλωσσικού μοντέλου στις δυο αυτές μορφές επεξεργασίας φυσικής γλώσσας. Στο $3^{ο}$ κεφάλαιο, δίνεται έμφαση στην προσαρμογή των γλωσσικών μοντέλων σε κείμενα με συναισθηματικό βάρος και περιγράφεται ο αλγόριθμος επαύξησης του γλωσσικού μοντέλου με δεδομένα που είναι προσανατολισμένα στην έκφραση συναισθημάτων. Τα αποτελέσματα και η βελτίωση που αποφέρει η χρήση αυτών των προσαρμοσμένων γλωσσικών μοντέλων περιγράφονται στο τέλος του τρίτου μέρους μαζί και η εφαρμογή της αναγνώρισης του συναισθηματικού λόγου σε εξελιγμένες διεπαφές ανθρώπου-μηχανής.

Στο τέταρτο κεφάλαιο περιγράφεται η χρήση του γλωσσικού μοντέλου στην διόρθωση προτάσεων με λάθη στην σειρά των λέξεων. Στην συγκεκριμένη περίπτωση το γλωσσικό

μοντέλο περιλαμβάνει N-grams με διγράμματα και τριγράμματα. Η αξιολόγηση του μεθόδου γίνεται με την χρήση διαφορετικών δεδομένων δοκιμής για την Αγγλική και την Ελληνική γλώσσα. Στην τελική ενότητα αξιολόγησης της μεθόδου παρουσιάζεται μια πειραματική άσκηση αναδιάταξης λέξεων που εκπονήθηκε με την αρωγή διαφόρων χρηστών, με σκοπό την αξιολόγηση των αποτελεσμάτων της μεθόδου.

Στο πέμπτο κεφάλαιο συνοψίζονται τα αποτελέσματα χρήσης των N-grams στους τομείς αναγνώρισης φωνής συναισθηματικού λόγου και διόρθωσης κειμένων και μαζί περιγράφονται τα θέματα που μπορούν να διερευνηθούν περαιτέρω.

Κεφάλαιο 2

2 Στατιστικές μέθοδοι μοντελοποίησης της γλώσσας

2.1 Εισαγωγή

Η στατιστική επεξεργασία της φυσικής γλώσσας αποσκοπεί να δώσει στατιστικό συμπέρασμα στην φυσική γλώσσα. Και όταν λέμε στατιστικό συμπέρασμα σημαίνει να λάβουμε κάποια δεδομένα και να προσπαθήσουμε να βγάλουμε συμπεράσματα για την κατανομή τους. Σε αυτό το κεφάλαιο θα εξετάσουμε την στατιστική μοντελοποίηση της γλώσσας, όπου το πρόβλημα είναι να προβλέψουμε την επόμενη λέξη βάσει των προηγούμενων. Αυτό το αντικείμενο είναι ιδιαίτερα διαδεδομένο στο τομέα της αναγνώρισης φωνής, στην οπτική αναγνώριση χαρακτήρων, στην μηχανική μετάφραση και ακόμη στην ορθογραφική διόρθωση. Κατά κάποιο τρόπο αυτή η διαδικασία είναι ανάλογη του παιχνιδιού Shannon (1951) που έχει σαν στόχο την πρόβλεψη του επόμενου γράμματος μέσα σε ένα κείμενο.

2.2 Στατιστικό γλωσσικό μοντέλο

Ο στόχος του στατιστικού γλωσσικού μοντέλου είναι να συγκεντρώσει του κανόνες της φυσικής γλώσσας με σκοπό να βελτιωθεί η απόδοση των εφαρμογών φυσικής γλώσσας. Η στατιστική μοντελοποίηση της γλώσσας επιτυγχάνεται με την εκτίμηση της κατανομής πιθανότητας διαφόρων γλωσσικών μονάδων όπως λέξεις και προτάσεις.

Ο ρόλος του στατιστικού γλωσσικού μοντέλου είναι πολύ σημαντικός για μια σειρά από εφαρμογές της γλωσσικής τεχνολογίας. Αυτές περιλαμβάνουν την αναγνώριση φωνής (όπου και το στατιστικό γλωσσικό μοντέλο ήταν ο μοχλός της εξέλιξης της), την μηχανική μετάφραση, την κατάταξη των εγγράφων και την δρομολόγηση τους, την οπτική αναγνώριση χαρακτήρων, την ανάκτηση πληροφοριών, την αναγνώριση χειρόγραφων, την διόρθωση των ορθογραφικών λαθών και πολλές άλλες εφαρμογές.

Το στατιστικό γλωσσικό μοντέλο χρησιμοποιεί τεχνικές στατιστικής εκτίμησης γλωσσικών δεδομένων εκπαίδευσης, που εφαρμόζονται σε εκτεταμένα κείμενα. Λόγω της

φύσης της γλώσσας και των μεγάλων λεξιλογίων που χρησιμοποιούν στην πράξη οι άνθρωποι, οι στατιστικές τεχνικές για να είναι αποδοτικές οφείλουν να εφαρμόζονται σε μεγάλο όγκο κειμένων.

Εδώ και είκοσι χρόνια διάφορα είδη σωμάτων κειμένων είναι διαθέσιμα στο διαδίκτυο προς ελεύθερη χρήση. Αυτό είχε σαν αποτέλεσμα, σε τομείς όπου υπήρχαν δεδομένα, η ποιότητα του γλωσσικού μοντέλου να βελτιωθεί σημαντικά. Εντούτοις, η βελτίωση αυτή ξεκίνησε να κινείται ασυμπτωτικά. Αυτό που παρατηρείται είναι ότι, αν και ο όγκος των σωμάτων κειμένων αυξάνεται με εκθετικό τρόπο (φυσικά αυτό εξαρτάται και από την πρόοδο του παγκόσμιου ιστού), η ποιότητα των στατιστικών γλωσσικών μοντέλων μοιάζει να μην βελτιώνεται σημαντικά. Κατά μια ανεπίσημη εκτίμηση από ερευνητές της IBM φαίνεται ότι τα στατιστικά γλωσσικά μοντέλα που βασίζονται σε διγράμματα μπορούν να προσδιοριστούν από σώματα κείμενων της τάξης των 100 εκατομμυρίων λέξεων.

Είναι όμως αντιφατικό ότι οι πιο επιτυχημένες τεχνικές προσδιορισμού στατιστικού γλωσσικού μοντέλου δεν χρησιμοποιούν καθόλου πληροφορία για το τι είναι στην πραγματικότητα η γλώσσα. Τα πιο δημοφιλή γλωσσικά μοντέλα (N-grams) δεν λαμβάνουν υπόψη τους καθόλου το γεγονός ότι αυτό που μοντελοποιούν είναι η ίδια η γλώσσα, θεωρώντας τις προτάσεις ως μια ακολουθία απλών συμβόλων που τυγχάνει να είναι λέξεις, χωρίς να συνυπολογίζεται η δομή και πληροφορία που φέρουν οι λέξεις καθαυτές.

Μια πιθανή εξήγηση για αυτήν την κατάσταση έχει να κάνει με την φτωχή γνώση της γλώσσας, και με το γεγονός ότι οι τεχνικές των N-grams έχουν ουσιαστική αποτελεσματικότητα, αποτρέποντας του ερευνητές με το να ασχοληθούν σοβαρά με την γνώση της γλώσσας.

Αλλά το ερώτημα που δημιουργείται είναι πόσο μακριά μπορεί κάποιος να φθάσει χωρίς να έχει βαθιά γνώση της γλώσσας; Όπως αναφέρει ο Jelinek (1995) πρέπει η γλώσσα να βρει την θέση της στο γλωσσικό μοντέλο. Δυστυχώς μόνο λίγες και μετρημένες στα δάχτυλα προσπάθειες έχουν γίνει μέχρι σήμερα ώστε να ενσωματωθεί η δομή και η γνώση της γλώσσας στον υπολογισμό του γλωσσικού μοντέλου.

2.3 Ορισμός και χρήση

Τα στατιστικά μοντέλα της γλώσσας δεν έχουν εφαρμογή μόνο στην αναγνώριση φωνής αλλά είναι εξίσου χρήσιμα και σημαντικά και σε άλλες εφαρμογές όπως π.χ η ορθογραφική διόρθωση και η μηχανική μετάφραση. Αυτές οι εφαρμογές όπως και άλλες μπορούν να ενταχθούν στο πλαίσιο του μοντέλου πηγής καναλιού που χρησιμοποιείται στην θεωρία πληροφοριών (Shannon, 1948).

Για την αναγνώριση φωνής ισχύει ότι: για ένα ακουστικό σήμα A που αντιστοιχεί σε μια πρόταση, χρειάζεται να βρούμε την πιο πιθανή μετεγγραφή T, ώστε

$$T = \arg_T \max P(T \mid A) \tag{2.1}$$

Εφαρμόζοντας τον κανόνα Bayes μπορούμε να γράψουμε την προηγούμενη σχέση ως εξής

$$T = \arg_T \max \frac{P(A \mid T)P(T)}{P(A)} = \arg_T \max P(A \mid T)P(T) \tag{2.2}$$

Η κατανομή πιθανότητας $P(T)$ καλείται γλωσσικό μοντέλο και περιγράφει πόσο πιθανή ή συχνά εμφανιζόμενη είναι η πρόταση T σε μια γλώσσα. Η κατανομή $P(A \mid T)$ αναπαριστά το ακουστικό μοντέλο και περιγράφει ποιο από τα ακουστικά σήματα A είναι πιθανόν να προέλθουν από την πρόταση T.

Το μοντέλο πηγής-καναλιού στην θεωρία πληροφοριών περιγράφει το πρόβλημα της ανάκτησης πληροφορίας που έχει μεταδοθεί μέσω ενός θορυβώδους καναλιού. Έτσι σε αυτό το πλαίσιο υπάρχει το μοντέλο της πληροφορίας της πηγής $P(I)$ και το μοντέλο του θορυβώδους καναλιού $P(O \mid I)$ που περιγράφει την πιθανότητα εμφάνισης της εξόδου O ενός καναλιού με είσοδο I, (Σε ένα τέλειο κανάλι θα μπορούσαμε να γράψουμε ότι $P(O \mid I) = 1$ όταν $O = I$ και $P(O \mid I) = 0$ σε κάθε άλλη περίπτωση). Ο στόχος είναι να επανακτήσουμε το μήνυμα I που εστάλη μέσω του θορυβώδους καναλιού βάσει της εξόδου O. Άρα η διαδικασία αυτή έχει σαν στόχο να βρεθεί το μήνυμα I με την μεγαλύτερη πιθανότητα για μια συγκεκριμένη έξοδο O.

Έτσι μπορούμε να γράψουμε ότι:

$$I = \arg_I \max P(I \mid O) = \arg_I \max \frac{P(O \mid I)P(I)}{P(O)} = \arg_I \max P(O \mid I)P(I) \tag{2.3}$$

Το μοντέλο πηγής-καναλιού μπορεί να επεκταθεί και σε άλλες εφαρμογές αλλάζοντας μόνο το μοντέλο του καναλιού (Brown et al., 1992). Στην οπτική αναγνώριση χαρακτήρων και στην αναγνώριση χειρογράφων (Hull, 1992; Srihari and Baltus, 1993) το κανάλι μπορεί να ερμηνευτεί ως η διαδικασία μετατροπής του κειμένου σε εικόνα αντί για την διαδικασία μετατροπής κειμένου σε ομιλία, και η εξίσωση που περιγράφει την διαδικασία αυτή είναι η εξής

$$T = \arg_T \max P(T)P(image \mid T) \tag{2.4}$$

Στην περίπτωση της ορθογραφικής διόρθωσης (Kernighan et al., 1990), το κανάλι μπορεί να ερμηνευτεί σαν μια δακτυλογράφο που γράφοντας ένα κείμενο T κάνει λάθη και παράγει τελικώς ένα άλλο θορυβώδες κείμενο T_n με ορθογραφικά λάθη, ώστε να ισχύει

$$T = \arg_T \max P(T)P(T_n \mid T) \tag{2.5}$$

Στην περίπτωση της μηχανικής μετάφρασης (Brown et al., 1990), το κανάλι μπορεί να ερμηνευτεί σαν μια μεταφράστρια που μεταφράζει ένα κείμενο T από μια γλώσσα σε ένα άλλο κείμενο T_f μιας ξένης γλώσσας, ώστε να ισχύει

$$T = \arg_T \max P(T \mid T_f) = \arg \max P(T_f \mid T)P(T) \tag{2.6}$$

Σε κάθε μια περίπτωση η προσπάθεια είναι να ανακτηθεί το αρχικό κείμενο T βάσει της εξόδου του θορυβώδους καναλιού, είτε όταν το κανάλι δίνει σαν έξοδο εικόνα είτε κείμενο με ορθογραφικά λάθη είτε κείμενο ξένης γλώσσας. Πρέπει να τονιστεί ότι σε όλες τις εφαρμογές ο στόχος είναι να δημιουργηθεί ένα γλωσσικό μοντέλο $P(T)$.

2.3.1 Γνωστές αδυναμίες των υπαρχόντων μοντέλων

Ακόμη και τα πιο απλά γλωσσικά μοντέλα έχουν καταλυτική επίδραση στην εφαρμογή από την οποία χρησιμοποιούνται (αυτό μπορεί να διαπιστωθεί από την αφαίρεση του γλωσσικού μοντέλου από την μηχανή αναγνώρισης φωνής). Παρόλ' αυτά τα υπάρχοντα γλωσσικά μοντέλα είναι άκρως ευαίσθητα σε αλλαγές θεματολογίας και ύφους των κειμένων που χρησιμοποιούνται για την εκπαίδευση τους. Για παράδειγμα, για να μοντελοποιηθούν καθημερινές τηλεφωνικές συνομιλίες είναι προτιμότερο να χρησιμοποιηθούν 2 εκατομμύρια λέξεις από μετεγγραφές τέτοιων συνομιλιών παρά 140 εκατομμύρια λέξεις από μετεγγραφές τηλεοπτικών και ραδιοφωνικών εκπομπών. Η επίδραση αυτή είναι τόσο ισχυρή ακόμη και σε αλλαγές που είναι επουσιώδεις για την ανθρώπινη αντίληψη. Ένα γλωσσικό μοντέλο που έχει εκπαιδευτεί με την χρήση κειμένων από τα δελτία τύπου του Dow Jones θα παρουσιάσει διπλάσια δυσκολία επιτυχούς αναγνώρισης όταν εφαρμοσθεί σε ένα άλλο σώμα κειμένου όπως αυτό των δελτίων τύπου του Associated Press.

2.4 Ποιοτικά μεγέθη αξιολόγησης γλωσσικών μοντέλων

Ο στόχος του γλωσσικού μοντέλου είναι να προβλέπει τις λέξεις σε μια ακολουθία, και ένα γλωσσικό μοντέλο είναι καλό όταν αποτελεί καλό προβλέπτη μιας λέξης σε κάθε δυνατή θέση βάσει των λέξεων που έχουν ήδη παρατηρηθεί. Για μια ακολουθία N λέξεων

$W = \{w_1, w_2, ..., w_N\}$ σε μια βάση δεδομένων εκπαίδευσης, η τιμή $P(W)$ για αυτήν την ακολουθία παρέχει πληροφορία για το πόσο ικανοποιητικά μπορεί να προβλεφθεί αυτή η ακολουθία λέξεων από το συγκεκριμένο γλωσσικό μοντέλο. Όσο μεγαλύτερη είναι η τιμή του $P(W)$ τόσο καλύτερο είναι το γλωσσικό μοντέλο στην πρόβλεψη λέξεων. Η εντροπία και η περιπλοκή είναι οι πιο συχνές μετρικές που χρησιμοποιούνται στην αξιολόγηση των N-grams μοντέλων. Στις 2 παρακάτω ενότητες περιγράφονται οι δυο αυτές μετρικές.

2.4.1 Εντροπία

Η εντροπία μπορεί να χρησιμοποιηθεί σαν ένα εργαλείο μέτρησης περί του πόση πληροφορία υπάρχει σε μια ειδική γραμματική, για το πόσο καλά μια γραμματική μοντελοποιεί την γλώσσα και πόσο καλά μπορεί να προβλεφθεί η επόμενη λέξη βάσει των μοντέλων N-grams. Δίνοντας δυο γραμματικές και ένα σώμα κειμένου, μπορούμε να μετρήσουμε την εντροπία ώστε να μας υποδείξει ποια γραμματική ταιριάζει καλύτερα στο σώμα κειμένου. Επίσης μπορούμε να χρησιμοποιήσουμε την εντροπία για να συγκρίνουμε πόσο δύσκολα είναι δυο αντικείμενα αναγνώρισης φωνής και επιπλέον να μετρήσουμε πόσο καλά μια στατιστική γραμματική μπορεί να ταιριάξει με την γραμματική που χρησιμοποιούν οι άνθρωποι.

Ο υπολογισμός της εντροπίας απαιτεί να διασφαλιστεί μια τυχαία μεταβλητή x η οποία αντιστοιχεί σε αυτό που θα περιγράψουμε (λέξεις, γράμματα, μέρη του λόγου, όλα αυτά ανήκουν σε ένα σύνολο χ) η οποία έχει μια ιδιαίτερη συνάρτηση πιθανότητας $p(x)$. Η εντροπία λοιπόν της τυχαίας μεταβλητής αυτής είναι

$$H(X) = -\sum_{x \in \chi} p(x) \cdot \log_2 p(x) \tag{2.7}$$

Το αποτέλεσμα της εντροπία μετριέται σε bits. Ο πιο διαισθητικός τρόπος για να οριστεί η εντροπία είναι να θεωρηθεί η εντροπία σαν το κατώτερο άκρο του αριθμού των bits που απαιτούνται για να κωδικοποιηθεί ένα κομμάτι πληροφορίας.

Οι Cover και Thomas (1991) πρότειναν το ακόλουθο παράδειγμα. Ας υποθέσουμε ότι θέλουμε να στείλουμε ένα μικρό μήνυμα σε ένα booker ώστε να στοιχηματίσει σε ένα αγώνα ιπποδρομίας 8 άλογων που λαμβάνει χώρα στην Αγγλία.

Ένας τρόπος είναι να στείλουμε την δυαδική αναπαράσταση του αριθμού του συγκεκριμένου αλόγου. Έτσι το άλογο νούμερο 1 θα έχει σαν κωδικό το 001, το 2 το 010, το 3 το 011 και το 8 το 000. Κατά συνέπεια αντιλαμβανόμαστε ότι θα απαιτηθούν κατά μέσο όρο 3 bits/αγώνα ώστε να στοιχηματίζουμε καθ' όλη την διάρκεια της ημέρας. Το ερώτημα που τίθεται είναι αν μπορούμε να στείλουμε μήνυμα με λιγότερα bits; για αυτόν τον λόγο θα

χρησιμοποιήσουμε την πληροφορία που έχουμε για τα προγνωστικά του κάθε αγώνα αναλογιζόμενοι ότι κάθε άλογο έχει μια προϊστορία νικών στους αγώνες.

Άλογο 1	1/2	**Άλογο 5**	1/64
Άλογο 2	1/4	**Άλογο 6**	1/64
Άλογο 3	1/8	**Άλογο 7**	1/64
Άλογο 4	1/16	**Άλογο 8**	1/64

Πίνακας 2.1 Η πιθανότητα νίκης του κάθε αλόγου στον αγώνα.

Η εντροπία της τυχαίας μεταβλητής X δίνει ένα κατώτατο άκρο αριθμού bits και ισούται με:

$$\begin{aligned} H(X) &= -\sum_{i=1}^{i=8} p(i)\cdot \log_2 p(i) \\ &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - 4(\frac{1}{64}\log_2\frac{1}{64}) \\ &= -2bits \end{aligned} \quad (2.8)$$

Άρα ο κώδικας που μπορεί να χρησιμοποιηθεί με μέσο όρο 2 bits/ αγώνα μπορεί να υλοποιηθεί με την χρήση λιγότερων bits για τα πιο πιθανά άλογα και με περισσότερο bits για τα λιγότερο πιθανά. Στην περίπτωση όπου όλα τα άλογα είναι ίσο-πιθανά να κερδίσουν το κάθε αγώνα, το μήκος του δυαδικού κώδικα κάθε αλόγου θα έχει τιμή ίση με 3 bits λόγω του ότι η πιθανότητα κάθε αλόγου ισούται με 1/8 και συνεπώς από την σχέση (2.7) εξάγεται το συμπέρασμα ότι η εντροπία επιλογής του κάθε αλόγου θα ισούται με 3 bits.

Μέχρι τώρα υπολογίστηκε μόνο η εντροπία μιας τυχαίας μεταβλητής, το ερώτημα που γεννάται είναι τι θα συμβεί στην περίπτωση όπου θέλουμε να υπολογίσουμε την εντροπία για μια ακολουθία λέξεων και όχι για μιας μόνο λέξης. Για παράδειγμα θέλουμε να υπολογίσουμε την εντροπία της ακόλουθης σειράς λέξεων $W = \{w_1, w_2, ..., w_N\}$. Μπορούμε λοιπόν να υπολογίσουμε την εντροπία μια τυχαίας μεταβλητής που κυμαίνεται σε όλες τις πεπερασμένες ακολουθίες λέξεων μήκους b για κάποια γλώσσα L με την παρακάτω σχέση:

$$H(w_1, w_2 ... w_N) = -\sum_{W_1^N \in L} P(W_1^N)\cdot \log_2 P(W_1^N) \quad (2.9)$$

Επιπλέον μπορούμε να ορίσουμε τον ρυθμό της εντροπίας (ή πιο σωστά θα μπορούσαμε να γράψουμε την ανά λέξη εντροπία) σαν την εντροπία της ακολουθίας διαιρούμενη με τον αριθμό των λέξεων:

$$\frac{1}{N}H(W_1^N) = -\frac{1}{N}\sum_{W_1^N \in L} P(W_1^N)\cdot \log_2 P(W_1^N) \tag{2.10}$$

Αν θέλουμε να υπολογίσουμε την πραγματική εντροπία θα πρέπει να θεωρήσουμε ακολουθίες λέξεων μη πεπερασμένες. Έτσι αν θεωρήσουμε την γλώσσα σαν μια στοχαστική διεργασία L η οποία παράγει ακολουθίες λέξεων τότε η ανα λέξη εντροπία μπορεί να γραφτεί ως εξής:

$$H(L) = \lim_{N\to\infty}\frac{1}{N}H(w_1, w_2 ... w_N) \tag{2.11}$$

$$H(L) = \lim_{N\to\infty} -\frac{1}{N}\sum_{W_1^N \in L} P(w_1, w_2 ... w_N)\cdot \log_2 P(w_1, w_2 ... w_N) \tag{2.12}$$

Το θεώρημα των Shannon-McMillan-Breiman (Jurafsky and Martin, 2000) αποδεικνύει ότι όταν μια γλώσσα είναι στάσιμη ισχύει ότι,

$$H(L) = \lim_{N\to\infty} -\frac{1}{N}\log_2 P(w_1, w_2 ... w_N) \tag{2.13}$$

Το θεώρημα μπορεί να εξηγηθεί αναλογιζόμενοι ότι μια ακολουθία λέξεων είναι τόσο μεγάλη σαν να είναι άθροισμα μικρότερων προτάσεων. Η ιδέα του θεωρήματος είναι ότι μια μεγάλη ακολουθία λέξεων περιλαμβάνει πολλές μικρές και κάθε μικρή επαναλαμβάνεται στην μεγαλύτερη ακολουθία ανάλογα με την συχνότητα εμφάνισης της.

Για να ανακεφαλαιώσουμε μπορούμε να πούμε ότι υπολογίζουμε την εντροπία μια οποιασδήποτε στοχαστικής διεργασίας παίρνοντας ένα μεγάλο δείγμα της εξόδου και υπολογίζοντας τον μέσο όρο της λογαριθμικής της πιθανότητας. Στην θεωρία πληροφοριών αυτό που πραγματικά μετράμε είναι η διασταυρωμένη εντροπία (cross entropy) των δεδομένων εκπαίδευσης για ένα συγκεκριμένο μοντέλο. Επειδή η διασταυρωμένη εντροπία είναι το πιο γνωστό είδος εντροπίας, από εδώ και πέρα με τον όρο εντροπία θα αναφερόμαστε στην διασταυρωμένη εντροπία.

Η διασταυρωμένη εντροπία ορίζεται ως το ανώτατο όριο της εντροπίας, καθώς το μήκος μιας ακολουθίας λέξεων πάει προς το άπειρο. Άρα λοιπόν χρειάζεται μια προσέγγιση της διασταυρωμένης εντροπίας που θα σχετίζεται με μια ακολουθία λέξεων πολύ μεγάλη αλλά με πεπερασμένο μήκος. Αυτή η προσέγγιση της διασταυρωμένης εντροπίας ενός μοντέλου $P(w_i \mid w_{i-N+1} ... w_{i-1})$, για μια ακολουθία λέξεων W είναι:

$$H(W) = -\frac{1}{N}\log P(w_i \mid w_{i-N+1}...w_{i-1}) \tag{2.14}$$

Όταν συγκρίνονται δυο γλωσσικά μοντέλα ως προς ένα συγκεκριμένο σύνολο δεδομένων δοκιμής, το γλωσσικό μοντέλο που δίνει την μεγαλύτερη πιθανότητα $P(W)$, θεωρείται ως το πιο αντιπροσωπευτικό για να το μοντελοποιήσει. Παράλληλα το μοντέλο αυτό παρουσιάζει και την μικρότερη διασταυρωμένη εντροπία. Η διασταυρωμένη εντροπία $H(W)$ ενός γλωσσικού μοντέλου για ένα σύνολο δεδομένων δοκιμής W που περιλαμβάνει N λέξεις, ορίζεται ως εξής:

$$H(W) = -\frac{1}{N}\log P(W) \tag{2.15}$$

Και μπορεί να ερμηνευτεί σαν το μέσο όρο των bits που χρειάζονται για να κωδικοποιηθεί κάθε λέξη στο σύνολο δεδομένων δοκιμής με την χρήση του γλωσσικού μοντέλου.

2.4.2 Περιπλοκή-Perplexity

Η περιπλοκή ενός μοντέλου είναι το αντίστροφο του μέσου όρου της πιθανότητας που αντιστοιχεί στην κάθε λέξη στο σύνολο των δεδομένων δοκιμής, και σχετίζεται με την διασταυρωμένη εντροπία βάσει της εξίσωσης,

$$PP(W) = 2^{H(W)} \tag{2.16}$$

Ξαναγράφοντας την σχέση 2.15 έχουμε ότι

$$H(W) = -\frac{1}{N}\log_2 P(w_1 w_2 ... w_N) \tag{2.17}$$

με συνέπεια να έχουμε:

$$2^{H(W)} = P(w_1 w_2 ... w_N)^{-\frac{1}{N}} \tag{2.18}$$

Με την βοήθεια της σχέση 2.16 έχουμε ότι :

$$PP(W) = P(w_1 w_2 ... w_N)^{-\frac{1}{N}} \tag{2.19}$$

και παίρνοντας την N -ιοστή ρίζα

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1 w_2 ... w_N)}} \tag{2.20}$$

Με την χρήση του κανόνα της αλυσίδας μπροούμε να επεκτείνουμε την πιθανότητα της ακολουθίας W :

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i \mid w_1 ... w_{i-1})}} \tag{2.21}$$

Στην περίπτωση που θέλουμε να υπολογίσουμε την περιπλοκή μιας ακολουθίας W με ένα διγραμμικό γλωσσικό μοντέλο, τότε έχουμε:

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i \mid w_{i-1})}} \tag{2.22}$$

Για κάθε γλωσσικό μοντέλο, είναι δυνατόν να υπολογισθεί η περιπλοκή για κάποιο σώμα κειμένων που θα αναγνωρισθεί. Για τεχνητά περιορισμένα θέματα με αυστηρή σύνταξη η περιπλοκή μπορεί να μεταφραστεί σαν να είναι ισοδύναμη με το μέσο όρο των διαφορετικών λέξεων που απαιτείται να διαχωριστούν σε κάθε σημείο της ακολουθίας, εάν όλες οι λέξεις σε κάθε σημείο είναι ίσο-πιθανές. Η περιπλοκή πολλές φορές παριστάνει ένα μέσο συντελεστή διακλάδωσης εναλλακτικών λέξεων. Η μικρότερη τιμή περιπλοκής είναι ίση με το 1, αλλά αυτό θα μπορούσε να συμβεί μόνο όταν όλες οι πιθανές λέξεις είχαν πιθανότητα 1, με αποτέλεσμα μόνο αυτή η ακολουθία λέξεων να μπορούσε να αναγνωριστεί.

Σε άλλη ακραία περίπτωση όταν κάθε λέξη σε μια ακολουθία έχει πιθανότητα ίση με 0, τότε η πιθανότητα ολόκληρης της ακολουθίας θα έχει τιμή 0 και η περιπλοκή θα είναι άπειρα μεγάλη. Όπως είναι κατανοητό, η μεγαλύτερη πρόκληση για ένα στατιστικό μοντέλο φυσικής γλώσσας είναι να μην έχει μηδενικές πιθανότητες που σημαίνει ότι κάποιες λέξεις αποκλείονται και παράλληλα να υπάρχουν λίγες εναλλακτικές λέξεις με μεγάλη πιθανότητα σε κάθε σημείο. Ένα καλό γλωσσικό μοντέλο πρέπει να δίνει μικρή περιπλοκή σε μεγάλα σώματα κειμένων (κείμενα που δεν έχουν χρησιμοποιηθεί στην φάση εκπαίδευσης του γλωσσικού μοντέλου).

Το μέγεθος της περιπλοκής παρέχει ένα τρόπο αξιολόγησης εναλλακτικών γλωσσικών μοντέλων σε κοινά δεδομένα δοκιμής χωρίς να χρειάζεται ολοκληρωμένο πείραμα αναγνώρισης. Η χρήση του μεγέθους της περιπλοκής επιτρέπει την αξιολόγηση του τμήματος του γλωσσικού μοντέλου ανεξάρτητα από το ακουστικό μοντέλο, χωρίς όμως να μπορεί να συμπεριλάβει τις όποιες αλληλοεπιδράσεις μεταξύ των δυο μοντέλων. Ένα αξιόπιστο γλωσσικό μοντέλο μπορεί να μην έχει επίδραση στην απόδοση της αναγνώρισης εφόσον οι λέξεις είναι

ακουστικά διακριτές αλλά έχει σημαντικό ρόλο σε ακουστικά συγκεχυμένες λέξεις. Οποιεσδήποτε επιδράσεις του αλγόριθμου αναζήτησης δεν μπορούν να χρησιμοποιηθούν στον υπολογισμό της περιπλοκής. Συνεπώς η περιπλοκή είναι ένα χρήσιμο εργαλείο για την σύγκριση διαφορετικών γλωσσικών μοντέλων αλλά δεν πρέπει να ξεχνάμε ότι το τελικό πείραμα πρέπει να γίνει βάσει του ποσοστού επιτυχίας όλου του συστήματος αναγνώρισης φωνής.

2.4.3 Ποσοστό λανθασμένων λέξεων (Word Error Rate-WER)

Η πιο συνηθισμένη μετρική αξιολόγησης της απόδοσης ενός συστήματος αναγνώρισης φωνής και κατά συνέπεια της ποιότητας του γλωσσικού μοντέλου είναι αυτή του ποσοστού λανθασμένων λέξεων. Το ποσοστό των λανθασμένων λέξεων βασίζεται στο κατά πόσον η ακολουθία των λέξεων που προήλθαν από την μηχανή αναγνώρισης φωνής (υποθετική πρόταση) αποκλίνει από αυτή την σωστή ακολουθία λέξεων (πρόταση αναφοράς). Κατά την αντιστοίχιση της υποθετικής πρότασης και της πρότασης αναφοράς θα δούμε ότι υπάρχουν τριών ειδών λάθη κατά την αναγνώριση μιας φράσης.

- ***Αντικαταστάσεις/Substitutions***--- λέξεις που αναγνωρίσθηκαν λάθος και στην θέση τους υπάρχει μια άλλη.
- ***Αφαιρέσεις/Deletions***--- λέξεις που παραλείφθηκαν να αναγνωριστούν.
- ***Εισαγωγές/Insertions***--- λέξεις που υπάρχουν σαν επιπρόσθετες χωρίς να έχουν ειπωθεί.

Συνεπώς το ποσοστό των λανθασμένων λέξεων μπορεί να περιγραφεί από την παρακάτω εξίσωση

$$WER = 100 \cdot \frac{C(Insertions) + C(Deletions) + C(Substitutions)}{N}\% \qquad (2.23)$$

Όπου N είναι ο συνολικός αριθμός των λέξεων της πρότασης αναφοράς, και $C(x)$ είναι ο αριθμός των λαθών κάθε τύπου x. Επειδή δεν υπάρχει αντιστοιχία ένα προς ένα μεταξύ των λέξεων της υποτιθέμενης πρότασης και της πρότασης αναφοράς απαιτείται η εφαρμογή μιας διαδικασίας δυναμικού προγραμματισμού ώστε να γίνει η αντιστοίχιση των λέξεων.

2.5 Επισκόπηση σε διαφορετικές τεχνικές στατιστικής μοντελοποίησης της γλώσσας

2.5.1 Υπολογισμός πιθανοτήτων απλών N-grams μοντέλων

Ο σκοπός ενός γλωσσικού μοντέλου είναι να ορισθεί η πιθανότητα $P(w_1^n)$ μιας ακολουθίας λέξεων $w_1^n = w_1, w_2, w_3, ..., w_n$. Μπορούμε να χρησιμοποιήσουμε τον κανόνα της αλυσίδας των πιθανοτήτων για να υπολογίσουμε την πιθανότητα:

$$P(w_1^n) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1^2)P(w_4 \mid w_1^3)...P(w_n \mid w_1^{n-1}) = \prod_{k=1}^{n} P(w_k \mid w_1^{k-1}) \quad (2.24)$$

Το ερώτημα που γεννάται είναι το πώς μπορούμε να υπολογίσουμε την πιθανότητα $P(w_n \mid w_1^{n-1})$. Από την στιγμή που είναι δύσκολο να υπολογισθεί η πιθανότητα του τύπου $P(w_n \mid w_1^{n-1})$ για μεγάλα n μπορούμε να υπολογίσουμε την πιθανότητα μιας λέξης που εξαρτάται μόνο από 2 προγενέστερες λέξεις, και μιλάμε για την θεωρία των τριγραμμάτων όπου ισχύει ότι $P(w_n \mid w_1^{n-1}) \approx P(w_n \mid w_{n-2}, w_{n-1})$, και στην πράξη δουλεύει ικανοποιητικά.

Η υπόθεση αυτή που θέλει την πιθανότητα της λέξης να εξαρτάται μόνο από τις προηγούμενες 2 λέξεις καλείται υπόθεση Markov. Η υπόθεση αυτή δέχεται ότι η πιθανότητα ενός μελλοντικού γεγονότος μπορεί να προβλεφθεί κοιτώντας το άμεσο παρελθόν του και όχι το πολύ μακρινό παρελθόν του. Το N-grams γλωσσικό μοντέλο, χρησιμοποιεί τις προηγούμενες $N-1$ λέξεις (τυπικά μια ή δύο) σαν μια προσέγγιση της ιστορίας. Συνεπώς ένα δίγραμμα καλείται σαν πρώτης τάξης Markov μοντέλο (κοιτάζει μια λέξη στο παρελθόν), ενώ το τρίγραμμα σαν δεύτερης τάξης Markov μοντέλο και σε γενικές γραμμές κάθε N-gram καλείται σαν N -1 τάξης Markov μοντέλο.

Άρα η γενική εξίσωση για κάθε N-gram υπολογίζοντας την υπό-συνθήκη πιθανότητα μιας επόμενης λέξης σε μια ακολουθία είναι

$$P(w_n \mid w_1^{n-1}) \approx P(w_n \mid w_{n-N+1}^{n-1}) \quad (2.25)$$

Η γενική εξίσωση αυτή δείχνει ότι η πιθανότητα μιας λέξης w_n βάσει όλων των προγενέστερων λέξεων μπορεί να προσεγγιστεί με την πιθανότητα βάσει των N τελευταίων λέξεων. Βάσει ενός διγραμμικού μοντέλου η συνολική πιθανότητα του string μέσω της προηγούμενης εξίσωσης, είναι ίση με:

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k \mid w_{k-1}) \tag{2.26}$$

Τα N-grams μοντέλα μπορούν να εκπαιδευτούν μετρώντας και κανονικοποιώντας τις εμφανίσεις τους σε ένα σώμα κειμένων εκπαίδευσης. Παίρνουμε ένα σώμα κειμένου και υπολογίζουμε τον αριθμό ενός συγκεκριμένου διγράμματος και τον διαιρούμε με το άθροισμα όλων των διγραμμάτων που μοιράζονται την ίδια πρώτη λέξη:

$$P(w_n \mid w_{n-1}) = \frac{c(w_{n-1}w_n)}{\sum_w c(w_{n-1}w)} \tag{2.27}$$

Το άθροισμα όλων των διγραμμάτων που ξεκινούν με αυτήν την λέξη ισούται με τον αριθμό εμφάνισης της λέξης. Άρα μπορούμε να γράψουμε ότι ισχύει:

$$P(w_n \mid w_{n-1}) = \frac{c(w_{n-1}w_n)}{c(w_{n-1})} \tag{2.28}$$

Και στην γενικευμένη μορφή της, στην περίπτωση των N-grams,

$$P(w_n \mid w_{n-N+1}^{n-1}) = \frac{c(w_{n-N+1}^{n-1}w_n)}{c(w_{n-N+1}^{n-1})} \tag{2.29}$$

Για τα τριγράμματα η πιθανότητα αυτή γράφεται ως εξής

$$P(w_3 \mid w_1, w_2) = \frac{C(w_1, w_2, w_3)}{C(w_1, w_2)} \tag{2.30}$$

Η τελευταία εξίσωση υπολογίζει την πιθανότητα του N-gram διαιρώντας τον αριθμό εμφάνισης μιας ακολουθίας λέξεων με τον αριθμό εμφάνισης των προηγούμενων λέξεων. Ο λόγος αυτός καλείται σχετική συχνότητα. Οι εκτιμήσεις της πιθανότητας N-gram μπορούν να υπολογισθούν χρησιμοποιώντας τις σχετικές πιθανότητες, που αποκαλούνται ως εκτιμήσεις μέγιστης πιθανότητας, δηλ. οι κανονικοποιημένοι αριθμοί εμφάνισης των N-grams σε ένα συγκεκριμένο σώμα κειμένου εκπαίδευσης.

2.5.2 Προβλήματα από την χρήση των N-grams μοντέλων

2.5.2.1 Σπανιότητα δεδομένων

Ακόμη και στα μεγαλύτερα σώματα κειμένων ποτέ δεν θα εμφανιστούν όλα τα πιθανά N-grams μιας γλώσσας. Είναι αναμενόμενο ότι απολύτως αποδεκτά N-grams μπορεί να μην εμφανιστούν ποτέ σε ένα σώμα κειμένου εκπαίδευσης. Επίσης, είναι γνωστό ότι χρησιμοποιώντας σχετικές

συχνότητες σαν τρόπο υπολογισμού των πιθανοτήτων, παράγονται ανεπαρκείς εκτιμήσεις όταν οι αριθμοί των N-grams είναι μικροί. Για να δημιουργηθούν ομοιόμορφες κατανομές, είναι αναγκαίο να εξομαλυνθούν οι πιθανότητες των N-grams. Για την επίλυση αυτού του ζητήματος αναπτύχθηκαν διάφορες τεχνικές εξομάλυνσης. Οι τεχνικές αυτές έχουν σχεδιαστεί για να εξομαλύνουν τις πιθανότητες των N-grams, ώστε κάθε ένα μηδενικής εμφάνισης N-gram να έχει μια μη μηδενική πιθανότητα. Οι τεχνικές εξομάλυνσης μπορούν να χωριστούν σε 2 κυρίως κατηγορίες:

- στις τεχνικές έκπτωσης με στόχο να αναδιανεμηθεί ένα μέρος των πιθανοτήτων σε μη εμφανιζόμενα N-grams.
- και σε τεχνικές που συνδυάζουν διαφορετικά επίπεδα μοντέλων (όπως η παρεμβολή και η τεχνική backoff).

2.5.2.2 Προβλήματα τοπικότητας

Η επιτυχία ενός N-gram στο να ενσωματώσει την απαραίτητη πληροφορία των δεδομένων εκπαίδευσης είναι άμεσα συνδεδεμένη με την κατάλληλη επιλογή του αριθμού N. Για μικρές τιμές N υπάρχει ο κίνδυνος της δημιουργίας γλωσσικού μοντέλου μικρής εμβέλειας ενώ για μεγάλες τιμές του N υπάρχει ο κίνδυνος να μην υπάρχουν αρκετά δεδομένα για την εκπαίδευση των παραμέτρων του γλωσσικού μοντέλου. Στην πράξη η απόδοση των N-grams βασίζεται στις μικρές τιμές του N, το οποίο όμως έχει σαν αποτέλεσμα τα N-grams να μην μπορούν να κωδικοποιήσουν εξαρτήσεις λέξεων μακράς απόστασης. Ας υποθέσουμε ότι θέλουμε να εξετάσουμε την πιθανότητα πρόβλεψης της λέξης “fell’ βάσει της λέξης “stocks” σε δυο ισοδύναμες φράσεις:

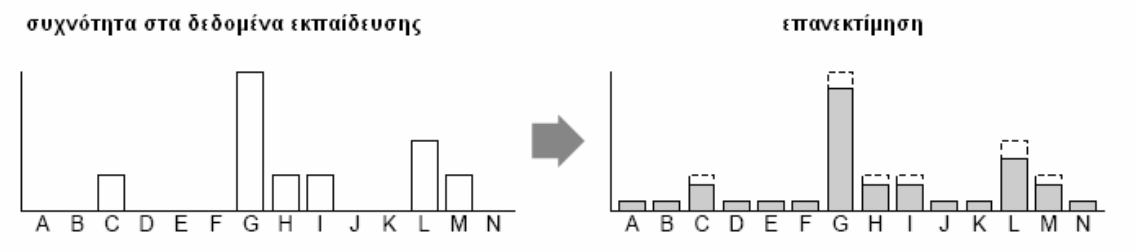
(1) Stocks fell sharply as a result of the announcement.

(2) Stocks, as a result of the announcement, sharply fell.

Στην περίπτωση της πρότασης (1) η πρόβλεψη γίνεται με την βοήθεια ενός διγραμμικού γλωσσικού μοντέλου (N=2). Στην περίπτωση (2) είναι φανερό ότι χρειάζεται ένα γλωσσικό μοντέλο με N=9, το οποίο είναι ουσιαστικά ανέφικτο να υπολογισθεί με την μέχρι τώρα συλλογή δεδομένων.

2.5.3 Τεχνικές βελτίωσης της απόδοσης των N-grams μοντέλων

Η εκτίμηση μέγιστης πιθανότητας των N-grams έχει σαν αποτέλεσμα την υπερεκτίμηση των πιθανοτήτων αυτών των N-grams που εμφανίζονται στα δεδομένα εκπαίδευσης. Οι πιθανότητες των μη εμφανιζόμενων N grams υποεκτιμούνται (παίρνουν τιμές≃0). Αυτό το πρόβλημα καλείται πρόβλημα μηδενικής πιθανότητας. Τεχνικές «έκπτωσης» όπως η τεχνική Good Turing και η τεχνική της απόλυτης έκπτωσης επιλύουν το πρόβλημα αυτό μειώνοντας τις εκτιμήσεις μέγιστης πιθανότητας και ανακατανέμοντας την αδέσμευτη μάζα πιθανότητας σε μη εμφανιζόμενα N-grams (βλέπε Σχήμα 2.1).



Σχήμα 2.1 Η μείωση μάζας ανακατανέμεται σε γεγονότα που δεν παρατηρήθηκαν στα δεδομένα εκπαίδευσης.

2.5.3.1 Δεδομένα δοκιμών και εκπαίδευσης

Ένα σημαντικό σφάλμα στην στατιστική επεξεργασία γλώσσας είναι να γίνονται δοκιμές με την χρήση των δεδομένων εκπαίδευσης. Η ιδέα του ελέγχου είναι να δούμε πόσο καλά δουλεύει ένα μοντέλο σε άγνωστα δεδομένα. Σε πολλές περιπτώσεις παρουσιάζεται το φαινόμενο της υπερ-εκπαίδευσης. Αυτό σημαίνει ότι τα προσδοκώμενα μελλοντικά συμβάντα είναι παρόμοια με αυτά που χρησιμοποιήθηκαν για την δοκιμή του μοντέλου. Γι' αυτό είναι αναγκαίος ο έλεγχος να γίνεται σε διαφορετικά δεδομένα. Αυτό επίσης ισχύει και για τον υπολογισμό της διασταυρωμένης εντροπίας. Για τον υπολογισμό της διασταυρωμένης εντροπίας λαμβάνεται ένα μεγάλο δείγμα κειμένων και υπολογίζεται η διασταυρωμένη εντροπία ανά λέξη, βάσει του μοντέλου. Η διασταυρωμένη εντροπία αποτελεί ένα μέτρο εκτίμησης της ποιότητας του μοντέλου και συνεπώς αυτή η διαδικασία είναι αληθής μόνο όταν τα δεδομένα εκπαίδευσης είναι ανεξάρτητα από τα δεδομένα δοκιμής, με την προϋπόθεση να είναι πολλά ώστε να είναι ενδεικτικά της πολυπλοκότητας της γλώσσας. Εάν υπολογίσουμε την διασταυρωμένη εντροπία πάνω σε δεδομένα εκπαίδευσης τότε η εντροπία θα είναι μικρότερη της πραγματικής. Συνεπώς

στο ξεκίνημα κάθε ελέγχου πρέπει να διαχωριστούν τα δεδομένα σε δεδομένα δοκιμών και σε δεδομένα εκπαίδευσης. Τα δεδομένα δοκιμών αποτελούν συνήθως ένα μικρό ποσοστό γύρω στο 5-10% των συνολικών δεδομένων, αλλά είναι αναγκαίο να είναι αρκετά για να είναι αξιόπιστα.

Πολλές φορές είναι ανάγκη να διαχωριστούν τα δεδομένα δοκιμών και εκπαίδευσης σε 2 μέρη. Σε πολλές μεθόδους στατιστικής επεξεργασίας φυσικής γλώσσας όπως αυτή της έκπτωσης των N-grams, ο αριθμός των εμφανίσεων αυτών εξομαλύνεται με την περαιτέρω χρήση των δεδομένων held out ή επικύρωσης. Τα δεδομένα held out πρέπει να είναι ανεξάρτητα και από τα δεδομένα δοκιμών και από τα δεδομένα εκπαίδευσης. Συνήθως τα δεδομένα held out χρησιμοποιούνται στην εκτίμηση πολύ λιγότερων παραμέτρων από αυτές που εκτιμώνται με τα πρωτεύοντα δεδομένα εκπαίδευσης. Και για αυτόν τον λόγο χρειάζεται να είναι πολύ μικρός ο όγκος τους περί του 10% σε σχέση με τα δεδομένα εκπαίδευσης. Επίσης και στην πλευρά των δεδομένων δοκιμής είναι αναγκαία η ύπαρξη δυο συνόλων δοκιμής, το ένα για την τελειοποίηση του αλγορίθμου και το άλλο για την τελική του αξιολόγηση πάνω σε πραγματικά δεδομένα.

2.5.3.2 Εμπειρική εκτίμηση

Πώς μπορεί κάποιος να ξέρει το ποσοστό της μάζας της πιθανότητας που είναι σωστό να διατεθεί υπέρ των μη εμφανιζόμενων N-grams. Ένας τρόπος για να ελεγχθεί αυτό είναι εμπειρικά. Μια εμπειρική προσέγγιση (συνήθως αναφέρεται σαν εκτίμηση held out) μπορεί να βασίζεται στην εξής ερώτηση «πόσο συχνά διγράμματα που εμφανίζονται r φορές σε ένα σύνολο δεδομένων εκπαίδευσης πρόκειται να συμβούν σε ένα σύνολο νέων δεδομένων (δεδομένα που λαμβάνονται από την ίδια πηγή όπως τα δεδομένα εκπαίδευσης);» Η πραγμάτωση αυτής της ιδέας περιγράφεται σαν held out εκτίμηση των Jelinek και Marcer (1985). Ας υποθέσουμε ότι u_r είναι ο αριθμός των N-grams που εμφανίζονται r φορές στα δεδομένα εκπαίδευσης, $c_t(\cdot)$ και $c_h(\cdot)$ είναι αντίστοιχα η συχνότητα στα δεδομένα εκπαίδευσης και στα held out δεδομένα. Η εκπτωτική συχνότητα για ένα δίγραμμα $w_{n-1}w_n$ μπορεί να υπολογισθεί βάσει της σχέσης

$$\hat{r}_{emp} = \frac{1}{u_r} \sum_{\substack{(w_{n-1}w_n): \\ c_t(w_{n-1}w_n)=r}} c_h(w_{n-1}w_n) \tag{2.31}$$

Εάν τα μεγέθη των δεδομένων εκπαίδευσης και held out είναι διαφορετικά τότε πρέπει να κανονικοποιηθούν. Η υπό συνθήκη σχετική συχνότητα διγράμματος είναι

$$\hat{f}_{emp}(w_{n-1}w_n) = \frac{\hat{r}_{emp}}{c_h(w_{n-1})} \tag{2.32}$$

2.5.3.3 Διασταυρωμένη επικύρωση

Η $\hat{f}_{emp}$ εκτιμάται κοιτώντας τι ακριβώς συμβαίνει στα δεδομένα εκπαίδευσης. Η όλη ιδέα της held out εκτίμησης είναι να λάβουμε το ίδιο αποτέλεσμα διαρρέοντας τα δεδομένα εκπαίδευσης σε 2 μέρη. Οι αρχικές εκτιμήσεις γίνονται στο ένα μέρος των δεδομένων και οι ακριβής υπολογισμός τους γίνεται με την χρήση της άλλης δεξαμενής δεδομένων που λέγεται held out δεδομένα. Το μόνο μειονέκτημα αυτής της μεθόδου είναι ότι τα δεδομένα εκπαίδευσης είναι λιγοστά και οι εκτιμήσεις θα είναι λιγότερο αξιόπιστες. Αντί να χρησιμοποιούμε τα δεδομένα εκπαίδευσης για τον υπολογισμό της συχνότητας των N-grams και τα held out δεδομένα για τον υπολογισμό των παραμέτρων εξομάλυνσης είναι προτιμότερο κάθε μικρότερο μέρος των δεδομένων εκπαίδευσης να χρησιμοποιείται τόσο σαν αρχικά δεδομένα εκπαίδευσης όσο και σαν δεδομένα held out. Οι μέθοδοι αυτοί καλούνται ως διασταυρωμένη επικύρωση. Η διασταυρωμένη επικύρωση είναι μια σχετική προσέγγιση προς αυτή της held out εκτίμηση, και μπορεί να εκτελεστεί με τα επόμενα βήματα.

1. διαχωρίζεται το σύνολο των δεδομένων εκπαίδευσης σε $K > 1$ μέρη.

2. για κάθε ένα $k = 1...K$, κρατάμε το μέρος k και

- από τα υπόλοιπα $K-1$ μέρη (αναφέρονται σαν δεδομένα ανάπτυξης) μαζεύουμε στατιστικά $u_r(k)$ ο αριθμός των N-grams που εμφανίζονται r φορές.

- Από το held out μέρος k μαζεύουμε τον συνολικό αριθμό εμφανίσεων $t_r(k)$ των N-grams που εμφανίζονται r φορές στα δεδομένα ανάπτυξης και κανονικοποιούμε το $t_r(k)$ βάσει των μεγεθών των δεδομένων ανάπτυξης και held out.

3. Υπολογίζουμε το μέσο όρο

$$\hat{r}_{cv} = \frac{\sum_{k=1...K} t_r(k)}{\sum_{k=1...K} u_r(k)} \tag{2.33}$$

2.5.4 Μοντέλα Εξομάλυνσης

Ένα από τα σημαντικότερα προβλήματα των N-grams είναι η ίδια η εκπαίδευσης τους, και θέλοντας να αποσαφηνίσουμε το τι σημαίνει αυτό θα πούμε ότι η απόδοση τους εξαρτάται άμεσα από την ποιότητα των σωμάτων κειμένων που χρησιμοποιούνται για την εκπαίδευση τους. Αυτό συμβαίνει γιατί κάθε ένα σώμα κειμένου είναι πεπερασμένο και μερικά από τα πιο αποδεκτά N-grams δεν βρίσκονται σ' αυτό. Ένα άλλο πρόβλημα που αντιμετωπίζουν τα N-grams είναι η αδυναμία τους να χρησιμοποιήσουν μακράς απόστασης συμφραζόμενα, με αποτέλεσμα να υπάρχει εμφανής τάση υποεκτίμησης της πιθανότητας των strings που κατά τύχη δεν συνυπάρχουν στα δεδομένα εκπαίδευσης. Για αυτόν το λόγο αναπτύχθηκαν τεχνικές που αναθέτουν μη μηδενικές πιθανότητες σε N-grams που εμφανίζουν μηδενικό αριθμό εμφανίσεων στα δεδομένα εκπαίδευσης. Η διαδικασία της επαναξιολόγησης των N-grams με μηδενική ή μικρή πιθανότητα καλείται εξομάλυνση. Στην επόμενη παράγραφο θα μελετήσουμε μερικές από τις μεθόδους εξομάλυνσης και τους τρόπους για να τις συνδυάσουμε ώστε να έχουμε καλύτερα αποτελέσματα.

Η εξομάλυνση είναι μια σημαντική παράμετρος ενίσχυσης της αξιοπιστίας του γλωσσικού μοντέλου με αποτέλεσμα αρκετοί αλγόριθμοι εξομάλυνσης να προτείνονται στην βιβλιογραφία. Το αντικείμενο της εξομάλυνσης του γλωσσικού μοντέλου είναι να παραχθούν περισσότερο ακριβείς πιθανότητες $P(W)$, ρυθμίζοντας κατάλληλα τις εκτιμήσεις μέγιστης πιθανότητας που υπολογίζονται με την προσέγγιση της σχετικής συχνότητας. Στην πράξη αυτό σημαίνει επανακαθορισμός, ή έκπτωση του αριθμού των N-grams σε ένα σώμα κειμένων. Με τον όρο «έκπτωση» περιγράφεται η διαδικασία μείωσης του αριθμού των N-grams με μη μηδενική τιμή, σύμφωνα με κάποια συνάρτηση «έκπτωσης», ώστε να «σωθεί» κάποια μάζα πιθανότητας και να αποδοθεί σε N-grams μηδενικής εμφάνισης ή ακόμη και σε αυτά με μικρό αριθμό εμφανίσεων μέσω της συνάρτησης ανακατανομής. Η συνάρτηση «έκπτωσης» και η συνάρτηση ανακατανομής συνδυάζονται χρησιμοποιώντας είτε την στρατηγική backoff (Katz, 1987) είτε την στρατηγική της παρεμβολής (Jelinek and Mercer, 1980). Και οι δύο αυτές τεχνικές εξομάλυνσης χρησιμοποιούν μικρότερης τάξης κατανομές για τον καθορισμό της πιθανότητας των N-grams με μηδενικό ή μικρό αριθμό εμφανίσεων.

2.5.4.1 Προσθετική εξομάλυνση

Ένας απλός αλλά αποδοτικός τρόπος εξομάλυνσης είναι να προσθέσουμε τον αριθμό ένα (1) σε όλους τους αριθμούς εμφάνισης των N-grams πριν αυτοί κανονικοποιηθούν για να γίνουν πιθανότητες. Η μέθοδος αυτή καλείται εξομάλυνση με προσθήκη ενός (add one smoothing) (Laplace, 1814;1995). Αν και ο αλγόριθμος δεν είναι αρκετά αξιόπιστος εντούτοις αποτελεί την

βασική αρχή πολλών αλγορίθμων εξομάλυνσης που θα δούμε στην συνέχεια. Ας υποθέσουμε ότι εφαρμόζουμε την εξομάλυνση με προσθήκη ενός σε πιθανότητες μονογραμμάτων. Άρα η εκτίμηση μέγιστης πιθανότητας ενός μονογράμματος θα είναι πριν την εξομάλυνση ίση με το αποτέλεσμα της διαίρεσης του αριθμού εμφάνισης του μονογράμματος δια τον συνολικό αριθμό των στοιχείων του κειμένου εκπαίδευσης N :

$$P(w_x) == \frac{c(w_x)}{\sum_i c(w_i)} = \frac{c(w_x)}{N} \tag{2.34}$$

Όλες οι διαδικασίες εξομάλυνσης στρέφονται στον υπολογισμό ενός νέου παράγοντα c^*.

Στην περίπτωση της εξομάλυνσης με προσθήκη ενός, ο αριθμός εμφανίσεων ενός μονογράμματος θα είναι ίσος με

$$P(w_x) = \frac{1 + c(w_x)}{\sum_{w_i} (1 + c(w_i))} \tag{2.35}$$

και ο νέος αριθμός εμφανίσεων του μονογράμματος θα είναι ίσος με

$$c_i^* = (c_i + 1)\frac{N}{|V| + N} \tag{2.36}$$

Όπου V είναι ο συνολικός αριθμός διαφορετικών ειδών λέξεων των δεδομένων εκπαίδευσης δηλαδή είναι ίσος με το μέγεθος του λεξιλογίου. Και στην περίπτωση των πιθανοτήτων αυτός ο αριθμός πρέπει να κανονικοποιηθεί με το N. Άρα μπορούμε να γράψουμε ότι

$$p_i^* = \frac{(1 + c_i)}{|V| + N} \tag{2.37}$$

Μπορούμε να καταλήξουμε στο συμπέρασμα ότι ο αλγόριθμος εξομάλυνσης είναι μια διαδικασία έκπτωσης μερικών μη μηδενικών εμφανίσεων N-grams με σκοπό να μετατοπιστεί μάζα πιθανοτήτων από τα μη μηδενικά στα μηδενικά N-grams. Σε πολλές περιπτώσεις οι αλγόριθμοι εξομάλυνσης αναφέρονται σαν αλγόριθμοι έκπτωσης με συντελεστή έκπτωσης ίσο με $d_c = \frac{c^*}{c}$. Όπου c^* είναι ο αριθμός των N-grams μετά την εξομάλυνση και c ο αριθμός των N-grams πριν την διαδικασία αυτή.

Το πρόβλημα στην χρήση της εξομάλυνσης με προσθήκη ενός είναι ότι λόγω της προσθήκης ενός αυθαίρετου αριθμού όπως το ένα μετακινείται αρκετή μάζα πιθανότητας σε N-grams που δεν εμφανίστηκαν στα δεδομένα εκπαίδευσης, με αποτέλεσμα να έχουν μια σχετική υπερ-εκτίμηση των μη εμφανιζόμενων N-grams στα δεδομένα εκπαίδευσης. Ένα παράδειγμα απλής αλλά ταυτόχρονα αποδοτικής διαδικασίας έκπτωσης είναι η προσθετική εξομάλυνση με την προσθήκη ενός αριθμού μικρότερου του ενός (Additive Interpolation) (Lindstone, 1920; Johnson, 1932; Jeffreys, 1948). Άρα η προηγούμενη σχέση μπορεί να γραφτεί τώρα ως εξής με την χρήση του παράγοντα $\lambda < 1$,

$$p_i^* = \frac{(c_i + \lambda)}{N + V\lambda} \tag{2.38}$$

2.5.4.2 Τεχνική Good-Turing

Η βασική προσέγγιση ανάμεσα σε όλες τις μεθόδους «έκπτωσης» είναι η τεχνική Good-Turing, η οποία πρώτο περιγράφτηκε από τον Good (1953) πάνω στην ιδέα του Turing και η πλήρης απόδειξη της παρουσιάστηκε από τους Church et al.(1991). Η βασική ιδέα αυτής της τεχνικής είναι ότι εκτιμάται το μέγεθος της μάζας πιθανότητας που θα ανατεθεί στα N-grams με μηδενική ή μικρή πιθανότητα βάσει των N-grams με μεγαλύτερο αριθμό εμφανίσεων. Με άλλα λόγια εξετάζεται το N_c, ο αριθμός των N-grams που εμφανίζονται c φορές. Με τον αριθμό των N-grams που εμφανίζονται c φορές αναφερόμαστε στην συχνότητα της συχνότητας c. Έτσι εφαρμόζοντας την ιδέα της εξομάλυνσης των διγραμμάτων, με N_0 αναφερόμαστε στα διγράμματα b με αριθμό εμφανίσεων το 0, με N_1 αναφερόμαστε στα διγράμματα b με αριθμό εμφανίσεων το 1, και ούτω καθεξής

$$N_c = \sum_{b:c(b)=c} 1$$

Με αυτήν την τεχνική ο μη μηδενικός αριθμός εμφάνισης N-grams υπολογίζεται ως εξής

$$c^* = (c+1)\frac{N_{c+1}}{N_c} \tag{2.39}$$

Όπου c^* συμβολίζεται ο αριθμός των εμφανίσεων με την χρήση της εξομάλυνσης και N_c ο αριθμός εμφάνισης των N-grams που εμφανίζονται c φορές. Ο υπολογισμός βάσει της τεχνικής Good-Turing για τα N-grams με μηδενική συχνότητα εμφάνισης είναι

$$c^* = \frac{N_1}{N_0} \tag{2.40}$$

Η οποία μπορεί να μεταφραστεί σε πιθανότητα κανονικοποιώντας τον αρχικό αριθμό με την κατανομή N :

$$P_{Good-Turing} = \frac{c^*}{N} \tag{2.41}$$

Όπως γίνεται αντιληπτό η εκτίμηση Good-Turing για τα διγράμματα που δεν έχουν εμφανιστεί στα δεδομένα εκπαίδευσης εκτιμάται με τον λόγο του αριθμού των διγραμμάτων που εμφανίστηκαν μια φορά προς αυτόν που δεν έχουν εμφανιστεί ακόμα. Το να εκτιμάμε γεγονότα που δεν έχουν συμβεί βάσει γεγονότων που συνέβησαν μια φορά είναι μια ιδέα που την χρησιμοποιεί όχι μόνο ο αλγόριθμος Good-Turing αλλά και η μέθοδος Witten-Bell που θα παρουσιάσουμε παρακάτω. Ο αριθμός N_o υπολογίζεται με την αφαίρεση του αριθμού των εμφανιζόμενων διγραμμάτων από τον συνολικό αριθμό διγραμμάτων που είναι ίσος με V^2. Για ένα λεξιλόγιο της τάξης των 65 χιλιάδων λέξεων , ο εκτιμώμενος αριθμός διγραμμάτων ο οποίος δεν εμφανίζεται στο σώμα κειμένων είναι όπου είναι ο αριθμός των εμφανιζόμενων διγραμμάτων. Για την καλή λειτουργία της τεχνικής Good-Turing, ο αλγόριθμος χρησιμοποιείται σε περιπτώσεις όπου ο αριθμός εμφάνισης των N-grams είναι μικρότερος από ένα όριο (στην πράξη το όριο αυτό τίθεται ίσο με το 5), ενώ για τιμές άνω του ορίου ο υπολογισμός των συχνοτήτων γίνεται βάσει της εκτίμησης της μέγιστης πιθανότητας.

2.5.4.3 Τεχνική Witten-Bell

Μια ακόμη συχνά εφαρμοζόμενη τεχνική έκπτωσης είναι αυτή των Witten-Bell (Witten and Bell, 1991), η οποία εκτιμά την πιθανότητα των μηδενικών N-grams, λαμβάνοντας υπόψη τα N-grams που εμφανίζονται για πρώτη φορά, και όχι αυτά που εμφανίζονται ακριβώς μια φορά όπως γίνεται με την μέθοδο Good-Turing. Η ιδέα πίσω από αυτή την τεχνική είναι ότι τα N-grams που εμφανίζουν μηδενική συχνότητα μπορούν να μοντελοποιηθούν με την πιθανότητα ενός N-gram που εμφανίζεται για πρώτη φορά. Το ερώτημα όμως είναι πώς θα υπολογίσουμε την πιθανότητα εμφάνισης για πρώτη φορά ενός N-gram. Η απάντηση δίνεται με τον υπολογισμό του αριθμού εμφάνισης των N-grams για πρώτη φορά στα δεδομένα εκπαίδευσης. Ο υπολογισμός αυτός είναι πολύ εύκολος από την στιγμή που ο αριθμός των πρώτο εμφανιζόμενων N-grams είναι ίσος με τον αριθμό των ειδών N-grams που υπάρχουν στα δεδομένα εκπαίδευσης.

Συνεπώς η συνολική πιθανότητα όλων των μηδενικών N-grams μπορεί να υπολογιστεί από τον λόγο του αριθμού των ειδών προς το άθροισμα του αριθμού των ειδών και του συνολικού αριθμού στοιχείων στα δεδομένα εκπαίδευσης.

$$\sum_{i:c_i=0} p_i^* = \frac{T}{N+T} \tag{2.42}$$

Η προηγούμενη σχέση δίνει την εκτίμηση μέγιστης πιθανότητας για γεγονότα εμφάνισης νέου είδους. Ας σημειωθεί στο σημείο αυτό ότι είναι διαφορετικός ο αριθμός των παρατηρούμενων ειδών T από τον συνολικό αριθμό ειδών (V) που χρησιμοποιείται στην τεχνική προσθήκης ενός. Ο αριθμός T συμβολίζει είδη που ήδη τα έχουμε δει στο κείμενο ενώ V είναι ο συνολικός αριθμός των ειδών που θα δούμε εντέλει. Αυτή είναι η συνολική πιθανότητα των μη εμφανιζόμενων N-grams. Ας υποθέσουμε ότι Z είναι ο συνολικός αριθμός ειδών των N-grams με αριθμό εμφάνισης το μηδέν, άρα η πιθανότητα του κάθε μονογράμματος μετά τον ισοκαταμερισμό της μάζας πιθανότητας θα είναι ίση με

$$p_i^* = \frac{T}{Z(N+T)} \tag{2.43}$$

$$\text{και } Z = \sum_{i:c_i=0} 1$$

Εάν η συνολική πιθανότητα των μηδενικών N-grams υπολογίζεται από την σχέση (2.30) τότε η παραπάνω πιθανότητα θα πρέπει να απορρέει από την έκπτωση των πιθανοτήτων όλων των εμφανιζόμενων N-grams .

$$\text{Άρα } p_i^* = \frac{c_i}{Z(N+T)} \quad if(c_i > 0) \tag{2.44}$$

Χρησιμοποιώντας $\frac{N}{N+T}$, όπου N είναι ο αρχικός αριθμός εμφάνισης του N-gram, σαν λόγο κανονικοποίησης, μετά την μείωση ο αριθμός των εμφανίσεων γίνεται ως εξής.

$$c_i^* = \begin{cases} \frac{T}{Z} \quad \frac{N}{N+T} & if \quad c_i = 0 \\ c_i \quad \frac{N}{N+T} & if \quad c_i > 0 \end{cases} \tag{2.45}$$

Ο αλγόριθμος Witten-Bell μοιάζει κατά πολύ με την εξομάλυνση προσθήκης ενός για τα μονογράμματα, αλλά αν προεκταθούμε στα διγράμματα τότε υπάρχει διαφορά. Αυτό συμβαίνει γιατί στα διγράμματα που έχουν κοινή την πρώτη λέξη αντιστοιχεί ένα είδος διγράμματος. Σε αυτήν περίπτωση για να υπολογίσουμε την πιθανότητα ενός διγράμματος $w_{i-1}w_i$ που δεν έχει

εμφανιστεί θα χρησιμοποιηθεί η πιθανότητα εμφάνισης ενός νέου διγράμματος που ξεκινάει από την λέξη w_{i-1}.

Μπορούμε λοιπόν να γράψουμε την παρακάτω σχέση ώστε να αποτυπώσουμε το αλγόριθμο Witten-Bell για τα διγράμματα,

$$\sum_{i:c(w_x w_i)} p_i^*(w_i \mid w_{i-1}) = \frac{T(w_x)}{N(w_x)+T(w_x)} \tag{2.46}$$

Με τον όρο $T(w_x)$ να αντιστοιχεί στον αριθμό των ειδών των διγραμμάτων και το $N(w_x)$ στον συνολικό αριθμό διγραμμάτων με πρώτη λέξη την w_x. Συνεπώς κάθε πρώην μηδενικό δίγραμμα μοιράζεται εξίσου την μάζα πιθανότητας που απόμεινε από την έκπτωση και έτσι έχουμε

$$p_i^*(w_i \mid w_{i-1}) = \frac{T(w_{i-1})}{Z(w_{i-1})(N+T(w_{i-1}))} \quad if\,(c_{w_{i-1}w_i}=0) \tag{2.47}$$

Έστω N ο αριθμός των διγραμμάτων που αρχίζουν με w_{i-1} και Z να είναι ο συνολικός αριθμός των διγραμμάτων με την πρώτη λέξη να έχει αριθμό εμφάνισης το μηδέν δηλαδή $Z(w_{i-1}) = V - T(w_{i-1})$,όπου V είναι ο αριθμός του λεξιλογίου και συνεπώς ο αριθμός των πιθανών ειδών των διγραμμάτων. Όσον αφορά τα μη μηδενικά διγράμματα ισχύει ότι

$$p_i^*(wi \mid w_{i-1}) = \frac{c(w_{i-1}w_i)}{c(w_{i-1})+T(w_{i-1}))} \quad if\,(c_{w_{i-1}w_i}>0) \tag{2.48}$$

2.5.5 Προηγμένες τεχνικές εξομάλυνσης

Μέχρι στιγμής οι μέθοδοι που έχουμε παρουσιάσει κάνουν χρήση της συχνότητας r για κάθε N-gram και προσπαθούν να εκτιμήσουν καλύτερα την πιθανότητα εμφάνισης του σε επιπλέον κείμενα. Αλλά αντί να δίνεται η ίδια εκτίμηση για όλα τα N-grams που δεν συμβαίνουν ή συμβαίνουν σπάνια, είναι καλύτερο να παράγουμε εκτιμήσεις πού στηρίζονται στην συχνότητα εμφάνισης των (N-1)-grams που βρίσκονται στα N-grams. Εάν και τα (N-1)-grams είναι απίθανο να συμβούν τότε και τα N-grams θα εκτιμηθούν με μικρές τιμές. Εάν τα (N-1)-grams είναι πιο πιθανά να συμβούν τότε και τα N-grams θα εκτιμηθούν με μεγαλύτερες τιμές. Οι Church και Gale (1991a) παρουσίασαν μια λεπτομερειακή μελέτη αυτής της ιδέας αποδεικνύοντας πώς οι πιθανότητες των μη εμφανιζόμενων διγραμμάτων μπορούν επανεκτιμηθούν με την χρήση των πιθανοτήτων των μονογραμμάτων που τα αποτελούν.

Σε αυτή την ενότητα θα περιγράψουμε τεχνικές συνδυασμού διαφορετικής τάξης μοντέλων. Για τα μοντέλα N-grams το μυστικό της επιτυχίας τους βρίσκεται στον συνδυασμό

μοντέλων διαφορετικής τάξης. Έτσι μπορούμε να συνδυάσουμε την εκτίμηση μέγιστης πιθανότητας N-grams διαφορετικών τάξεων (με κάποια ανοχή σε μη εμφανιζόμενες λέξεις) χρησιμοποιώντας την τεχνική γραμμικής εξομάλυνσης με πολύ θετικά αποτελέσματα (Chen και Goodman 1996). Υπάρχουν φυσικά πολλές άλλες τεχνικές που δίνουν ακόμη καλύτερα αποτελέσματα και αυτές θα παρουσιάσουμε παρακάτω.

Ένας τρόπος επίλυσης του προβλήματος της σπανιότητας των τριγραμμάτων είναι η ανάμειξη τους με μοντέλα διγραμμάτων και μονογραμμάτων τα οποία υποφέρουν λιγότερο από την σπανιότητα δεδομένων. Στην πράξη οι τεχνικές έκπτωσης που παρουσιάστηκαν παραπάνω εφαρμόζονται στο πλαίσιο ενός συνδυαστικού μοντέλου, που είτε κάνει χρήση της στρατηγικής της υποχώρησης είτε της στρατηγικής της γραμμικής εξομάλυνσης μεγάλων τάξεων μοντέλων N-grams με αυτά μικρότερης τάξης μοντέλα N-grams. Και οι δύο αυτές στρατηγικές χρησιμοποιούν την εξής ιδέα: τα μοντέλα N-grams μικρότερης τάξης μπορούν να παρέχουν σημαντική πληροφορία για τον υπολογισμό της πιθανότητας των μοντέλων μεγαλύτερης τάξης, ιδιαίτερα όταν δεν υπάρχουν ή είναι περιορισμένα και δεν είναι ικανοποιητικά για την εκτίμηση της πιθανότητας ενός N-gram μοντέλου μεγάλης τάξης.

2.5.5.1 Εξομάλυνση παρεμβολής με διαγραφές

Ας υποθέσουμε ότι ένα διγραμμικό μοντέλο παράγεται από δεδομένα εκπαίδευσης όπου το πλήθος εμφανίσεων των παρακάτω διγραμμάτων είναι: c(burnish the)=0 και c(burnish thou)=0. Σύμφωνα με την τεχνική της προσθετικής εξομάλυνσης και της τεχνικής Good-Turing η πιθανότητα και των δυο θα ήταν η ίδια:

$$p(the \mid burnish) = p(thou \mid burnish) \tag{2.49}$$

Αν και είναι φανερό ότι $p(the \mid burnish) > p(thou \mid burnish)$ από την στιγμή όπου η λέξη “the” είναι πιο συχνή από την λέξη “thou”. Στην περίπτωση όπου χρησιμοποιήσουμε την γραμμική παρεμβολή ενός μονογράμματος και ενός διγράμματος τότε θα έχουμε

$$p_{\text{int}erp}(w_i \mid w_{i-1}) = \lambda p_{ML}(w_i \mid w_{i-1}) + (1-\lambda) P_{ML}(w_i) \text{ με } 0 \le \lambda \le 1 \tag{2.50}$$

Αλλά επειδή $p(the \mid burnish) = p(thou \mid burnish)$ και $p(the) \gg p(thou)$ θα ισχύει ότι

$$p_{\text{int}erp}(the \mid burnish) > p_{\text{int}erp}(thou \mid burnish) \tag{2.51}$$

Σε γενικές γραμμές η τεχνική που περιγράψαμε είναι πολύ χρήσιμη αφού όταν λείπουν μεγάλης τάξης N-grams χρησιμοποιούμε μικρότερης τάξης N-grams μιας που είναι περισσότερο πιθανά να εμφανιστούν σε ένα σώμα κειμένου. Με την όρο παρεμβολή, που συχνά αναφέρεται με την ονομασία Jelinek-Mercer (Jelinek and Mercer, 1980), οι πιθανότητες των μονογραμμάτων, των διγραμμάτων και των τριγραμμάτων συνδυάζονται και σταθμίζονται με ένα ιδιαίτερο βάρος (λ). Για την εκτίμηση της πιθανότητας ενός τριγράμματος η εξίσωση της παρεμβολής παίρνει την ακόλουθη μορφή:

$$\hat{P}(w_n \mid w_{n-2}w_{n-1}) = \lambda_1 P(w_n \mid w_{n-2}w_{n-1}) + \lambda_2 P(w_n \mid w_{n-1}) + \lambda_3 P(w_n) \quad (2.52)$$

$$\text{με } \sum_i \lambda_i = 1$$

Στην πράξη με τον αλγόριθμο της παρεμβολής με διαγραφές δεν απαιτείται μόνο η εκπαίδευση των τριών λ στην περίπτωση ενός τριγραμμικού μοντέλου. Σε αυτή την περίπτωση το κάθε λ δίνεται ως συνάρτηση των συμφραζομένων. Με αυτόν τον τρόπο όταν έχουν ακριβή αριθμό διγραμμάτων θεωρούμε ότι ο αριθμός του τριγράμματος που βασίζεται στο δίγραμμα είναι πολύ αξιόπιστος και δίνουμε μεγαλύτερη τιμή στο λ του τριγράμματος δίνοντας του έτσι μεγαλύτερη βαρύτητα στην παρεμβολή. Συνεπώς μια πιο λεπτομερής έκδοση της σχέση παρεμβολής μπορεί να είναι η ακόλουθη:

$$\hat{P}(w_n \mid w_{n-2}w_{n-1}) = \lambda_1(w_{n-2}^{n-1})P(w_n \mid w_{n-2}w_{n-1}) + \lambda_2(w_{n-2}^{n-1})P(w_n \mid w_{n-1})\lambda_3(w_{n-2}^{n-1})P(w_n) \quad (2.53)$$

Το ερώτημα που τίθεται εδώ είναι πώς θα οριστούν οι τιμές λ. Στην περίπτωση της γραμμικής παρεμβολής τα λ εκπαιδεύονται από τα held-out δεδομένα. Τα δεδομένα held-out είναι ένα κομμάτι από τα δεδομένα εκπαίδευσης, τα οποία, δεν χρησιμοποιούνται για τον υπολογισμό των σχετικών πιθανοτήτων αλλά για τον ορισμό των διάφορων παραμέτρων που εξυπηρετούν την βελτιστοποίηση των τεχνικών εξομάλυνσης. Έτσι επιλέγονται λ τα οποία μεγιστοποιούν την πιθανότητα των N-grams στα δεδομένα held-out (Baum, 1972; Dempster et al., 1997; Jelinek and Mercer, 1980; Bahl et al., 1983).

2.5.5.2 Katz's backing off

Η εξομάλυνση Katz (Katz, 1987) είναι αποκλειστικά μια μορφή Backoff εξομάλυνσης μη γραμμική. Βάσει αυτής της τεχνικής πραγματοποιείται διαδικασία υποχώρησης σε μικρότερης τάξης μοντέλα. π.χ (N-1)-grams. όταν απαιτείται να προσδιορισθούν μηδενικά εμφανιζόμενα N-grams. Η αρχική φάση της εξομάλυνσης περιλαμβάνει και την τεχνική Good Turing. Η ουσιαστική διαφορά μεταξύ της παρεμβολής και της υποχώρησης είναι ότι για τα N-grams με μη μηδενικές εμφανίσεις ισχύει ότι η παρεμβολή πάντα χρησιμοποιεί την πληροφορία των μικρότερης τάξης N-grams ενώ η υποχώρηση όχι. Στην περίπτωση της υποχώρησης Katz, όταν ένα N-gram έχει μηδενική πιθανότητα τότε χρησιμοποιείται ένα (N-1)-gram και ενδεχομένως ένα μοντέλο μικρότερης τάξης μέχρι να φθάσει το σημείο να μπορεί να προσδιοριστεί. Έτσι στην γενική περίπτωση μπορούμε να γράψουμε ότι:

$$P_{katz}(w_n \mid w_{n-N+1}^{n-1}) = \begin{cases} P^*(w_n \mid w_{n-N+1}^{n-1}) & \text{εάν} \quad C(w_{n-N+1}^{n-1}) > 0 \\ a(w_{n-N+1}^{n-1}) P_{katz}(w_n \mid w_{n-N+2}^{n-1}) & \text{αλλίως} \end{cases} \quad (2.54)$$

2.5.5.3 Συνδυασμός τεχνικών υποχώρησης και έκπτωσης

Προηγουμένως αναφερθήκαμε στην τεχνική έκπτωσης με την οποίαν μεταφέρεται μάζα πιθανότητας σε μη εμφανιζόμενα γεγονότα. Χάριν απλότητας θεωρούμε ότι όλα τα γεγονότα είναι ισοπίθανα με συνέπεια η μάζα πιθανότητας που εξασφαλίζεται να μοιράζεται σε αυτά ισόποσα. Σε αυτήν την ενότητα θα δούμε πως μπορούμε να μοιράσουμε την πιθανότητα με πιο έξυπνο τρόπο συνδυάζοντας τις τεχνικές έκπτωσης και υποχώρησης. Έτσι ο αλγόριθμος έκπτωσης είναι αρμόδιος για τον καθορισμό της πιθανότητας που θα διατεθεί σε μη εμφανιζόμενα γεγονότα και ο αλγόριθμος της υποχώρησης είναι αρμόδιος για τον τρόπο κατανομής της πιθανότητας σε αυτά.

Θέλοντας να ερμηνεύσουμε την παρουσία της μεταβλητής α που είδαμε προηγουμένως στην εξίσωση πρέπει να τονίσουμε ότι χωρίς αυτή την μεταβλητή το αποτέλεσμα της εξίσωσης δεν θα ήταν μια πραγματική πιθανότητα. Οι εκτιμήσεις μέγιστης πιθανότητας $P(w_n \mid w_{n-N+1}^{n-1})$ είναι πραγματικές πιθανότητες, και αν προσθέσουμε όλες τις πιθανότητες μιας λέξης w_i σε συμφραζόμενα N-gram θα πρέπει να ισχύει ότι:

$$\sum_{i,j} P(w_n \mid w_i w_j) = 1 \tag{2.55}$$

Άρα λοιπόν με την χρήση των εκτιμήσεων μέγιστης πιθανότητας μαζί με την υποχώρηση σε μικρότερης τάξης μοντέλα εφόσον οι πιθανότητες είναι μηδενικές, θα έχουμε την προσθήκη επιπλέον πιθανότητας, με το άθροισμα των πιθανοτήτων μιας συγκεκριμένης λέξης να είναι μεγαλύτερο από 1. Για αυτό λοιπόν θα πρέπει το γλωσσικό μοντέλο υποχώρησης να υποστεί προηγουμένως έκπτωση. Αυτό εξηγεί τον ρόλο της μεταβλητής α και του P^*. Η πιθανότητα P^* προέρχεται από την έκπτωση της εκτίμησης μέγιστης πιθανότητας, ώστε να εξασφαλιστεί ένα μέρος πιθανότητας για την υποχώρηση σε μικρότερης τάξης μοντέλα. Η μεταβλητή α παίζει τον ρόλο του εγγυητή ότι το άθροισμα των πιθανοτήτων των μικρότερης τάξης μοντέλων ισοδυναμεί με την πιθανότητα που εξασφαλίστηκε με την χρήση του αλγόριθμου της έκπτωσης των μεγαλύτερων τάξης μοντέλων.

Άρα ορίζουμε το P^* σαν την μειωμένη εκτίμηση μέγιστης πιθανότητας της υπό-συνθήκης πιθανότητας ενός N-gram:

$$P^*(w_n \mid w_{n-N+1}^{n-1}) = \frac{c^*(w_{n-N+1}^n)}{c(w_{n-N+1}^{n-1})} \tag{2.56}$$

Και επειδή το κατά μέσο όρο ισχύει ότι το $c^* < c$ θα έχουμε ότι η πιθανότητα P^* είναι λίγο μικρότερη από την εκτίμηση μέγιστης πιθανότητας που ισούται με $\frac{c(w_{n-N+1}^n)}{c(w_{n-N+1}^{n-1})}$.

Έτσι λοιπόν η μάζα πιθανότητας που θα περισσέψει θα διατεθεί στα μικρότερης τάξης N-grams, η οποία με την σειρά της θα αναδιανεμηθεί ανάλογα με τα βάρη α. Για να περιγράψουμε το συνολικό ποσό της εναπομένουσας μάζας πιθανότητας χρησιμοποιούμε με μια συνάρτηση β, δηλ. συνάρτηση των συμφραζομένων (N-1)-grams. Για τα συμφραζόμενα (N-1)-grams η συνολική εναπομένουσα μάζα πιθανότητας μπορεί να υπολογιστεί με την αφαίρεση από το 1 της πιθανότητας που προέκυψε από την έκπτωση για όλα τα N grams που ξεκινάνε με αυτά τα συμφραζόμενα.

Άρα $$\beta(w_{n-N+1}^{n-1}) = 1 - \sum_{w_n : c(w_{n-N+1}^n) > 0} P^*(w_n \mid w_{n-N+1}^{n-1}) \tag{2.57}$$

Αυτό συμβολίζει το ποσό της πιθανότητας που θα διατεθεί στα N-1 grams. Το καθένα (N-1)-grams θα πάρει μόνο ένα μέρος της συνολικής πιθανότητας με αποτέλεσμα να χρειάζεται να κανονικοποιηθεί το β με την συνολική πιθανότητα των (N-1)-grams (Chen και Goodman, 1998).

Η τελική εξίσωση που δείχνει πόση μάζα πιθανότητας θα διανεμηθεί από ένα N-gram σε ένα (N-1)-gram παριστάνεται με την συνάρτηση του α .

$$a(w_{n-N+1}^{n-1}) = \frac{\beta(w_n \mid w_{n-N+1}^{n-1})}{\sum_{w_n : c(w_{n-N+1}^{n})=0} P^*(w_n \mid w_{n-N+2}^{n-1})}$$

$$a(w_{n-N+1}^{n-1}) = \frac{1 - \sum_{w_n : c(w_{n-N+1}^{n})>0} P^*(w_n \mid w_{n-N+1}^{n-1})}{1 - \sum_{w_n : c(w_{n-N+1}^{n})>0} P^*(w_n \mid w_{n-N+2}^{n-1})} \tag{2.58}$$

Ας σημειωθεί ότι το α είναι συνάρτηση της ακολουθίας των προηγουμένων λέξεων, w_{n-N+1}^{n-1}. Στην περίπτωση όπου το πλήθος των εμφανίσεων για ένα N-1 gram είναι 0 (δηλ. $c(w_{n-N+1}^{m-1}) = 0$), η σχέση τροποποιείται ως εξής:

$$P^*(w_n \mid w_{n-N+1}^{n-1}) = 0$$

$$\text{και } \tilde{\beta}(w_{n-N+1}^{n-1}) = 1$$

για τον υπολογισμό του P^* θα χρησιμοποιηθεί τεχνική Good-Turing. Για τα τριγράμματα, η μέθοδος υποχώρησης μπορεί να γραφτεί ως εξής:

$$P_{katz}(w_i \mid w_{i-2}w_{i-1}) = \begin{cases} P^*(w_i \mid w_{i-2}w_{i-1}) & \text{εάν } C(w_{i-2}w_{i-1}w_i) > 0 \\ a(w_{i-2}^{i-1})P^*(w_i \mid w_{i-1}) & \text{εάν } C(w_{i-2}w_{i-1}w_i) = 0 \text{ και } C(w_{i-1}w_i) > 0 \\ a(w_{i-1})P^*(w_i) & \text{αλλιώς} \end{cases} \tag{2.59}$$

2.5.5.4 Εξομάλυνση Kneser–Ney

Οι Chen και Goodman (1998) υλοποίησαν τις τεχνικές εξομάλυνσης τόσο σε επίπεδο υποχώρησης όσο και σε επίπεδο παρεμβολής, κάνοντας επιπλέον και μια εκτεταμένη σύγκριση των υφιστάμενων τεχνικών εξομάλυνσης. Ένα από τα ευρήματα τους είναι ότι η τροποποιημένη τεχνική των Kneser και Ney συμπεριφέρεται καλύτερα σε σχέση με τις υπόλοιπες τεχνικές. Η τεχνική των Kneser και Ney είναι κατά κάποιον τρόπο μια προέκταση της απόλυτης έκπτωσης και προσπαθεί να βελτιστοποιήσει τον συνδυασμό των μικρής και μεγάλης τάξης μοντέλων N-grams σε περιπτώσεις όπου υπάρχουν λίγα ή καθόλου μεγάλης τάξης μοντέλα. Αυτή η τεχνική στην βιβλιογραφία περιγράφεται με το παράδειγμα του "San Francisco". Το "San Francisco"είναι πολύ συχνό δίγραμμα, με το μονόγραμμα "Francisco" συνήθως να ακολουθεί

την λέξη "San". Καθώς το μονόγραμμα θα έχει μεγάλη πιθανότητα $P_{katz}(Francisco)$, ένα σχήμα έκπτωσης θα αποδώσει στο δίγραμμα μεγάλη πιθανότητα. Ας δούμε την περίπτωση της υποχώρησης Katz όπου η πιθανότητα θα ισούται με

$$P_{katz}(on \quad Francisco) = \begin{cases} \dfrac{disc(c(on \quad Francisco))}{c(on)} & \varepsilon\acute{\alpha}\nu \quad c(on \quad Francisco) > 0 \\ \alpha(on) \cdot P_{katz}(Francisco) & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \tag{2.60}$$

$$P_{katz}(on \quad Francisco) = \alpha(on) \cdot P_{katz}(Fransisco) \tag{2.61}$$

Κάποιος μπορεί να ισχυριστεί ότι αυτή η ισότητα είναι παραπλανητική ως προς το αποτέλεσμα της από την στιγμή που η λέξη "Francisco" συνήθως βρίσκεται σε ένα συγκεκριμένο περιβάλλον συμφραζόμενων και είναι σπάνιο να βρεθεί σε διαφορετικό περιβάλλον. Με στόχο να βελτιωθεί η πιθανότητα ενός διγράμματος σε τέτοιες περιπτώσεις, η τεχνική των Kneser και Ney δεν χρησιμοποιεί την πιθανότητα του μονογράμματος που είναι αναλογικό του αριθμού εμφάνισης της λέξης, αλλά αντί αυτού, τον αριθμό των διαφορετικών συμφραζομένων της λέξης (περιβάλλον της λέξης). Συνεπώς η πιθανότητα $P_{KN}(on \quad Francisco)$ θα έχει σχετικά μικρότερη τιμή ενώ μια αντίστοιχη πιθανότητα $P_{KN}(on \quad Thuesday)$ θα έχει μεγαλύτερη τιμή λόγω της μεγάλης παρουσίας της λέξης Tuesday σε διαφορετικά περιβάλλοντα.

Η τεχνική των Kneser και Ney χρησιμοποιεί ένα πιο απλό σχήμα έκπτωσης απ' ότι η μέθοδος Katz, αντί λοιπόν να χρησιμοποιούμε εκπτώσεις τύπου Good-Turing χρησιμοποιείται ένας συγκεκριμένος παράγοντας έκπτωσης D. Οι Chen και Goodman πρότειναν μια τροποποίηση της τεχνικής Kneser-Ney ονομάζοντας την σαν τροποποιημένη τεχνική εξομάλυνσης Kneser-Ney, η οποία χρησιμοποιεί τρεις διαφορετικές παραμέτρους D_1, $D_2, \kappa\alpha\iota$ D_{3+} εφαρμόζοντας τα σε N-grams με μια, δυο και τρεις η παραπάνω εμφανίσεις, αντί για ένα γενικό D σε όλες τις περιπτώσεις για μη μηδενικά N-grams.

2.6 Εναλλακτικές τεχνικές υπολογισμού γλωσσικών μοντέλων

2.6.1 Μοντέλα Παράλειψης (Skipping Models)

Ουσιαστικά είναι απίθανο μια αλληλουχία πέντε λέξεων που εμφανίζεται στα δεδομένα εκπαίδευσης να συμβεί και στα πειραματικά δεδομένα. Παρόλ'αυτά είναι λιγότερο απίθανο παρόμοιες αλληλουχίες ίδιων λέξεων να εμφανιστούν. Ειδικότερα, με την χρήση των μοντέλων παράλειψης γίνεται προσπάθεια να υπολογισθούν οι πιθανότητες του τύπου

$P(w_i \mid w_{i-4}w_{i-3}w_{i-1})$ και $P(w_i \mid w_{i-4}w_{i-2}w_{i-1})$. Όπως γίνεται κατανοητό σε μια σειρά πέντε λέξεων οι δυνατοί συνδυασμοί που μπορούν να επιτευχθούν στο πλαίσιο της τεχνικής παράλειψης, είναι πολλοί.

2.6.2 Μοντέλα Ομαδοποίησης (Clustering Models)

Ας θεωρήσουμε την πιθανότητα ενός τριγράμματος $P(Thuesday \mid party \quad on)$. Πιθανόν τα δεδομένα εκπαίδευσης δεν περιλαμβάνουν την φράση "party on Thuesday", αν και άλλες φράσεις του τύπου "party on Friday" μπορεί να περιλαμβάνονται. Συνεπώς αν βάλουμε ομάδες λέξεων και τοποθετήσουμε την λέξη Tuesday στην ομάδα λέξεων Weekday μπορούμε να υπολογίσουμε την πιθανότητα αποσυνθέτοντας την ως εξής

$$\begin{aligned} & p(Thuesday \mid party \quad on) \\ & = p(Weekday \mid party \quad on) \times p(Thuesday \mid party \quad on \quad Weekday) \end{aligned} \tag{2.62}$$

2.6.3 Μοντέλα Μνήμης (Caching Models)

Εάν θεωρηθεί ότι ένας ομιλητής χρησιμοποιεί μια λέξη είναι πιθανόν ότι αυτή τη λέξη θα την χρησιμοποιήσει και στο μέλλον. Αυτή η παρατήρηση είναι η αρχή των μοντέλων μνήμης (Kuhn, 1988; Kuhn and De Mori, 1990; Kuhn and De Mori, 1992; Kupiec, 1989;Jelinek et al., 1991). Ειδικότερα, στην περίπτωση ενός πίνακα μνήμης μονογραμμάτων, σχηματίζεται ένα μοντέλο μονογραμμάτων από τις πιο πρόσφατες λέξεις (αυτές που συνοδεύουν ένα συγκεκριμένο άρθρο ή αν δεν υπάρχει κάτι τέτοιο ένα συγκεκριμένο αριθμό από προγενέστερες λέξεις). Αυτό το μοντέλο μνήμης μονογραμμάτων μπορεί να παρεμβληθεί γραμμικά με ένα συμβατικό N-gram. Επίσης μπορεί να χρησιμοποιηθούν και άλλα είδη από μνήμες, παράδειγμα ένα τρίγραμμα εξομάλυνσης μπορεί να σχηματιστεί από προηγούμενες λέξεις και να παρεμβληθεί.

2.6.4 Λέξεις πρόκλησης (Word triggers)

Μια γενίκευση της ιδέας των μοντέλων μνήμης για την συσχέτιση 2 διαφορετικών λέξεων οδηγεί σε μια νέα περιοχή έρευνας που ονομάζεται λέξεις πρόκλησης (Rosenfeld 1996, Beeferman et al. 1997). Επί της αρχής, οι σχέσεις μεταξύ οποιουδήποτε ζευγαριού λέξεων ή φράσεων μπορεί να μοντελοποιηθεί. Δυστυχώς, οι υπολογιστικές απαιτήσεις για την εκπαίδευση τέτοιου μοντέλου αυξάνει υπέρ-γραμμικά σε σχέση με τον αριθμό των ανεξάρτητων μοντελοποιημένων ζευγαριών λέξεων πρόκλησης, κάνοντας ακόμη και αποτρεπτική την χρήση ενός μικρού αριθμού ζευγαριών.

2.6.5 Μοντέλα μίξης προτάσεων (Sentence mixture models)

Οι Iyer και Ostendorf (1999) παρατήρησαν ότι μέσα σε ένα σώμα κειμένων υπάρχουν διαφορετικοί τύποι προτάσεων που μπορεί να ομαδοποιηθούν βάσει του θέματος τους, του ύφους τους και άλλων κριτηρίων. Ας υποθέσουμε ότι στα δεδομένα της Wall Street Journal υπάρχουν τρία διαφορετικά είδη προτάσεων, π.χ προτάσεις οικονομικού ενδιαφέροντος (με λέξεις αριθμούς και ονόματα μετοχών), προτάσεις επιχειρηματικού ενδιαφέροντος (με λέξεις συνενώσεις εταιρειών, τρόποι προβολής) και προτάσεις με διάφορες ιστορίες. Έτσι μπορεί να υπολογισθεί η πιθανότητα μιας πρότασης βάσει κάθε ενός τύπου πρότασης και μετά να ληφθεί ένα άθροισμα με διαφορετικά βάρη για κάθε μια από αυτές της πιθανότητες.

2.6.6 Προσαρμογή γλωσσικού μοντέλου

Σε γενικές γραμμές η τεχνική των N-grams δίνει καλά αποτελέσματα όταν παρέχεται ικανοποιητικός όγκος δεδομένων εκπαίδευσης σε ένα συγκεκριμένο αντικείμενο. Όταν τα δεδομένα εκπαίδευσης είναι διαφορετικά από τα δεδομένα δοκιμής τότε τα αποτελέσματα δεν είναι ικανοποιητικά και η απόδοση της τεχνικής υστερεί. Για να αντιμετωπιστεί το πρόβλημα αυτό έχουν γίνει πολλές προσπάθειες να χρησιμοποιηθούν μείγματα γλωσσικών μοντέλων ώστε να βελτιωθεί η απόδοση τους. Μια τυπική μέθοδος μίξης των γλωσσικών μοντέλων είναι αυτή που απορρέει από τον συνδυασμό γλωσσικών μοντέλων που έχουν εκπαιδευτεί σε ένα μικρό κομμάτι δεδομένων εκπαίδευσης, συγκεκριμένου αντικειμένου με γλωσσικά μοντέλα που βασίζονται σε μεγαλύτερα και λιγότερο εξειδικευμένα δεδομένα εκπαίδευσης. Για παράδειγμα ένα γλωσσικό μοντέλο που βασίζεται σε μεγάλο όγκο δεδομένων από εφημερίδες μπορεί να παρεμβληθεί με γλωσσικά μοντέλα που έχουν δομηθεί πάνω σε λιγοστά δεδομένα από κείμενα ραδιοφωνικής και τηλεοπτικής ειδησεογραφίας. Τα βάρη που θα χρησιμοποιηθούν στην παρεμβολή μπορούν να υπολογισθούν με την χρήση κάποιων δεδομένων, ή αλλιώς μπορούν να προσαρμοστούν δυναμικά σαν συνάρτηση της τρέχουσας ιστορίας. Η εφαρμογή της τρέχουσας ιστορίας εμφανίζεται και σε τεχνικές βελτίωσης των απλών γλωσσικών μοντέλων όπως αυτό το μοντέλο μνήμης που περιγράφεται παραπάνω.

2.7 Τεχνικές στατιστικής μοντελοποίησης της γλώσσας με την χρήση γραμματικής πληροφορίας

2.7.1 Μοντέλα βασισμένα σε μέρη του λόγου (POS)

Σε ένα μοντέλο N-gram, το λεξικό είναι μια λίστα από λέξεις που ανήκουν σε πάνω από μια κατηγορίες. Το σίγουρο είναι ότι σε μια γλώσσα οι λέξεις σχηματίζουν σύνθετες και πολλές

φορές μη κατανοητές λεξικές σχέσεις. Όπως καταλαβαίνει κανείς η λέξη "THUESDAY" είναι πιο κοντά στην λέξη "WEDNESDAY" παρά στην λέξη "CHAIR".

Η πιο απλή και αποτελεσματική μέθοδος για να αναδειχθεί η λεξική σχέση των λέξεων είναι η χρήση της πληροφορίας του παρέχεται με την ιδιότητα του Μέρους του Λόγου, Part of Speech (POS). Άρα η πληροφορία που φέρει ένας POS tagger μπορεί να συμπεριληφθεί στον υπολογισμό των N-grams όπως αναφέρει ο Jelinek (1989). Παράδειγμα για ένα τρίγραμμα μπορούμε να γράψουμε το εξής:

$$P(w_n \mid w_{n-2}w_{n-1}) = P(w_n \mid POS_n) \cdot P(POS_n \mid POS_{n-2}, POS_{n-1}) \tag{2.63}$$

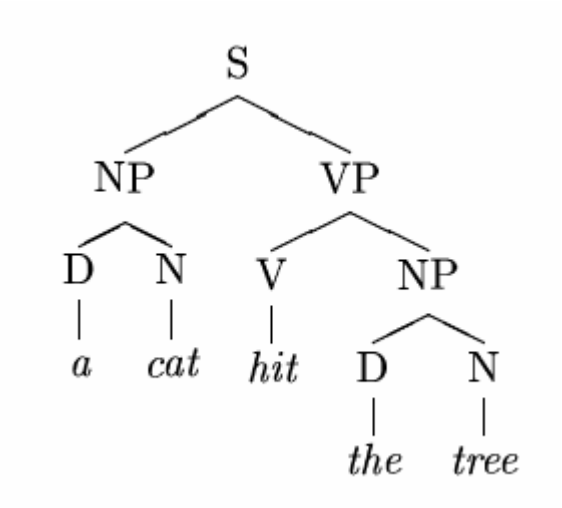
Όπου POS_n είναι η POS κατηγορία του w_n.

Το κύριο κίνητρο της ενσωμάτωσης ενός POS σε N-grams είναι να μειωθεί ο αριθμός των παραμέτρων και κατά συνέπεια η διασπορά της εκτίμησης. Ένα από τα πιο πρακτικά προβλήματα της γλώσσας είναι η πολυσημία των λέξεων με αποτέλεσμα η κατάταξη μιας λέξης σε μια κατηγορία βάσει του POS να είναι πολύ δύσκολη. Ένας πολύ σύγχρονος POS αναλυτής παρουσιάζει ποσοστό ακρίβειας γύρω στο 95-97% κάτω από ιδανικές συνθήκες. Όμως, διάφορα πειράματα αποδεικνύουν ότι τα μοντέλα αυτά συνήθως δεν αποδίδουν ικανοποιητικά, όπως αυτό φαίνεται από την βελτίωση του βαθμού αβεβαιότητας σε σχέση με τα κανονικά N-grams που βασίζονται μόνο σε λέξεις. Το συμπέρασμα που εξάγεται είναι ότι αυτό που είναι χρήσιμο στην γλωσσική πληροφορία δεν είναι και κατ' επέκταση αποτελεσματικό στις τεχνικές πρόβλεψης λέξεων.

2.7.2 Συντακτική δομή

Πολλές προσπάθειες έχουν γίνει στο να ενσωματωθούν πληροφορίες σχετικά με την σύνταξη στο γλωσσικό μοντέλο. Τα N-grams μοντέλα εκμεταλλεύονται τη στατιστική συν-εμφάνισης των λέξεων στα σώματα κειμένων, αλλά δεν έχουν αντίληψη της γραμματικής σε αποστάσεις μεγαλύτερης από N. Ένα εναλλακτικό γλωσσικό μοντέλο, είναι η Ανεξάρτητη από τα Συμφραζόμενα Πιθανοτική Γραμματική (Probabilistic Context Free Grammar) (PCFG), η οποία αποτελείται από μια Ανεξάρτητη από τα Συμφραζόμενα Γραμματική (CFG) μέσα στην οποία κάθε κανόνας αναδιατύπωσης έχει μια συσχετισμένη πιθανότητα. Το άθροισμα όλων των πιθανοτήτων για όλους του κανόνες με την ίδια την αριστερή πλευρά ισούται με 1. Στο μοντέλο PCFG, η πιθανότητα μιας συμβολοσειράς $P(w_n)$, είναι απλώς το άθροισμα των πιθανοτήτων των συντακτικών δένδρων της. Η πιθανότητα ενός δεδομένου δένδρου είναι το γινόμενο των πιθανοτήτων όλων των κανόνων που απαρτίζουν τους κόμβους του δένδρου. Το σχήμα δείχνει τον τρόπο υπολογισμού της πιθανότητας μιας πρότασης. Μπορούμε να υπολογίσουμε αυτή την

πιθανότητα χρησιμοποιώντας έναν διαγραμματικό αναλυτή CFG για να απαριθμήσουμε τις δυνατές αναλύσεις και στην συνέχεια προσθέτουμε τις πιθανότητες.



Σχήμα 2.2 Η συντακτική ανάλυση της πρότασης "a cat hit the tree".

S	→	NP VP	(1.0)
VP	→	V NP	(1.0)
NP	→	D N	(1.0)
D	→	*a*	(0.6)
D	→	*the*	(0.4)
N	→	*boat*	(0.5)
N	→	*cat*	(0.3)
N	→	*tree*	(0.2)
V	→	*hit*	(0.7)
V	→	*missed*	(0.3)

Σχήμα 2.3 Οι πιθανότητες που απεικονίζονται δείχνουν την συχνότητα εμφάνισης του κάθε κανόνα.

Το πρόβλημα με τις PCFG είναι ότι είναι ανεξάρτητες από τα συμφραζόμενα. Αυτό σημαίνει ότι η διαφορά του τριγράμματος $P(eat\ \ a\ \ banana)$ και του τριγράμματος $P(eat\ \ a\ \ bandanna)$ είναι μόνο οι διαφορετικές πιθανότητες των $P(banana)$ και $P(bandanna)$ και όχι από την σχέση της λέξης eat με τα αντίστοιχα αντικείμενα. Για να φθάσουμε σε αυτού του είδους τη σχέση πρέπει να χρησιμοποιήσουμε ένα μοντέλο που εξαρτάται από τα συμφραζόμενα, όπως η λεξικοποιημένη PCFG (Lexicalised PCFG). Σε αυτό το μοντέλο η κεφαλή μιας φράσης μπορεί να παίξει ρόλο στην πιθανότητα μιας φράσης που την περιέχει. Αν έχουμε αρκετά δεδομένα εκπαίδευσης, μπορούμε να κάνουμε τον κανόνα για το

ΡΦ→ΡΦ ΟΦ να ορίζεται υπό συνθήκη ως προς την κεφαλή της ΡΦ(eat) και της ΟΦ(banana). Συνεπώς η λεξικοποιημένη PCFG μπορεί να συλλαμβάνει μερικούς από τους περιορισμούς συν-εμφάνισης των N-grams μοντέλων, μαζί με τους γραμματικούς περιορισμούς των μοντέλων CFG. Μια ενδιαφέρουσα προσπάθεια να συνδυαστούν τα N-grams με PCFG αναφέρθηκε από τον Miller (1995), χωρίς όμως να αναφερθεί ουδεμία σημαντική βελτίωση στην απόδοση των N-grams.

2.8 Χαρακτηριστικά του γλωσσικού μοντέλου που εφαρμόζεται στην παρούσα εργασία

Το γλωσσικό μοντέλο που χρησιμοποιείται στο πλαίσιο αυτής της εργασίας έχει βασισθεί στον υπολογισμό των N-grams μοντέλων (Jelinek, 1995). Όπως γνωρίζουμε τα N-grams παρέχουν την δυνατότητα εκτίμησης της πιθανότητας εμφάνισης μιας ακολουθίας λέξεων και παράλληλα μπορούν να κωδικοποιήσουν την συντακτική πληροφορία των προτάσεων κάθε γλώσσας. Η ισχύς των N-grams μοντέλων βασίζεται στην τοπική αλληλο-εξάρτηση των λέξεων μιας πρότασης, και αυτή φαίνεται κυρίως σε γλώσσες που η σειρά των λέξεων ακολουθεί μια αυστηρή διάταξη. Συνεπώς ένα γλωσσικό μοντέλο μπορεί να περιγράψει στατιστικά τους περιορισμούς που εμφανίζονται στην σειρά των λέξεων μιας πρότασης. Έτσι μια ακολουθία λέξεων που η σειρά της δεν είναι αρκετά συχνή εμφανίζει μικρή πιθανότητα ενώ σε άλλη περίπτωση ακολουθίες λέξεων με συχνές εμφανίσεις σε ένα σώμα κειμένου θα συνοδεύονται από σχετικά υψηλές πιθανότητες.

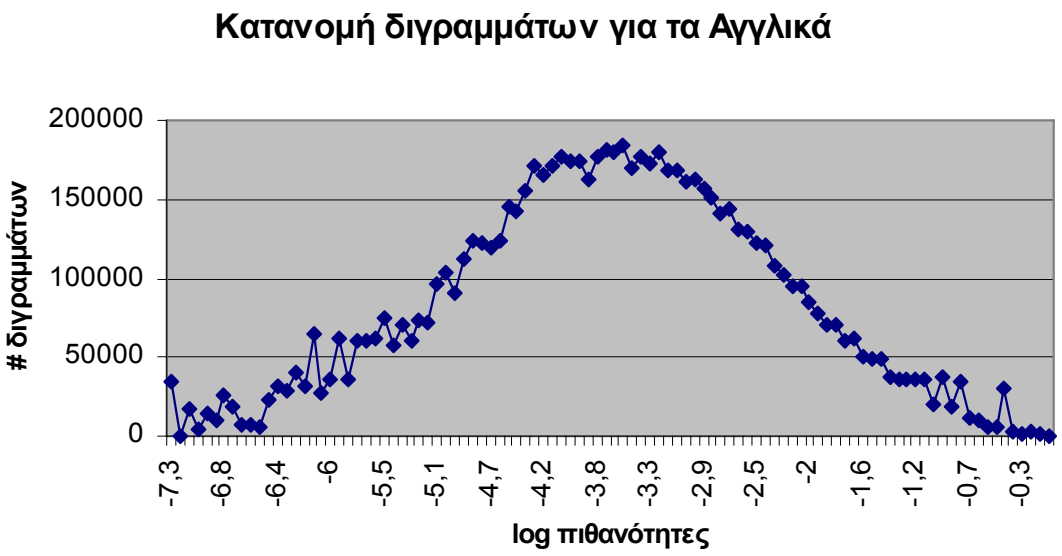
Ένα σημαντικό πρόβλημα με την τεχνική των συμβατικών N-grams μοντέλων είναι ότι πρέπει να εκπαιδευτούν από ένα συγκεκριμένο και πεπερασμένο σώμα κειμένου με συνέπεια κάποια αποδεκτά N-grams να λείπουν από αυτό και να λαμβάνουν μηδενική πιθανότητα εμφάνισης. Για αυτόν τον λόγο πρέπει να χρησιμοποιηθούν τεχνικές που κατά κάποιο τρόπο να εξομαλύνουν τις πιθανότητες των N-grams. Στην περίπτωση μας θα χρησιμοποιήσουμε την τεχνική Good-Turing (1953) και την τεχνική Katz backoff (1987). Το υφιστάμενο γλωσσικό μοντέλο εκπαιδεύτηκε με εκτενή σώματα κείμενων, με την χρήση εξειδικευμένων εργαλείων παραγωγής γλωσσικών μοντέλων (Stolcke, 2002) και αποτελείται από ακολουθίες λέξεων που καλούνται διγράμματα (2 λέξεις) και τριγράμματα (3 λέξεις).

Το σώμα κειμένου για τη Αγγλική γλώσσα περιλαμβάνει 6.25 εκατομμύρια προτάσεις και 100 εκατομμύρια στοιχεία-λέξεις, ενώ το αντίστοιχο για την Ελληνική γλώσσα 130 εκατομμύρια στοιχεία-λέξεις. Οι παρακάτω πίνακες παρουσιάζουν το πλήθος των μονογραμμάτων, διγραμμάτων και τριγραμμάτων για κάθε μια γλώσσα ξεχωριστά. Στα παρακάτω σχήματα φαίνεται αναλυτικά η κατανομή των διαφορετικών N-grams για την

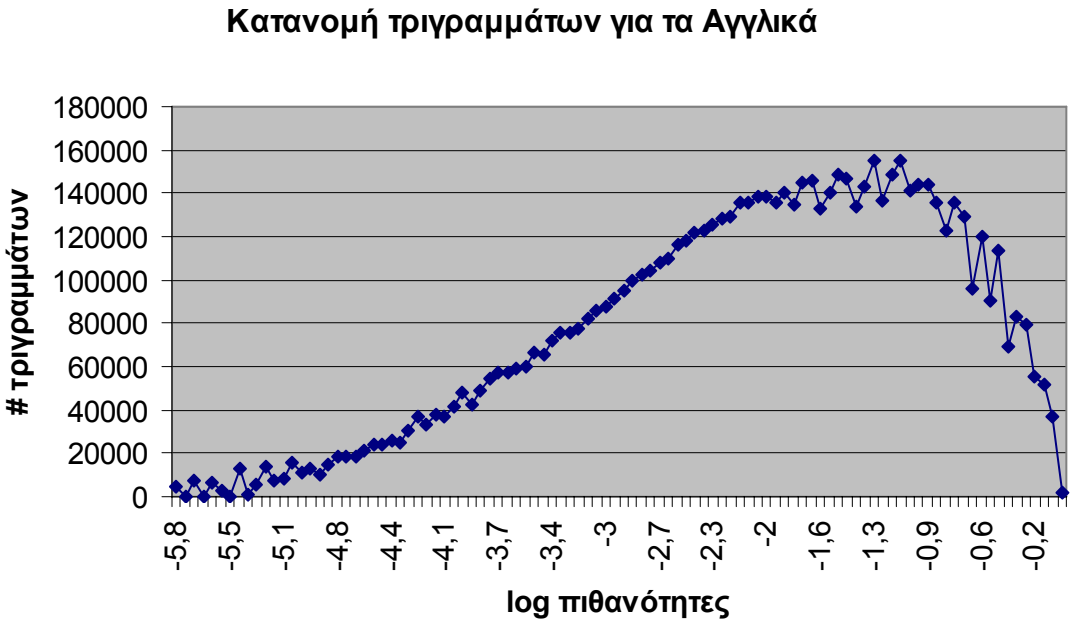
Αγγλική και Ελληνική γλώσσα. Η κατανομή των διγραμμάτων και τριγραμμάτων γίνεται βάσει των λογαριθμικών πιθανοτήτων τους.

Στοιχεία του γλωσσικού μοντέλου για την Αγγλική γλώσσα	Πλήθος στοιχείων
Μονογράμματα	126062
Διγράμματα	8166674
Τριγράμματα	8033315

Πίνακας 2.2 Το πλήθος των στοιχείων του γλωσσικού μοντέλου για την Αγγλική γλώσσα.



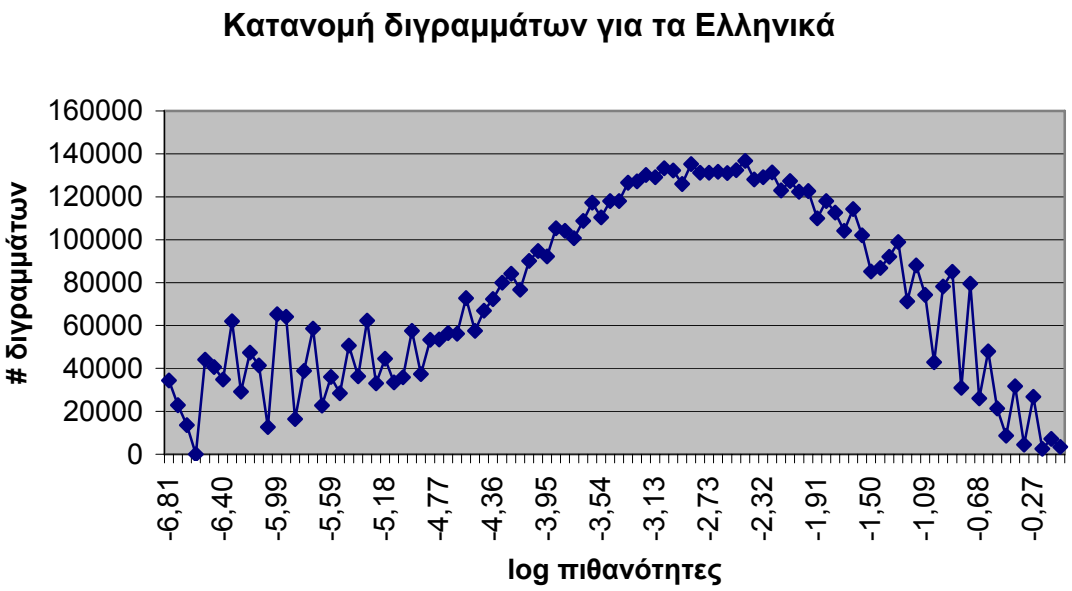
Σχήμα 2.4 Η κατανομή των διγραμμάτων βάσει του λογάριθμου της πιθανότητας τους.



Σχήμα 2.5 Η κατανομή των τριγραμμάτων βάσει της λογαριθμικής πιθανότητας.

Στοιχεία του γλωσσικού μοντέλου για την Ελληνική γλώσσα	Πλήθος στοιχείων
Μονογράμματα	317403
Διγράμματα	7652952
Τριγράμματα	7029801

Πίνακας 2.3 Το πλήθος των στοιχείων του γλωσσικού μοντέλου για την Ελληνική γλώσσα.



Σχήμα 2.6 Η κατανομή των διγραμμάτων βάσει της λογαριθμικής πιθανότητας.

Σχήμα 2.7 Η κατανομή των τριγραμμάτων βάσει της λογαριθμικής πιθανότητας

Κεφάλαιο 3

3 Η επίδραση του συναισθηματικά προσανατολισμένου γλωσσικού μοντέλου στην αναγνώριση φωνής

3.1 Εισαγωγή

Τα τελευταία χρόνια στο χώρο των διαλογικών συστημάτων, υπάρχει μεγάλο ενδιαφέρον για τις φωνητικές διεπαφές που λαμβάνουν υπόψη τους το συναίσθημα του ομιλητή (André Dybkjaer and Minker, 2004). Από την στιγμή που το συναίσθημα παίζει σημαντικό ρόλο στην ανθρώπινη επικοινωνία, είναι πολύ φυσικό το να προσπαθεί κανείς να συμπεριλάβει αυτή την πληροφορία σ' αυτό που λέγεται πραγματικό σύστημα επεξεργασίας φυσικής γλώσσας. Με τον όρο σύστημα φυσικής επικοινωνία ανθρώπου-μηχανής εννοούμε ένα σύστημα φιλικό προς τον χρήστη που του παρέχει την δυνατότητα της απευθείας συνομιλίας με τον υπολογιστή σε τέτοιο επίπεδο που δεν θα χρειάζεται να μετριάζει την συμπεριφορά του, αλλά μπορεί ελεύθερα να εκφράζει τα συναισθήματα του μέσα από τα λεγόμενα του (Cowie and Schröder, 2005).

Αν και η γνώση σε αυτήν την περιοχή αναπτύσσεται ταχύτατα, σε αρκετά σημεία υστερεί και είναι περιορισμένη ώστε να μπορέσει να βελτιώσει την επικοινωνία ανθρώπου μηχανής. Αυτό το κεφάλαιο φωτίζει δυο από τα προβλήματα που καθιστούν την επεξεργασία του συναισθηματικού λόγου, μια δύσκολη διαδικασία.

Θέλοντας να βρούμε τα αίτια για την περιορισμένη απόδοση ανάλογων συστημάτων που σχετίζονται με την επεξεργασία και αναγνώριση συναισθηματικού λόγου, πρέπει να αναφέρουμε πρώτον το ότι δεν υπάρχει ακόμη ιδιαίτερη γνώση αυτού που ονομάζεται συναισθηματικά φορτισμένος λόγος. Όπως γίνεται κατανοητό για να υπάρξει πρόοδος στην αναγνώριση συναισθηματικού λόγου πρέπει να δημιουργηθούν αρχικά βάσεις δεδομένων με ανάλογο περιεχόμενο. Η μελέτη και η κατανόηση των γλωσσικών και των παρα-γλωσσικών

δομών του συναισθηματικού λόγου είναι εφικτή μόνο μέσα από την συγκέντρωση ανάλογου υλικού. Μέχρι σήμερα η όποια έρευνα γίνεται στον τομέα αυτό αφορά την επεξεργασία δεδομένων που προκύπτουν από την καταγραφή κάποιων ηθοποιών που αποδίδουν συγκεκριμένους ακραίους συναισθηματικά ρόλους (Douglas-Cowie et al., 2003a). Φυσικά, αυτό αποτελεί μια λύση στο πρόβλημα που λέγεται συγκέντρωση υλικού με συναισθηματικό λόγο. Αλλά το ερώτημα που τίθεται είναι κατά πόσον το περιεχόμενο μιας τέτοιας βάση δεδομένων συμβαδίζει με την καθημερινή συμπεριφορά ενός ανθρώπου (André, 2004). Θέλοντας οι ερευνητές να μαζέψουν πιο φυσικό υλικό επιχειρήθηκε να γίνει καταγραφή των κλήσεων σε τηλεφωνικά κέντρα – (call centers) χρησιμοποιώντας πραγματικά δεδομένα και δεδομένα προσομοίωσης (Batliner et al., 2003; Batliner et al., 2004). Είναι φανερό όμως ότι τέτοιο είδους εκφράσεις περιέχουν μικρή ποικιλία συναισθημάτων. Αντίθετα με όλα τα παραπάνω, η εργασία αυτή επικεντρώνεται στην χρήση ενός υλικού με συναισθηματικά φορτισμένο λόγο που παράγεται από χρήστες που συνδιαλέγονται με ένα σύστημα πρόκλησης συναισθημάτων όπως είναι το Sensitive Artificial Listener (SAL) (Douglas-Cowie et al., 2003b).

Το δεύτερο σημείο το οποίο έχει καθηλώσει την έρευνα γύρω από την αναγνώριση του συναισθητικού λόγου είναι ότι κυρίως επικεντρώνεται στην αναγνώριση της συναισθηματικής και μόνον κατάστασης του ομιλητή. Όπως γίνεται κατανοητό αυτό είναι μόνο το μισό του συνολικού προβλήματος. Το άλλο μισό έχει σχέση με το κλασσικό πρόβλημα της αυτόματης αναγνώρισης λόγου που έχει σαν στόχο όχι μόνο να αναγνωρίσει την συναισθηματική κατάσταση του ομιλητή αλλά και να μπορέσει να ανακτήσει το σύνολο της γλωσσικής πληροφορίας που φέρει ο συναισθηματικά φορτισμένος λόγος.

Ο υπαγορευτικός και ο μη-ελεύθερος λόγος που αποδίδεται με προσεκτικό τρόπο χωρίς υπερβολές, (Labov, 1974) μπορεί να αναγνωριστεί με αρκετά μεγάλο ποσοστό επιτυχίας χρησιμοποιώντας την σύγχρονη τεχνολογία της αναγνώρισης φωνής. Αλλά παρά την σημαντική βελτίωση στην απόδοση τέτοιων συστημάτων μεγάλου λεξιλογίου, ακόμη οι ερευνητές βρίσκονται σε πρώιμο στάδιο στον τομέα της αναγνώρισης του συναισθηματικού λόγου. Οι αρχικές ενδείξεις αναφέρουν ότι πράγματι είναι πολύ δύσκολο να ανακτηθεί η γλωσσική πληροφορία του εκάστοτε ομιλητή που παρουσιάζει οξυμμένη συναισθηματική συμπεριφορά. Το αντικείμενο αυτής της εργασίας είναι μεταξύ άλλων να βελτιστοποιήσει την απόδοση ενός κλασσικού συστήματος αναγνώρισης φωνής όταν γίνεται χρήση συναισθηματικά φορτισμένου λόγου.

Σε αυτό το κεφάλαιο θα ασχοληθούμε με μια άλλη στρατηγική αντιμετώπισης του προβλήματος, η οποία είναι συμπληρωματική της μεθόδου που υλοποιεί ο Polzin και Waibel (1998) χρησιμοποιώντας την πληροφορία της προσωδίας. Η όλη ιδέα αυτής της στρατηγικής

είναι ότι το συναίσθημα ενός ομιλητή δεν εκφράζεται μόνο από την χροιά του λόγου του, αλλά και από τις λέξεις που χρησιμοποιεί και από την σειρά με την οποία τις διατυπώνει. Όπως γνωρίζουμε η μεγάλη επιτυχία όλων των συστημάτων αναγνώρισης φωνής οφείλεται στην χρήση και κατασκευή γλωσσικών μοντέλων, τα οποία μοντελοποιούν την γλώσσα με σκοπό να περιοριστεί το πρόβλημα με την αμφισημία ομόηχων λέξεων. Το ζητούμενο λοιπόν είναι να χρησιμοποιηθεί και στην αναγνώριση συναισθηματικού λόγου κάποιο ανάλογο γλωσσικό μοντέλο. Για την δημιουργία ενός γλωσσικού μοντέλου απαιτείται ένα μεγάλο σώμα κειμένου που θεωρητικά περιλαμβάνει το μεγαλύτερο όγκο των λέξεων και των συνδυασμών τους. Στον τομέα της αναγνώρισης υπαγορευτικού λόγου η χρήση σωμάτων κειμένων που συμπεριλαμβάνονται σε εφημερίδες περιοδικά και ακόμα στο Internet αποτελεί μονόδρομο για την επιτυχία ενός ανάλογου συστήματος. Άρα λοιπόν πιθανολογείται ότι η επιτυχία και ενός συστήματος αναγνώρισης συναισθηματικού λόγου μπορεί να εξαρτάται και από το είδος των σωμάτων κειμένων που θα χρησιμοποιηθούν για την εκπαίδευση του. Πώς όμως θα συλλέξουμε κείμενα με συναισθηματικό περιεχόμενο; Πώς θα ορίσουμε ποια κείμενα έχουν μεγαλύτερη συναισθηματική βαρύτητα από άλλα, ώστε να τα επιλέξουμε; Η απάντηση σε αυτό το ερώτημα θα ήταν απλή αν αναλογιστούμε ότι μπορεί να επιτευχθεί με την μετεγγραφή ηχητικών σημάτων με συναισθηματικά φορτισμένες εκφράσεις, που όμως είναι πολύ χρονοβόρα διαδικασία. Φυσικά εκτός από το ότι είναι χρονοβόρα και επίπονη διαδικασία είναι και πολύ δαπανηρή αν αναλογιστούμε το κόστος μιας τέτοιας διαδικασίας σε επίπεδο άνθρωπο-ωρών. Άρα η λύση αυτού του προβλήματος δεν είναι άλλη από την αυτόματη συλλογή κειμένων με συναισθηματικό υπόβαθρο έτσι ώστε να αποτελέσουν την βάση για την εκπαίδευση των στατιστικών γλωσσικών μοντέλων.

Το κεφάλαιο αυτό εξηγεί πώς ένα συναισθηματικά προσανατολισμένο γλωσσικό μοντέλο με έμφαση στο συναίσθημα μπορεί να παραχθεί από ένα κανονικό σώμα κειμένου. Το να μετεγγραφεί συναισθηματικά φορτισμένος λόγος για την δημιουργία κειμένου με στόχο την ενίσχυση του γλωσσικού μοντέλου είναι όπως ήδη αναφέραμε πολύ χρονοβόρο. Αυτή η εργασία παρουσιάζει έναν εναλλακτικό τρόπο για να δημιουργηθεί συναισθηματικά φορτισμένο κείμενο με την χρήση ενός υφιστάμενου κειμένου όπως αυτό του Εθνικού Βρετανικού σώματος Κειμένου. Γι' αυτόν το λόγο χρησιμοποιείται ένα λεξιλόγιο με όρους-λέξεις που παρουσιάζουν συναισθηματικό φορτίο σαν εργαλείο για την εξαγωγή προτάσεων από το Εθνικό Βρετανικό Σώμα Κειμένου, με συναισθηματικό περιεχόμενο. Το παραγόμενο κείμενο δημιουργείται από τον συνδυασμό του Εθνικού Βρετανικού σώματος Κειμένου και των εξαγόμενων προτάσεων με συναισθηματικό περιεχόμενο. Συνεπώς το κείμενο αυτό χρησιμοποιείται για την παραγωγή γλωσσικού μοντέλου με έμφαση στο συναίσθημα, που με την σειρά του το γλωσσικό μοντέλο

θα χρησιμοποιηθεί για να βελτιώσει την απόδοση του κλασσικού συστήματος αναγνώρισης φωνής με δεδομένα εισόδου συναισθηματικά φορτισμένες εκφράσεις.

Το κεφάλαιο 3 οργανώνεται ως εξής: στην ενότητα 3.2 γίνεται σύντομη περιγραφή του κλασσικού συστήματος αναγνώρισης φωνής που χρησιμοποιείται για την καταγραφή του υπαγορευτικού λόγου. Στην επόμενη ενότητα παρουσιάζονται τα προβλήματα που ανακύπτουν κατά την αναγνώριση του συναισθηματικού λόγου και οι τεχνικές που έχουν μέχρι σήμερα χρησιμοποιηθεί για να επιλύσουν το πρόβλημα. Μια σύντομη επισκόπηση στον τομέα έρευνας γύρω από τα συναισθήματα και τις συναισθηματικές καταστάσεις παρουσιάζεται στην ενότητα 3.4, ενώ στην ενότητα 3.5 αναλύονται τα βήματα που θα ακολουθηθούν για την επαύξηση του γλωσσικού μοντέλου. Τα πειραματικά δεδομένα μαζί με τα αποτελέσματα αναλύονται στην ενότητα 3.6 ενώ η εφαρμογή της αναγνώρισης φωνής σε πολυμεσικά συστήματα ανάλυσης συναισθηματική κατάστασης περιγράφεται στην ενότητα 3.7. Τέλος τα συμπεράσματα της μεθόδου ενίσχυσης του γλωσσικού μοντέλου παρατίθενται στην ενότητα 3.8.

3.2 Αρχιτεκτονική του συστήματος αναγνώρισης φωνής

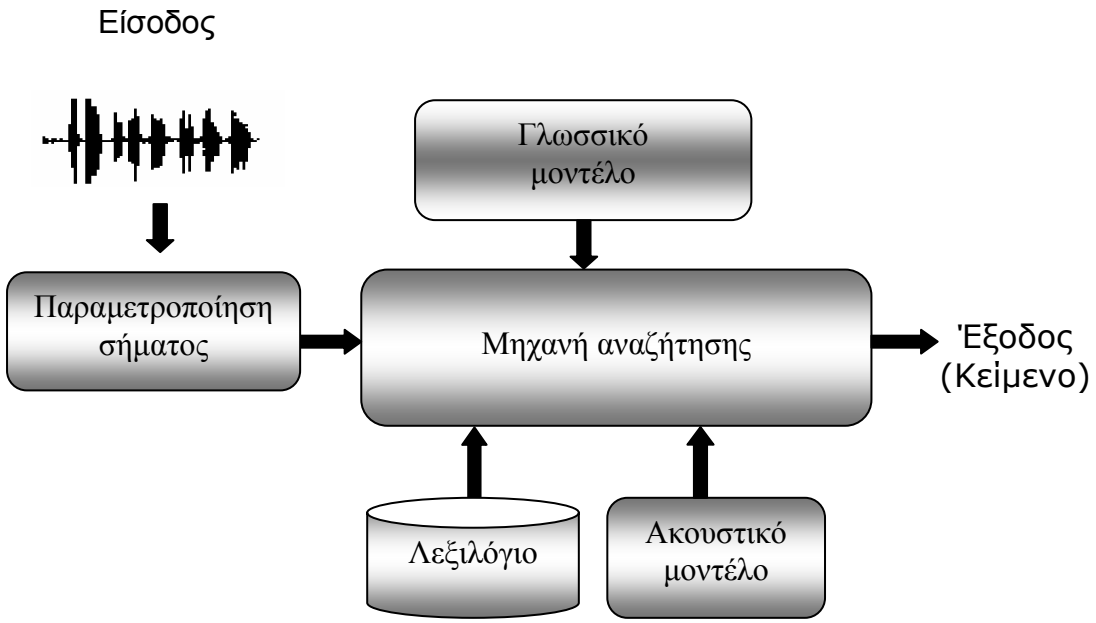
3.2.1 Περιγραφή του συστήματος

Το σύστημα αναγνώρισης φωνής μεγάλου λεξιλογίου που θα χρησιμοποιηθεί εν προκειμένω, βασίζεται στην τεχνική των κρυφών Μαρκοβιανών Μοντέλων (HMM), (Jelinek, 1995). Η άγνωστη είσοδος στο σύστημα μετατρέπεται σε μια ακολουθία ακουστικών διανυσμάτων $\mathrm{Y} = y_1, y_2, ..., y_n$, μέσω του υποσυστήματος της μηχανής αναγνώρισης φωνής που καλείται υποσύστημα παραμετροποίησης του σήματος. Ο στόχος ενός συστήματος αναγνώρισης φωνής είναι να ορίσει την πιο πιθανή ακολουθία λέξεων $\hat{W}$ δεδομένου του παρατηρούμενου ακουστικού σήματος Y, η τεχνική αυτή βασίζεται στο κανόνα του Bayes περί αποσύνθεσης της πιθανότητας $\mathrm{P}(W \mid \mathrm{Y})$ σε δύο όρους, και αυτή η σχέση μπορεί να γραφτεί ως εξής

$$\hat{W} = \arg\max_{W} P(W/Y) = \arg\max_{W} \frac{P(W)P(Y/W)}{P(Y)} \tag{3.1}$$

Η πιθανότητα $\mathrm{P}(W)$ ορίζεται απευθείας απο το γλωσσικό μοντέλο. Η πιθανότητα $P(Y \mid W)$ υπολογίζεται με την χρήση ενός Κρυφού Μαρκοβιανού Μοντέλου που αναπαριστά τον σχηματισμό της λέξης W με απλά κρυφά Μαρκοβιανά Μοντέλα που μοντελοποιούν τα

φωνήματα σύμφωνα με το τρόπο προφοράς κάθε λέξης. Το επόμενο σχήμα αναπαριστά τα πιο σημαντικά μέρη του συστήματος αναγνώρισης φωνής.



Σχήμα 3.1 Η αρχιτεκτονική ενός κλασσικού συστήματος αναγνώρισης φωνής που χρησιμοποιείται στο πλαίσιο αυτής της εργασίας. Το γλωσσικό μοντέλο είναι το στοιχείο του συστήματος στο οποίο θα εφαρμοσθεί μια νέα τεχνική για την βελτίωση της απόδοσης του.

3.2.2 Παραμετροποίηση σήματος

Η πρωταρχική λειτουργία του υποσυστήματος της παραμετροποίησης του σήματος είναι να διαχωρίσει το σήμα σε τμήματα και για καθένα τέτοιο τμήμα να παράγει μια εκτίμηση του λειασμένου φάσματος. Η απόσταση μεταξύ δυο τμημάτων είναι 10 msecs και τα τμήματα συνήθως αλληλο-καλύπτονται για να δώσουν ένα μεγαλύτερο παράθυρο ανάλυσης της τάξεως των 25 ms. Τα δείγματα του παράθυρου ανάλυσης πολλαπλασιαζόμενα με ένα παράθυρο Hamming χρησιμοποιούνται για τον υπολογισμό των συντελεστών Mel-Frequency Cepstral Coefficients (MFCCs). Οι συντελεστές αυτοί χρησιμοποιούνται για να περιγράψουν το φασματικά χαρακτηριστικά του σήματος.

3.2.3 Ακουστικό μοντέλο

Ο σκοπός του ακουστικού μοντέλου είναι να δώσει μια μέθοδο υπολογισμού της πιθανότητας μιας οποιασδήποτε ακολουθίας ακουστικών διανυσμάτων δεδομένης μίας λέξης W. Για ένα σύστημα αναγνώρισης φωνής μικρού λεξιλογίου και αναγνώρισης ψηφίων, μπορεί να υπάρχουν μοντέλα για κάθε μια λέξη, επιτυγχάνοντας έτσι πολύ καλή απόδοση. Αλλά για συστήματα αναγνώρισης φωνής μεγάλου λεξιλογίου αυτό είναι τυπικά αδύνατον. Γι' αυτό, κάθε λέξη διασπάται σε μια ακολουθία βασικών ήχων που αποκαλούνται φωνήματα. Κάθε ένα φώνημα μοντελοποιείται από ένα Κρυφό Μαρκοβιανό Μοντέλο. Τα μοντέλα φωνημάτων με κρυφά Μαρκοβιανά Μοντέλα αποτελούνται από τρεις καταστάσεις εξόδου και έχουν τοπολογία που κινείται απο αριστερά στα δεξιά. Για την αγγλική γλώσσα υπάρχουν 45 φωνήματα για να περιγράφουν όλες οι προφορές των λέξεων. Τα αντίστοιχα κρυφά Μαρκοβιανά Μοντέλα για την ανάγκη του συγκεκριμένου πειράματος εκπαιδεύτηκαν με το Αγγλικό-Βρετανική βάση δεδομένων της WSJ (Robinson, 1995) με 8000 προτάσεις και 92 ομιλητές.

3.2.4 Γλωσσικό μοντέλο

Το γλωσσικό μοντέλο που χρησιμοποιείται από τα συστήματα αναγνώρισης φωνής μεγάλου λεξιλογίου βασίζεται στα μοντέλα N-grams (Stolcke, 2002). Τα N-grams μοντέλα δίνουν μια εκτίμηση της πιθανότητας $\mathrm{P}(W)$, δηλαδή της πιθανότητας να παρατηρηθεί η συγκεκριμένη ακολουθία λέξεων W. Αν υποθέσουμε ότι η πιθανότητα μια δεδομένης λέξης σε μια πρόταση εξαρτάται απο το πεπερασμένο πλήθος λέξεων που προηγούνται αυτής, η πιθανότητα του τμήματος πρότασης N λέξεων μπορεί να αποτυπωθεί με την παρακάτω σχέση:

$$P(W_1^N) = \prod_{k=1}^{N} P(W_k \mid W_1^{k-1}) \tag{3.2}$$

Τα N-grams ταυτόχρονα συνοψίζουν συντακτική πληροφορία επικεντρώνοντας το ενδιαφέρον τους σε τοπικές αλληλο-εξαρτήσεις. Συνεπώς τα N-grams είναι πολύ αποτελεσματικά για γλώσσες που έχουν ισχυρούς κανόνες στην διάταξη των λέξεων και η μία λέξη επηρεάζεται από την γειτονία των λέξεων που τις περιβάλλουν. Το στατιστικό γλωσσικό μοντέλο περιγράφει με πιθανότητες τους περιορισμούς και τις ασυμβατότητες στην σειρά των λέξεων που υποβόσκουν σε κάθε γλώσσα. Έτσι οι τυπικές αλληλουχίες λέξεων παρουσιάζουν υψηλή πιθανότητα σε αντίθεση με την με άλλες αλληλουχίες που έχουν μικρή πιθανότητα. Τα

N- grams έχουν επίσης επιλεγεί, για το λόγο ότι η κατανομή των πιθανοτήτων τους μπορεί να υπολογισθεί απευθείας από ένα σώμα κειμένου, χωρίς να υπάρχει περαιτέρω ανάγκη για συντακτικούς, γραμματικούς και άλλους κανόνες. Το στατιστικό γλωσσικό μοντέλο που χρησιμοποιείται για τους πειραματικούς λόγους αυτής της εργασίας αποτελείται από διγράμματα και τριγράμματα (εκτενής αναφορά στα χαρακτηριστικά του γλωσσικού μοντέλου στην ενότητα 2.8).

3.2.5 Αποκωδικοποίηση Viterbi

Το βασικό και καίριο πρόβλημα της αναγνώρισης φωνής είναι να βρει την διάταξη των λέξεων που μεγιστοποιεί την πιθανότητα της σχέσης 3.1. Τα συστήματα αναγνώρισης φωνής μεγάλου λεξιλογίου είναι τόσο σύνθετα που απαιτούν περιορισμό του χώρου έρευνας. Η μηχανή αναζήτησης που χρησιμοποιείται εδώ υλοποιεί των αλγόριθμο Viterbi. Τα κλαδιά του συνολικού δέντρου (χώρος έρευνας) αποτελούνται από κόμβους που οι συνδέσεις αντιστοιχούν σε συνδέσεις μεταξύ των καταστάσεων των κρυφών Μαρκοβιανών Μοντέλων και οι κόμβοι στο τέλος κάθε λέξης ενώνονται με μεταβάσεις σε άλλες λέξεις. Κάθε μονοπάτι από τον αρχικό κόμβο σε κάθε τυχαίο σημείο του δέντρου χαρακτηρίζεται από ένα στέλεχος που τοποθετείται στον κόμβο που τερματίζει το κάθε μονοπάτι. Η πιθανότητα του κάθε στελέχους είναι η τελική λογαριθμική πιθανότητα ως αυτό το σημείο, και κάθε φορά που η διαδρομή περνάει από ένα κόμβο, αυτομάτως καταχωρείται στο ιστορικό του κάθε στελέχους. Σε κάθε χρονική περίοδο η καλύτερη πιθανότητα κάθε στελέχους σημειώνεται και κάθε στέλεχος που έχει πιθανότητα μικρότερη από την καλύτερη καταστρέφεται.

3.3 Αναγνώριση φωνής συναισθηματικού λόγου

3.3.1 Χαρακτηριστικά του συναισθηματικού λόγου

Όπως είναι γνωστό, ο ανθρώπινος λόγος μεταφέρει 2 ειδών πληροφορίες. Την γλωσσική πληροφορία που έχει να κάνει με την δομή της γλώσσας, δηλαδή τους κανόνες που διέπουν την γλώσσα και την παρα-γλωσσική πληροφορία που σχετίζεται με τα φωνητικά χαρακτηριστικά του λόγου δηλαδή, την ενέργεια , την θεμελιώδη συχνότητα, την ένταση και την διάρκεια των φωνημάτων. Το συναίσθημα μοιάζει να επηρεάζει τόσο την δομή της γλώσσας όσο και τα προσωδιακά χαρακτηριστικά της. Από την βιβλιογραφία φαίνεται ότι το συναίσθημα επηρεάζει την επιλογή των λέξεων, την σειρά με την οποίον τις διατυπώνουμε, την ένταση του ήχου της ομιλίας (άλλοτε ακούγεται δυνατά και άλλοτε σιγά) και την διάρκεια των λέξεων και των

φωνημάτων. Με την παρούσα εργασία θα ασχοληθούμε μόνο με το σκέλος της επίδρασης του συναισθήματος στην γλωσσική πληροφορία.

3.3.2 Αναγνώριση συναισθηματικού λόγου με την χρήση προσωδίας

Τα πειράματα που έχουν γίνει με συστήματα αναγνώρισης φωνής επιβεβαιώνουν ότι το συναίσθημα επηρεάζει κατά πολύ την απόδοση τους. Ο Πίνακας 3.1 συνοψίζει τα αποτελέσματα ενός συστηματικού πειράματος που έγινε με σκοπό τον προσδιορισμό της απόδοσης ενός συστήματος αναγνώρισης φωνής σε συνθήκες θορύβου και ησυχίας. Μία ακόμη κατηγοριοποίηση των αποτελεσμάτων γίνεται λαμβάνοντας υπόψη την συναισθηματική κατάσταση του χρήστη. Οι επιδράσεις του συναισθήματος είναι περισσότερο σημαντικές από τις άλλες καταστάσεις που συνήθως μελετώνται (Steeneken και Hansen, 1999).

Κάποιος θα ανέμενε ότι εκπαιδεύοντας το σύστημα με λόγο που παρέχει πληροφορία για την προσωδία θα μπορούσε να βελτιώσει την απόδοση ενός συστήματος αναγνώρισης φωνής. Ο Πίνακας 3.2 όπως αναφέρουν οι Polzin και Waibel (1998), δείχνει ότι μια τέτοια τεχνική μπορεί να φέρει θετικά αποτελέσματα. Η πρώτη στήλη ανταποκρίνεται στις διαφορετικές συναισθηματικές καταστάσεις ενώ η δεύτερη στήλη δείχνει ότι όταν το σύστημα εκπαιδευτεί τόσο με φωνήματα όσο και με την προσωδία μπορεί να παράγει βελτιωμένα αποτελέσματα.

Άρα λοιπόν είναι προφανές ότι η ενίσχυση του συστήματος με προσωδία είναι μια λύση στο πρόβλημα που λέγεται αναγνώριση του συναισθηματικά φορτισμένου λόγου. Η μελέτη που διεξάχθηκε από τους Polzin και Waibel (1998) περιλάμβανε ηθοποιούς που απέδωσαν με διαφορετικά συναισθήματα με την χρήση προεπιλεγμένων προτάσεων. Σε αυτό το πλαίσιο αναμένει κανείς ότι κάποια στοιχεία του συναισθήματος, ιδιαίτερα, τα αμβλυμμένα προσωδιακά χαρακτηριστικά του θα ενισχυθούν ενώ από την άλλη πλευρά κάποια άλλα θα παραμερισθούν συστηματικά, με σκοπό τον επιτυχή έλεγχο της απόδοσης του συστήματος.

	Συναισθηματική κατάσταση	
Συνθήκες Θορύβου	Ηρεμία	Θυμός
Ησυχία	48%	20%
Θόρυβος	33%	15%

Πίνακας 3.1 Συγκριτικός πίνακας για διαφορετικές καταστάσεις.

Συναισθηματική κατάσταση	Βελτίωση με Συναισθηματικό Μοντέλο
Ευτυχία	15,30%
Φόβος	21,90%
Θυμός	-0,40%
Λύπη	24,50%
Ηρεμία	5,10%

Πίνακας 3.2 Το ποσοστό βελτίωσης των αποτελεσμάτων της αναγνώρισης φωνής χρησιμοποιώντας την πληροφορία της προσωδίας.

3.4 Χαρτογράφηση των συναισθηματικών καταστάσεων

3.4.1 Κατηγορίες συναισθημάτων

Ξεκινώντας την διαδικασία περιγραφής των κυριότερων συναισθημάτων θα ήταν σκόπιμο να προσδιορίσουμε το έννοια της λέξης «συναίσθημα». Έτσι λοιπόν με τον όρο συναίσθημα θεωρούμε μια πολυσύνθετη αντιδραστική τάση ως προς κάποιο ερέθισμα που εκδηλώνεται σε τακτά χρονικά διαστήματα. Η εκδήλωση - εμφάνιση του ερεθίσματος είναι η αφετηρία, εν συνεχεία έχουμε την πρόσληψη (συνειδητή ή ασυνείδητη) και την υποκειμενική ερμηνεία του ατόμου για αυτό και η τελική κατάληξη είναι μια μορφή αντίδρασης σε ποικίλα επίπεδα π.χ. φυσιογνωμικό, εκφραστικό, γνωσιακό, φυσιολογικό-βιολογικό, νευρολογικό, εμπειρικό κ.α. Το συναίσθημα δηλαδή είναι μια μορφή αντίδρασης του οργανισμού μας στις εξελίξεις του περιβάλλοντος μας και όχι μόνο. Αν υπάρχει το ερέθισμα (εσωτερικό ή εξωτερικό) και προσληφθεί από το άτομο άλλα αυτό δεν προχωρήσει σε καμία αντίδραση σε κανένα επίπεδο, μιλάμε για απουσία συναισθήματος. Το συναίσθημα λοιπόν αποτελεί αντίδραση και πρωταρχικά αντίδραση ψυχική προς κάποιο ερέθισμα (Boutri and Stalikas, 2004). .

Στον χώρο της θεωρίας των συναισθημάτων επικρατούν δύο κυρίαρχες τάσεις. Στην πρώτη συγκαταλέγονται αυτοί που υποστηρίζουν ότι τα συναισθήματα είναι ανεξάρτητες διακριτές οντότητες, με ξεχωριστή δομή και λειτουργία, και από στην δεύτερη υπάρχουν αυτοί που αντιλαμβάνονται τα συναισθήματα ως προβολές επάνω σε δύο διπολικούς άξονες.

3.4.2 Βασικά συναισθήματα

Η κεντρική ιδέα αυτών των θεωριών είναι ότι υπάρχουν κάποια συναισθήματα, τα οποία είναι διακριτά, βασικά και παγκόσμια. Το κάθε ένα από αυτά τα συναισθήματα θεωρείται ότι έχει μοναδικά πρότυπα φυσιολογικής διέγερσης, μοναδικές συμπεριφορικές εκφράσεις, μοναδικό τρόπο διοργάνωσης των γνώσεων και της αντίληψης, μοναδικό τρόπο κινητοποίησης του οργανισμού.

Έτσι για κάποια συναισθήματα τα βασικά χαρακτηριστικά τους παρατηρούνται κατ' επανάληψη σε όλους τους ανθρώπινους πολιτισμούς, καθώς και σε κάποια ανώτερα θηλαστικά. Τα συναισθήματα αυτά φαίνεται ότι συσχετίζονται με ιδιαίτερες εκφράσεις του προσώπου, αναγνωρίσιμες παγκοσμίως, ενώ έχουν εντοπιστεί και οι μοναδικές βιολογικές λειτουργίες που εξυπηρετούν ως προς την επιβίωση του ατόμου και του είδους. Αντίθετα υπάρχουν άλλα συναισθήματα τα οποία φαίνεται πως προσδιορίζονται από κοινωνικούς και πολιτισμικούς παράγοντες και τα οποία διαφέρουν ανά είδος. Αναλογιζόμενοι αυτές τις παρατηρήσεις, η επόμενη λογική κίνηση είναι η αναζήτηση νευροφυσιολογικών και ανατομικών δομών που συνδέονται με το κάθε ένα από τα βασικά συναισθήματα.

Πρώτος ο Δαρβίνος (1872/1965) περιέγραψε αναλυτικά περισσότερα από δώδεκα συναισθήματα, τόσο θετικά όσο και αρνητικά, υποστηρίζοντας ότι πολλά από αυτά αναπτύχθηκαν από κεντρικά λειτουργικά συστήματα. Για τον Δαρβίνο, η έκφραση των βασικών συναισθημάτων εξυπηρετούσε προσαρμοστικούς και εξελικτικούς σκοπούς που αποσκοπούσαν στην επιβίωση του είδους. Για κάθε ένα, λοιπόν, από τα συναισθήματα που περιγράφει, αναφέρει τους φυσιολογικούς μηχανισμούς διέγερσης, τις ιδιαίτερες συσπάσεις των μυών του προσώπου και τις κινήσεις των μελών τους σώματος καθώς και το σκοπό που εξυπηρετεί η έκφρασή τους. Για παράδειγμα, το συναίσθημα του φόβου έχει ως χαρακτηριστικές εκφράσεις προσώπου το ανασήκωμα των βλεφάρων, το άνοιγμα του στόματος και το γούρλωμα των ματιών. Με αυτόν τον τρόπο το άτομο παρακολουθεί καλύτερα το περιβάλλον του, οπτικά και ακουστικά, και ανιχνεύει για περαιτέρω πληροφορίες που σχετίζονται με το φοβογόνο ερέθισμα. Επιπλέον, ο οργανισμός προετοιμάζεται να αντιδράσει σε αυτό το φοβογόνο ερέθισμα είτε με επίθεση είτε με φυγή. Η αντίδραση αυτή έχει ξεκάθαρο προσαρμοστικό χαρακτήρα, καθώς βοήθησε το ανθρώπινο είδος στον εντοπισμό και την αποφυγή των απειλών του περιβάλλοντος, και κατά συνέπεια στην επιβίωση και εξέλιξη του ανθρώπινου είδους. Παρ' όλο που οι συνθήκες διαβίωσης έχουν αλλάξει δραματικά από τον καιρό της δημιουργίας του ανθρώπινου είδους, φαίνεται ότι τα συναισθήματα διατηρούν ακόμη την προσαρμοστική τους αξία. Ο Δαρβίνος διέκρινε 2 βασικές προσαρμοστικές λειτουργίες της συναισθηματικής έκφρασης: την επικοινωνία με το κοινωνικό περιβάλλον και τη ρύθμιση του

συναισθήματος. Η πρώτη λειτουργία αφορά την έκφραση του συναισθήματος για τη σωστή κοινωνικοποίηση και ανατροφή των παιδιών (για παράδειγμα, το μητρικό χαμόγελο επιβράβευσης ή η έκφραση απογοήτευσης). Η δεύτερη λειτουργία αφορά την εκφόρτιση του συναισθήματος μέσα από την ελεύθερη και πλήρη συναισθηματική έκφραση (Boutri and Stalikas, 2004).

Για τον Williams James το σώμα παίζει βασικό ρόλο για το συναίσθημα. Οι κινήσεις του σώματος ακολουθούν κάποια ερεθίσματα αυτόματα και το συναίσθημα εκδηλώνεται μέσω την αντίληψης αυτών των αλλαγών. Συνεπώς χωρίς την αντίληψη από το σώμα δεν θα υπήρχε συναίσθημα. Τέλος βάσει της θεωρίας του James η έκφραση του προσώπου ενός ατόμου έχει επίδραση στην υποκειμενική συναισθηματική του εμπειρία, για παράδειγμα, εάν ένα άτομο έχει χαρούμενο πρόσωπο τότε το άτομο αυτό συμπεριφέρεται περισσότερο χαρούμενα.

Η κεντρική ιδέα της Γνωστικής θεωρίας είναι ότι η σκέψη και το συναίσθημα είναι αλληλο-εξαρτώμενα. Περισσότερο συγκεκριμένα όλα τα συναισθήματα υπό το πρίσμα αυτής της θεωρίας συνδέονται με την αξιολόγηση, μια διαδικασία κρίσης ενός ερεθίσματος ως αρνητικού ή θετικού για το κάθε άτομο. Αυτή η παράμετρος καθορίζει πόσο σημαντικό είναι ένα ερέθισμα για κάθε άτομο και αυξάνει την πιθανότητα της αντίδραση του.

Στην Κοινωνική θεωρία τα συναισθήματα λαμβάνονται σαν πρότυπα κοινωνικά παραγόμενα τα οποία μαθαίνονται και διαχέονται στους πολιτισμούς. Τα συναισθήματα παίζουν έναν πολύ σημαντικό ρόλο αφού είναι αυτά που ρυθμίζουν την επικοινωνία μεταξύ των ανθρώπων. Οι εκφράσεις των συναισθημάτων και τα συναισθήματα από μόνα τους περιγράφονται σαν παράγοντας πολιτισμού. Αν και η φύση των συναισθημάτων είναι βιολογική, αυτό που έχει μεγάλο ενδιαφέρον είναι οι μηχανισμοί διάπλασης των συναισθημάτων από τις κοινωνίες.

Από μια πρώτη εκτίμηση όλες αυτές οι θεωρίες μοιάζουν να είναι αλληλο-συγκρουόμενες. Κοιτώντας όμως πιο προσεκτικά κάποιος μπορεί να υποστηρίξει ότι οι διαφορετικές θεωρίες δεν είναι τίποτα άλλο από θεωρίες που εξετάζουν διαφορετικές πλευρές του συναισθήματος. Η Darwinian θεωρία αποτελεί το εξελικτικό περιβάλλον του συναισθήματος, η Jamesian θεωρία εξετάζει την σχέση σώματος και συναισθήματος. Η Γνωστική θεωρία μελετά τα φυσιολογικά φαινόμενα που σχετίζονται με τα συναισθήματα, ενώ η κοινωνική θεωρία προβάλλει το συναίσθημα σαν κοινωνική λειτουργία σε ευρύτερο περιβάλλον.

3.4.3 Στοιχεία επιβεβαίωσης των διακριτών συναισθημάτων

Βάσει της θεωρίας των διακριτών συναισθημάτων υπάρχουν τουλάχιστον 6 συναισθημάτων: χαρά, έκπληξη, φόβος, λύπη, θυμός, αηδία μαζί με περιφρόνηση, που αντανακλώνται στις εκφράσεις του προσώπου.

Υπάρχουν ερευνητικά δεδομένα τα οποία στηρίζουν την άποψη ότι κάποια συναισθήματα βιώνονται τόσο από τον άνθρωπο όσο και από τα ανώτερα θηλαστικά (Νευροφυσιολογικά δεδομένα). Τα συναισθήματα αυτά φαίνεται ότι ενεργοποιούνται απευθείας από υποφλοιώδεις εγκεφαλικές δομές -κοινές στους ανθρώπους και κάποια ζώα- χωρίς τη διαμεσολάβηση γνωστικών μηχανισμών και μάθησης. Επιπλέον, ο Ekman (1992), παραθέτει δεδομένα σύμφωνα με τα οποία υπάρχουν φυσιολογικές αποδείξεις ότι το αυτόνομο νευρικό σύστημα ακολουθεί ένα διαφορετικό πρότυπο ενεργοποίησης για διαφορετικά συναισθήματα όπως του φόβου, του θυμού, της αηδίας, και της λύπης.

Στα πρώτα στάδια της παιδικής ηλικίας η ενεργοποίηση του συναισθήματος οδηγεί κατά κανόνα σε άμεση έκφραση του. Καθώς το παιδί αποκτάει έλεγχο στους σωματικούς μυς που σχετίζονται με τη συναισθηματική έκφραση, και καθώς κοινωνικοποιείται και ενημερώνεται για τους κοινωνικούς κανόνες συμπεριφοράς, σταδιακά μαθαίνει να ρυθμίζει και να διαφοροποιεί την έκφραση των συναισθημάτων του (Συναισθηματική Εξέλιξη). Μορφοποιώντας την συναισθηματική έκφραση με τον καιρό μορφοποιείται και το κάθ' αυτό συναισθηματικό βίωμα. Έτσι διαμορφώνονται οι «συναισθηματικές-γνωστικές δομές», δηλαδή, ένας σύνδεσμος μεταξύ συναισθηματικού βιώματος και γνώσεων. Αυτές οι δομές θεωρούνται οι θεμέλιοι λίθοι της σκέψης και της μνήμης. Αντίστοιχα, η μη-προσαρμοστική συμπεριφορά αντανακλά προβλήματα στο σύνδεσμο μεταξύ συναισθήματος και γνώσεων, και στα πρότυπα συναισθήματος- γνώσεων- δράσης (Boutri and Stalikas, 2004).

3.4.4 Αναπαράσταση συναισθημάτων πάνω σε διπολικούς άξονες

Τα συναισθήματα μπορούν να γίνουν αντιληπτά ως σημεία επάνω σε δύο διπολικούς άξονες. Κατά συνέπεια, τα συναισθήματα δεν είναι εντελώς ανεξάρτητα μεταξύ τους, ούτε λειτουργούν ως αυτοτελείς οντότητες, αλλά είναι αλληλένδετα.

Τα βασικά επιχειρήματα των ερευνητών αυτής της άποψης προέρχονται από τον χώρο της σημασιολογίας. Στην καθημερινή γλώσσα η άνθρωποι μιλούν με τους όρους θετικό - αρνητικό, ή ευχάριστο - δυσάρεστο. Επιπλέον, η άποψη αυτή έχει υποστηριχθεί από ερευνητικά δεδομένα σχετικά με τις γνωστικές διεργασίες (θεωρίες αποτίμησης/ appraisal theories) που οδηγούν στο βίωμα κάποιου συναισθήματος, την ένταση του συναισθήματος, την αντίληψη του

συναισθήματος στηριζόμενοι σε φωνητικά στοιχεία ή στην έκφραση του προσώπου. Παράδειγμα τέτοιων ερευνών είναι το εύρημα ότι δείχνοντας σε κάποιον συμμετέχοντα δύο φωτογραφίες προσώπων όπου το ένα έχει θλιμμένη έκφραση και το άλλο ουδέτερη, ο συμμετέχων θα χαρακτηρίσει την ουδέτερη έκφραση ως χαρούμενη, δείχνοντας έτσι την διπολικότητα των δύο συναισθημάτων.

Οι ρίζες της συγκεκριμένης θεώρησης βρίσκονται στον Spencer (1880), ο οποίος αντιλήφθηκε τα συναισθήματα ως διαστάσεις της συνείδησης. Σύντομα, ο Wundt (1897), προέκτεινε την άποψη του Spencer, υποστηρίζοντας ότι όλα τα συναισθήματα μπορούν να γίνουν κατανοητά εάν τοποθετηθούν σε δυο άξονες: ευχαρίστηση-δυσαρέσκεια (pleasantness-unpleasantness), χαλάρωση-ένταση (relaxation-tension).

Από τότε πολλοί θεωρητικοί και ερευνητές στον χώρο του συναισθήματος έχουν υποστηρίξει την ύπαρξη των συναισθηματικών διαστάσεων. Οι βασικές διαστάσεις στις οποίες κινούνται οι ερευνητές είναι το θετικό-αρνητικό συναίσθημα, και η ένταση του συναισθήματος. Ίσως ο κυριότερος σύγχρονος εκπρόσωπος είναι ο Russell, του οποίου η θεωρία γίνεται αποδεκτή από την πλειοψηφία των θεωρητικών αυτού του ρεύματος. Η θεωρία του είναι η *Θεωρία Ευχαρίστησης - Διέγερσης (Pleasure- Arousal Theory -PAT)*. Σύμφωνα με το Russell, τα ερευνητικά δεδομένα του υποδεικνύουν την ύπαρξη τουλάχιστον 2 βασικών διπολικών αξόνων: ευχαρίστησης έναντι δυσαρέσκειας (pleasantness vs unpleasantness), και ενεργοποίησης έναντι απενεργοποίησης (activation vs deactivation). Για τους Barret & Russell (1998), όλα τα συναισθήματα μπορούν να τοποθετηθούν σε έναν χώρο που αποτελείται από αυτές τις διαστάσεις. Σύμφωνα με την διάταξη αυτή, όλα τα συναισθήματα που παρουσιάζονται έχουν και το αντίθετο τους (180^{o} απόκλιση). Επιπλέον, για κάθε συναίσθημα που απεικονίζεται υπάρχει ένα άλλο με το οποίο είναι ανεξάρτητο και βρίσκεται σε απόκλιση 90^{o}. (Boutri and Stalikas, 2004).

3.4.5 Τροχός συναισθημάτων Whissel

Η Whissel (1989) προσέγγισε το θέμα της αναπαράστασης του συναισθήματος με το επίπεδο Δραστηριοποίησης και Αποτίμησης (Activation-evaluation space), που είναι ανάλογο της θεωρίας *Ευχαρίστησης - Διέγερσης*. Το επίπεδο Δραστηριοποίησης και Αποτίμησης (Activation-evaluation space) αναπαριστά τις συναισθηματικές καταστάσεις σε 2 διαστάσεις.

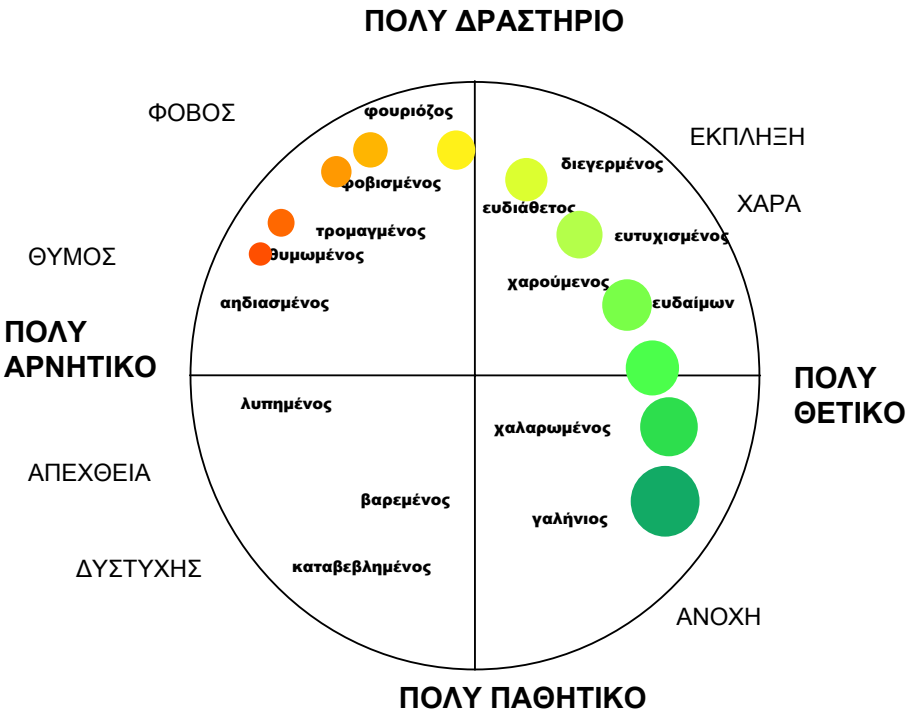
- Ο όρος «Δραστηριοποίηση» (Activation) αναπαριστά πόσο δυναμικό έχει η συναισθηματική κατάσταση του ομιλητή. Η ευτυχία έχει υψηλό δείκτη

δραστηριοποίησης (Activation) ενώ στον αντίποδα η ανία έχει πολύ χαμηλό δείκτη.

- Ο όρος «Αποτίμηση» (Evaluation) είναι πολύ σφαιρικός και αναπαριστά το κατά πόσον είναι θετικό η αρνητικό το αίσθημα που απορρέει από μια συναισθηματική κατάσταση.

Σε αυτό το επίπεδο υπάρχουν δυο κύριοι άξονες, ο άξονας της δραστηριοποίησης που ξεκινάει από το πολύ ενεργό και καταλήγει στο πολύ παθητικό, και ο άξονας της αποτίμησης που ξεκινάει από το πολύ θετικό και καταλήγει στο πολύ αρνητικό. Το κέντρο του κύκλου δείχνει κατά κάποιον τρόπο την συναισθηματική ουδετερότητα. Το επίπεδο περιλαμβάνει και άλλα χαρακτηριστικά όπως το χρώμα σε κάθε θέση βάσει της χρωματικής κωδικοποίησης που εισήγαγε ο Plutchik (1997). Όταν ο δείκτης βρίσκεται σε μια θέση που αναφέρεται σε θετικά συναισθήματα χρωματίζεται πράσινος. Κόκκινος είναι σε περιοχές όπου υπάρχουν αρνητικά συναισθήματα και κίτρινο σε θέσεις όπου τα συναισθήματα είναι ενεργά και μπλε εκεί που είναι παθητικά.

Για παράδειγμα, η χαρά έχει πολύ θετική αποτίμηση ενώ η απόγνωση παρουσιάζει πολύ αρνητική αποτίμηση. Διάφορες τεχνικές καταλήγουν στο συμπέρασμα ότι ο κάθε συναισθηματικός όρος-λέξη έχει μια θέση πάνω στον συναισθηματικό χώρο και κάθε σημείο περιγράφεται από 2 συνιστώσες που είναι οι 2 διαστάσεις. Θεωρητικά, το επίπεδο είναι κυκλικό και η περίμετρος του ορίζεται από τις ακραίες συναισθηματικές καταστάσεις. Οι καταστάσεις αυτές περιγράφονται από σημεία που είναι ισαπέχοντα από το ουδέτερα συναισθηματικό σημείο (αρχή των αξόνων) (Σχήμα 3.2).



Σχήμα 3.2 Η αναπαράσταση του επιπέδου activation-evaluation.

3.4.6 Συλλογή δεδομένων συναισθηματικού λόγου

Το μεγαλύτερο ποσοστό των βάσεων δεδομένων που έχουν χρησιμοποιηθεί γύρω από την έρευνα της ανίχνευσης της συναισθηματικής κατάστασης ενός ομιλητή έχουν τρία είδη χαρακτηριστικών.

Στις περισσότερες από αυτές, το συναίσθημα προσομοιώνεται από έναν ηθοποιό (όχι απαραίτητα εκπαιδευμένο), ο ηθοποιός διαβάζει προκαθορισμένα κείμενα, και έχει σαν στόχο να προσομοιώσει έντονα συναισθήματα (full-blown). Πάρα πολλές φορές γίνεται προσπάθεια να γίνουν τα δεδομένα πιο φυσικά, κάνοντας τα πιο συναφή με το συναίσθημα, δηλαδή χρησιμοποιώντας υλικό του οποίου το περιεχόμενο του είναι συναισθηματικά φορτισμένο.
Αυτού του είδους το υλικό έχει αρκετό ενδιαφέρον, πρώτον γιατί είναι εύκολο να συλλεγεί, και επίσης μπορεί να χρησιμοποιηθεί σε συγκεκριμένες μελέτες. Όμως δεν πρέπει να ξεχνάμε ότι ένα τέτοιο υλικό δεν περιλαμβάνει φυσική ομιλία και δεν είναι αντιπροσωπευτικό του καθημερινού συναισθηματικού λόγου.

Ακόμη υπάρχουν άλλες βάσεις όπου έχει γίνει καταγραφή τηλεοπτικών και ραδιοφωνικών εκπομπών με τυχαία δείγματα λόγου που περιλαμβάνουν μεγάλο φάσμα συναισθηματικών καταστάσεων. Μερικά τέτοια παραδείγματα αποτελούν η βάση των Roach et

al., (1998) που χρησιμοποίησαν τμήματα από ραδιοφωνικές εκπομπές με έντονα συναισθηματικό λόγο και η βάση δεδομένων του Belfast που είναι οπτικό-ακουστική και σε αυτήν χρησιμοποιήθηκαν διάλογοι με έντονα συναισθηματικό λόγο κυρίως καταγεγραμμένοι από την τηλεόραση. Και οι δυο αυτές βάσεις υστερούν στο γεγονός ότι είναι δύσκολα διαθέσιμες λόγω της προστασίας των προσωπικών δεδομένων αυτών που μιλούσαν στην τηλεόραση ή στο ραδιόφωνο. Επίσης ένα άλλο σημαντικό πρόβλημα που παρουσιάζεται είναι η ποιότητα του καταγεγραμμένου ηχητικού σήματος.

Παρόμοιες συλλογές δεδομένων, με μικρότερο όμως φάσμα συναισθημάτων περιλαμβάνουν υλικό που καταγράφεται σε συγκεκριμένες στιγμές όπως η ενασχόληση με ηλεκτρονικά παιχνίδια, η χρήση προσομοιωτών πτήσεων, και ρεπορτάζ δημοσιογράφων κάτω από την επίδραση διαφόρων γεγονότων.

Κάποιοι ερευνητές ισχυρίζονται ότι η δημιουργία βάσεων δεδομένων με πραγματικούς διαλόγους ανθρώπου-μηχανής, αποτελεί ουσιαστικά την καλύτερη και αποτελεσματικότερη λύση. Υπάρχουν κάποια στοιχεία που κάνουν αυτού του είδους των βάσεων πολύ ελκυστικές. Ο πρώτος λόγος είναι ότι τα συναισθήματα είναι πραγματικά και όχι προσποιητά. Δεύτερον είναι προϊόν διαλόγου και όχι μονολόγου που δημιουργεί τεχνητά συναισθήματα. Τρίτον, είναι πολύ κοντά στα δεδομένα που επεξεργάζονται οι εφαρμογές που έχουν στόχο την αναγνώριση συναισθήματος. Αλλά όπως είναι φυσικό μαζί με τα πλεονεκτήματα έπονται και οι περιορισμοί. Η συχνότητα εμφάνισης έντονων συναισθημάτων είναι χαμηλή με πιο συχνό συναίσθημα ήταν η απογοήτευση. Επιπλέον λόγω της φύσης της διασύνδεσης δημιουργούνται κάποιοι περιορισμοί στις προτάσεις και πιθανόν στον τρόπο με τον οποίο εκφράζεται το συναίσθημα μέσα από αυτές.

Ανάλογες δράσεις σαν και αυτές που αναφέραμε περιλαμβάνουν την μελέτη των Ang et al (2002) όπου γίνεται χρήση του DARPA Communicator Corpus. Με το σύστημα αυτό οι χρήστες καλούν ένα τηλεφωνικό κέντρο για να κάνουν ταξιδιωτικές κρατήσεις. Οι Lee and Narayanan (2003) επεδίωξαν την ανίχνευση αρνητικών και μη αρνητικών συναισθημάτων χρησιμοποιώντας εκφράσεις που έχουν καταγραφεί μέσα από διαλόγους ανθρώπου μηχανής. Η πλειοψηφία των διαλόγων αποτελείται από μια πρόταση ανά έκφραση. Οι Boozer et al. (2003) εργάστηκαν στην κατεύθυνση αναγνώρισης ουδέτερων, χαρούμενων και λυπημένων καταστάσεων με την χρήση διαλογών ανθρώπου μηχανής που έχουν δημιουργηθεί από το Mercury- ένα σύστημα τηλεφωνικής επικοινωνίας για κρατήσεις αεροπορικών θέσεων. Ένας εναλλακτικός τρόπος για την δημιουργία μιας πραγματικά φυσικής βάσης δεδομένων με μεγάλο φάσμα συναισθηματικών καταστάσεων, είναι οι πολύωρες ηχογραφήσεις. Ο Campbell έχει περιγράψει τον τρόπο που πρέπει να ακολουθηθεί για να γίνει μια τέτοια ηχογράφηση (Douglas-Cowie et al 2003a). Η εργασία αυτή είναι σε εξέλιξη.

Τέλος έχουν αναπτυχθεί κάποιες τεχνικές με τις οποίες διεγείρονται τα συναισθήματα των ανθρώπων με αποτέλεσμα οι παραγόμενες καταστάσεις να περικλείουν συναισθήματα που εκφράζονται πιθανώς και με τον λόγο. Πρώιμα παραδείγματα τέτοιων τεχνικών περιλαμβάνουν ένα πείραμα όπου παρουσιάζονται σε διάφορα άτομα κάποιες δυσάρεστες εικόνες (Tollkmitt & Scherer, 1986) και κάποια μέρη από την βάση δεδομένων SUSAS (Speech under Simulation and Actual Stress database), όπου χρησιμοποιείται λόγος παρακινούμενος από στρεσιογόνες καταστάσεις, (Hansen & Bou-Ghazale, 1997). Πιο σύγχρονα πειράματα περιλαμβάνουν προσομοιώσεις τηλεφωνικών κέντρων που είναι σχεδιασμένα να παράγουν εκνευρισμό στους χρήστες (Batliner et al., 2003; Mitchell et al., 2000); έντονο στρες πάνω στην οδήγηση (Fernandez & Picard, 2003); και άλλα πιθανά θέματα που παράγουν ανία και πλήξη στους χρήστες (Cowie et al, 2003).

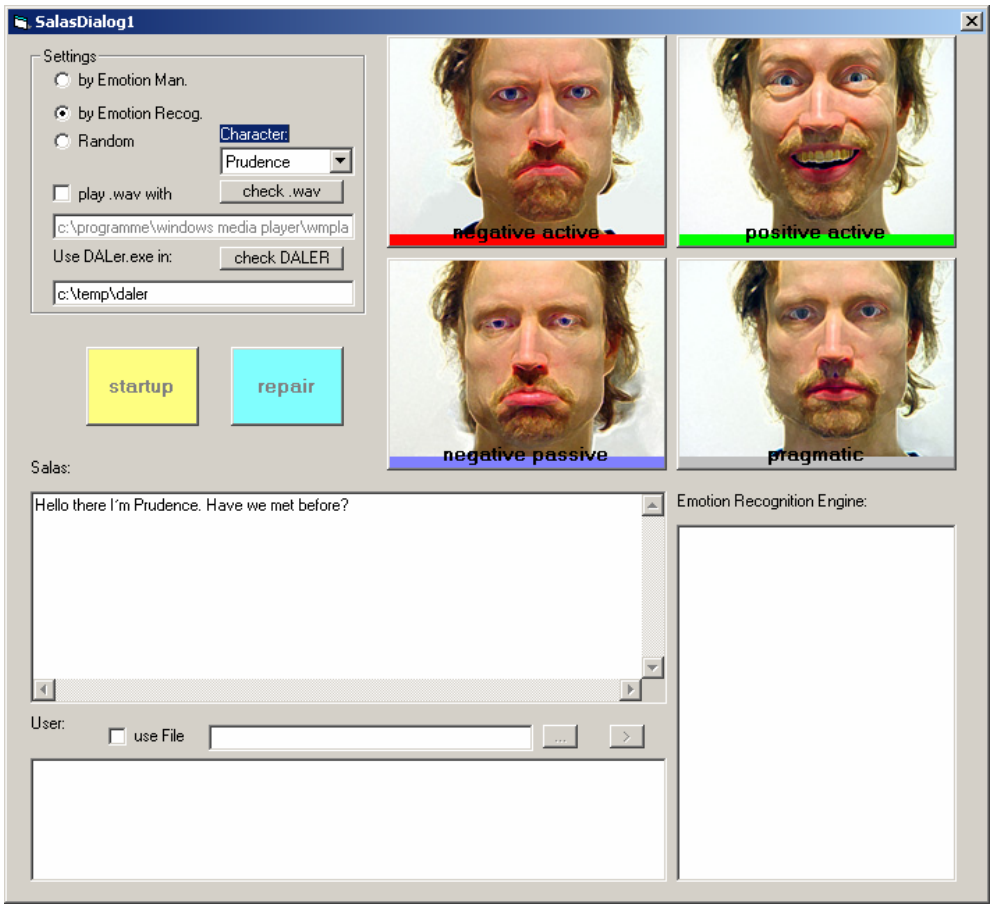
Τα περισσότερα από αυτά τα πειράματα δίνουν καλύτερο έλεγχο και υψηλότερο ρυθμό δεδομένων από ότι τα πειράματα με τα τηλεφωνικά κέντρα, αλλά πάλι τείνουν να παράγουν αδύναμα αρνητικά συναισθήματα, και συνήθως βάζουν περιορισμούς στο γλωσσικό περιεχόμενο του λόγου. Παρόλα αυτά, έγιναν προσπάθειες να αναπτυχθούν ανάλογες βάσεις σε ευρύτερη κλίμακα όπως αυτή των Kwon et al (2003), που χρησιμοποίησαν ηχογραφήσεις από την βάση δεδομένων ψυχαγωγίας AIBO στα γερμανικά. Στο πλαίσιο αυτής της εργασίας έχει γίνει χρήση της βάσης δεδομένων με συναισθηματικό λόγο που ονομάζεται SALAS.

Η βάση SALAS βασίζεται στην ανάπτυξη της ιδέας ELIZA του Weizenbaum (1966). Ο χρήστης επικοινωνεί με το σύστημα πρόκλησης συναισθήματος. Είναι γεγονός ότι το σύστημα δεν λαμβάνει υπόψη το νόημα των λεγομένων του χρήστη. Απλώς ανασύρει μια απάντηση βάσει των επιφανειακών ερεθισμάτων που δίνουν τα λεγόμενα του χρήστη. Στην περίπτωση του SALAS, τα επιφανειακά ερεθίσματα περιλαμβάνουν συναισθηματικό τόνο. Σε παραθυρικό περιβάλλον, ένας διαχειριστής αναγνωρίζει τον συναισθηματικό τόνο, και το χρησιμοποιεί για να διαλέξει την περιοχή της βάσης που θα δώσει την κατάλληλη απάντηση του συστήματος. Αρχικά ο χρήστης επιλέγει ένα από τους τέσσερις τεχνητούς ακροατές (χαρακτήρες) ώστε να συνομιλήσει. Κάθε ένας από αυτούς τους χαρακτήρες οδηγεί με τις ερωτήσεις τον χρήστη προς μια συγκεκριμένη συναισθηματική κατάσταση:

- Ο **«Οξύθυμος»** θα προσπαθήσει να στρέψει τον χρήστη προς τον θυμό.
- Ο **«Ευχάριστος»** θα προσπαθήσει να τον κάνει ευδιάθετο, χαρούμενο.
- Ο **«Μελαγχολικός»** προσπαθεί να τον στρέψει προς την μελαγχολία, σε μια στάση παθητική, απαισιόδοξη.

- Ο **«Προσγειωμένος»** που είναι ρεαλιστής κινούμενος σε λογικά πλαίσια και προσπαθεί να τον προσγειώσει.

Το σύστημα SAL (Σχήμα 3.3) παρέχει ένα πλαίσιο στο οποίο οι χρήστες μπορούν να εκφράσουν μεγάλο φάσμα συναισθημάτων με τρόπο που να μην τους εμποδίζει να μιλάνε ελεύθερα. Η επιτυχία του συστήματος έγκειται στην συνεργασία που θα επιδείξει ο χρήστης. Το σύστημα αποτελεί ένα τρόπο εξάσκησης της έκφρασης συναισθημάτων του χρήστη με συνέπεια να μπορεί να συλλέγει ευρεία κλίμακα συναισθημάτων κατά την διάρκεια της συνομιλίας.



Σχήμα 3.3 Η διαπροσωπεία του SAL

Σε όλες τις βάσεις ανάλογων δεδομένων το πιο σημαντικό στοιχείο είναι το πώς θα γίνουν οι ηχογραφήσεις. Σε αυτό το σημείο πρέπει να τονιστεί ότι υπάρχει μια λεπτή ισορροπία μεταξύ του τρόπου καταγραφής των δεδομένων και των μεθόδων ανάλυσης τους. Οι ηχογραφήσεις πρέπει να γίνουν κατά τέτοιο τρόπο ώστε η ποιότητα των οπτικό ακουστικών δεδομένων να είναι ικανοποιητική και να επιτρέπουν στο χρήστη να συμπεριφερθεί φυσιολογικά αναπτύσσοντας μια ευρεία γκάμα συναισθημάτων. Συνήθως οι άνθρωποι δεν

έχουν την διάθεση να συμπεριφερθούν φυσιολογικά απέναντι σε διαφορετικά ερεθίσματα όταν γίνεται προσπάθεια ηχογράφησης, με συνέπεια να περιορίζουν τις εκφράσεις τους. Αυτό είναι αποτέλεσμα της αντίληψης που έχει ο καθένας ότι παρατηρείται και έτσι δεν συμπεριφέρεται φυσιολογικά. Από την άλλη αν ο χρήστης του συστήματος αφεθεί ελεύθερος να συμπεριφερθεί όπως εκείνος το επιθυμεί οι μέθοδοι ανάλυσης των δεδομένων δεν αποδίδουν αφού δεν τηρούνται οι βασικές ρυθμίσεις για ποιοτικές ηχογραφήσεις.

3.5 Ο αλγόριθμος ενίσχυσης του γλωσσικού μοντέλου με συναισθηματικά χαρακτηριστικά

3.5.1 DAL: Dictionary of Affect language, Λεξικό με συναισθηματικούς όρους

Η ερώτηση που τίθεται είναι κατά πόσον είναι δυνατόν να περιγράψουμε με ένα μέγεθος την έννοια του συναισθήματος που φέρει μια οποιαδήποτε λέξη. Για να μπορεί ένα μέγεθος να περιγράψει ικανοποιητικά την έννοια του συναισθήματος πρέπει να συνυπολογισθεί ότι η έννοια του συναισθήματος δεν έχει ορισθεί καταλλήλως, καθένα μέγεθος πρέπει να συνοδεύεται με βαθμό αξιοπιστίας και εγκυρότητας, το μέγεθος πρέπει να περιγράφεται με όρους που είναι χρονικά εξαρτώμενοι, δεν είναι ένα μέγεθος μοναδικό και αποκλειστικό για την έννοια του συναισθήματος. Ένα τέτοιο μέγεθος που καλύπτει όλες αυτές τις συνθήκες είναι το λεξιλόγιο με το συναισθηματικό περιεχόμενο, Dictionary of Affect language (DAL) (Whissel et al., 1989).

Το λεξιλόγιο με το συναισθηματικό περιεχόμενο (DAL) που κατασκευάστηκε από την Whissell και άλλους συνεργάτες της και αποτελείται από 8742 λέξεις που έχουν αξιολογηθεί πάνω στις 2 διαστάσεις, της «Δραστηριοποίησης» και «Αποτίμησης», που ανταποκρίνονται στην έρευνα του Russell. Κάθε λέξη έχει μια τιμή πάνω σε κάθε διάσταση που απορρέει από την υποκειμενική βαθμολόγηση διαφόρων ερωτηθέντων (βαθμίδα αξιολόγησης, Πίνακας 3.3). Όλα τα άτομα που χρησιμοποιήθηκαν στην διαδικασία συγκέντρωσης αυτού του υλικού δεν αξιολόγησαν την ίδια λίστα λέξεων αλλά κάθε λέξη βαθμολογήθηκε από τουλάχιστον 4 άτομα. Ο μέσος όρος στην διάσταση δραστηριοποίησης-διέγερσης ήταν 1,67 με σταθερή απόκλιση ίση με 0,36, και με κλίμακα από το 1 (μη ενεργός) στο 3 (ενεργός), ενώ ο μέσος όρος αποτίμησης- ευχαρίστησης ήταν 1,85 με σταθερή απόκλιση 0,36 , και με κλίμακα από το 1 (μη χαρούμενος) στο 3 (χαρούμενος).

Οι λέξεις οι οποίες περιλαμβάνονται στο λεξιλόγιο αυτό επελέγησαν με αντικειμενικό τρόπο ακολουθώντας μια διαδικασία τριών βημάτων.

1. Το σώμα κείμενων των Kucera και Francis (1967) προέκυψε από την συλλογή κειμένων από διαφορετικά έντυπα μέσα, στις αρχές της δεκαετίας ’60. Όσες λέξεις από αυτό το σώμα κειμένου είχαν συχνότητα εμφάνισης μεγαλύτερη από 10 και εμφανιζόταν τουλάχιστον σε ένα από τα κείμενα συμπεριλήφθηκαν στην λίστα των λέξεων DAL. Με την μέθοδο αυτή διασφαλίστηκε ότι οι λέξεις που χρησιμοποιηθήκαν δεν ήταν σπάνιες. Επιπλέον τα κύρια ονόματα αφαιρέθηκαν από τα διάφορα κείμενα ύστερα από κατάλληλη επεξεργασία τους.

2. Το σύνολο των λέξεων συγκρίθηκε με 4 δείγματα κειμένων που δημιουργήθηκαν από διάφορα άτομα και δεν ήταν προϊόν συλλογής από έντυπα μέσα. Επίσης το σύνολο των λέξεων αυτών συγκρίθηκε και με ένα μεγάλο δείγμα από κείμενα που ανήκουν στην νεανική λογοτεχνία. Μοναδικές λέξεις που βρέθηκαν από αυτές τις πηγές προστέθηκαν στην λίστα.
    - Ιστορίες φοιτητών, 16,309 λέξεις
    - Συνεντεύξεις με θέμα την παρενόχληση και την κακομεταχείριση, 6,085 λέξεις
    - Περιγραφή συναισθημάτων εφήβων, 15,929 λέξεις
    - Γραπτές εκθέσεις φοιτητών, 14,807 λέξεις
    - Νεανικές ιστορίες της δεκαετίας ’50, ’60 ,’70, ’80, ’90.

3. Η λίστα με τις λέξεις του DAL στο τέλος του δεύτερου σταδίου ελέγχθηκε πάνω σε 16 νέα τυχαία επιλεγμένα δείγματα. Επιπλέον, ελέγχθηκε σε ένα Αγγλικικό σώμα κειμένου 350,000 λέξεων που ήταν προϊόν συλλογής της ίδιας της Whissel από διαφορετικές πηγές. Το 90% των λέξεων του λεξικού DAL περιλαμβάνονταν σε αυτά τα κείμενα.

ΔΡΑΣΤΗΡΙΟΠΟΙΗΣΗ	(1) παθητικό	(2) ενδιάμεσο	(3) ενεργητικό
ΑΠΟΤΙΜΗΣΗ	(1) μη χαρούμενο	(2) ενδιάμεσο	(3) χαρούμενο

Πίνακας 3.3 Η βαθμίδα αξιολόγησης των λέξεων με κλίμακα από 1-3.

Περίπου 200 εθελοντές χρησιμοποιήθηκαν για να εξετάσουν το σύνολο λέξεων. Οι περισσότεροι εθελοντές είχαν την δυνατότητα να κάνουν περί των 200 αξιολογήσεων πριν εμφανίσουν σημάδια κόπωσης και πλήξης. Τα δεδομένα που χρησιμοποιήθηκαν για την

δημιουργία του DAL περιλαμβάνουν περισσότερες από 186,000 διαφορετικές αξιολογήσεις για το σύνολο των λέξεων. Κάθε λέξη αξιολογήθηκε για τις δυο διαστάσεις κατά μέσο όρο 8 φορές.

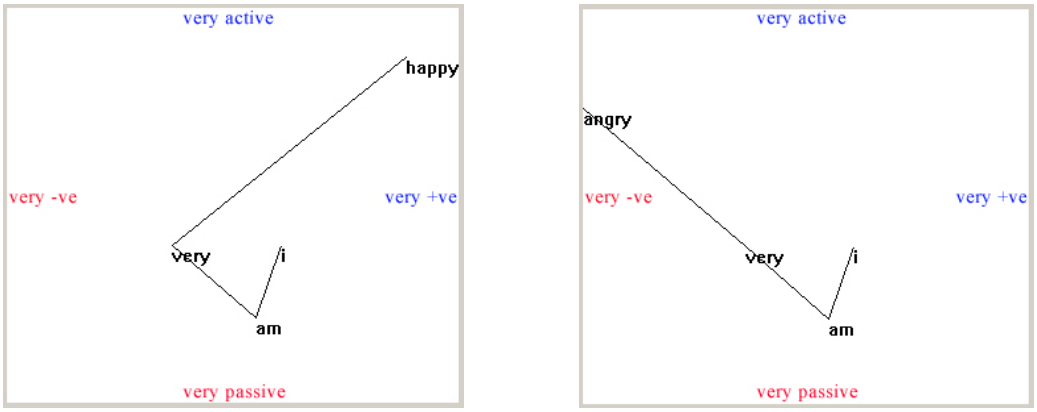
Ένα δείγμα του λεξικού φαίνεται στον Πίνακας 3.4, όπου οι περισσότερες λέξεις είναι κοντά στο κέντρο της διάστασης «αποτίμηση». Φαίνεται λογικό η λέξη "a" να είναι κοντά στο κέντρο, όπως και η λέξη "absent" όπου συνδέεται περισσότερο με αρνητική αποτίμηση.

ΛΕΞΗ	**ΑΠΟΤΙΜΗΣΗ (1)-(3)**	**ΔΡΑΣΤΗΡΙΟΠΟΙΗΣΗ (1)-(3)**
a	2	1.3846
abandon	1	2.375
abandoned	1.1429	2.1
abandonment	1	2
abated	1.6667	1.3333
abilities	2.5	2.1111
ability	2.5714	2.5
able	2.2	1.625
abnormal	1	2
aboard	1.8	1.875
abolition	1.5	2.1818
abortion	1	2.7273
about	1.7143	1.3
above	2.2	1.25
abroad	2.6	1.75
abrupt	1.2857	2.3
abruptly	1.1429	2.2
abscess	1.125	1.5455

absence	1.5	1.5556
absent	1	1.3
absolute	1.6667	1.4444
absolutely	1.6	1.5
absorb	1.8	1.75
absorbed	1.4	1.625
absorption	1.7778	1.6667
abstract	1.6667	1.4444
abstraction	1.4286	1.4
absurd	1	1.5
abundance	2.6667	1.5556
abuse	1.4286	2.5
abusers	1.25	2.7273
abusing	1.25	2.8182
abusive	1.6667	2.6667

Πίνακας 3.4 Οι πρώτες λέξεις του DAL από την C.Whissell.

Το παρακάτω σχήμα δείχνει την τροχιά που διαγραφεί το συναίσθημα κάθε μιας από τις δυο προτάσεις “I am angry” και “I am happy”που αντιστοιχούν σε δυο τελείως διαφορετικές συναισθηματικές καταστάσεις. Η αναπαράσταση της τροχιάς κάθε πρότασης γίνεται πάνω στον δυσδιάστατο χώρο της δραστηριοποίησης-αποτίμησης.



Σχήμα 3.4Η τροχιά του συναισθήματος για δυο αντίθετες συναισθηματικά προτάσεις.

3.5.2 Αλγόριθμος παραγωγής ενισχυμένου γλωσσικού μοντέλου

Η αναγνώριση συναισθηματικά φορτισμένου λόγου επιβάλει το χειρισμό κάποιων γλωσσικών φαινόμενων που δεν παρουσιάζονται στο λόγο υπαγόρευσης. Αν και αυτά όλα τα φαινόμενα δεν προκαλούν δυσκολία στην κατανόηση του νοήματος των λόγων στον άνθρωπο, είναι φανερό ότι υποβαθμίζουν την απόδοση ενός κλασσικού συστήματος αναγνώρισης φωνής. Γι' αυτόν τον λόγο αυτή η εργασία προτείνει ένα αλγόριθμο που βελτιώνει την απόδοση του συστήματος με την χρήση ενός γλωσσικού μοντέλου ενισχυμένου με συναισθηματικά δεδομένα (βλέπε Σχήμα 3.5). Ακολουθώντας τα βήματα του αλγορίθμου για την εξαγωγή κειμένου με συναισθηματικά δεδομένα χρησιμοποιούμε σαν σώμα κειμένου αναφοράς το Εθνικό Βρετανικό Σώμα Κειμένου. Με την βοήθεια του λεξιλογίου Whissell (1989) που περιλαμβάνει 8742 όρους-λέξεις που είναι βαθμολογημένοι βάσει των παραμέτρων Δραστηριοποίησης και Αποτίμησης εξάγονται προτάσεις με συναισθηματικό περιεχόμενο από το Εθνικό Βρετανικό Σώμα Κειμένου. Είναι γνωστό ότι η συναισθηματικότητα του ομιλητή επηρεάζει τόσο τα προσωδιακά του χαρακτηριστικά όσο και το περιεχόμενο των όρων-λέξεων που χρησιμοποιεί. Συνεπώς ένας πρακτικός τρόπος να μοντελοποιηθεί η χρήση των λέξεων βάσει της συναισθηματικής κατάστασης του ομιλητή είναι να ενισχυθεί ένα σώμα κειμένου, εν προκειμένω, το Εθνικό Βρετανικό Σώμα Κειμένου, με δεδομένα που δεν είναι τίποτα άλλο από προτάσεις με έντονο συναισθηματικό υπόβαθρο. Το συναισθηματικά εμπλουτισμένο σώμα κειμένου χρησιμοποιείται για την κατασκευή του γλωσσικού μοντέλου.

Το πρώτο βήμα του αλγορίθμου περιλαμβάνει την εξαγωγή από το Εθνικό Βρετανικό Σώμα Κειμένου, προτάσεων με όρους – λέξεις που ανήκουν επίσης και στο σύνολο των λέξεων του λεξιλογίου Whissell. Το παραγώμενο σώμα κειμένου ονομάζεται Whissell Σώμα Κειμένου.

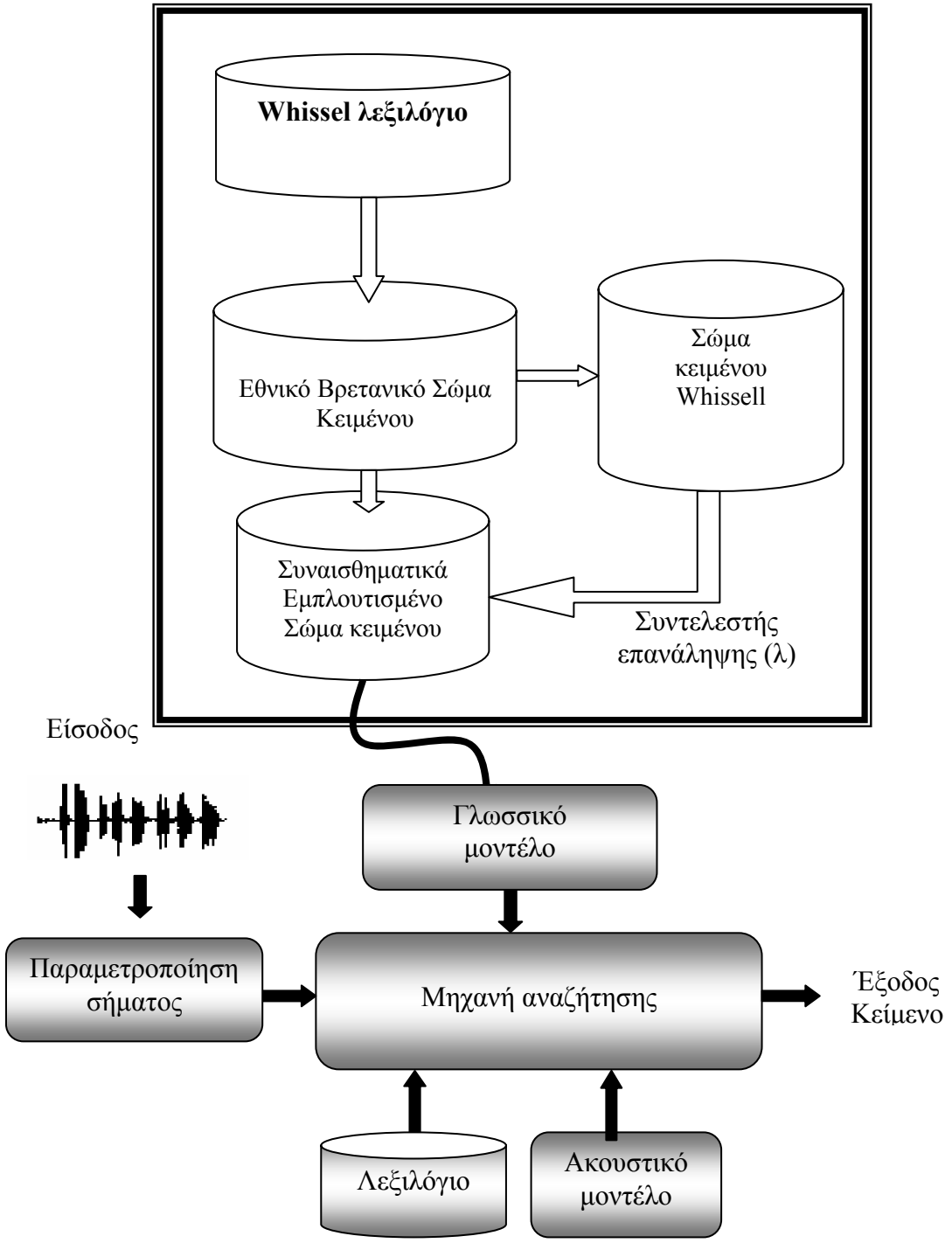
Κατόπιν, το Whissell Σώμα Κειμένου αφού επαναληφθεί λ φορές προσαρτάται στο Εθνικό Βρετανικό Σώμα Κειμένου, ώστε να σχηματιστεί ένα σώμα κειμένου εμπλουτισμένο με συναισθηματικά δεδομένα (σώμα κειμένου με συναισθηματική βαρύτητα). Το συναισθηματικά εμπλουτισμένο σώμα κειμένου χρησιμοποιείται σαν βάση για την εκπαίδευση του επαυξημένου γλωσσικού μοντέλου. Ο συντελεστής επανάληψης λ αυξάνεται βηματικά ώστε να μεγιστοποιηθεί η απόδοση του συστήματος αναγνώρισης φωνής (για λ=0, έχουμε μόνο το βασικό γλωσσικό μοντέλο). Η παρακάτω σχέση αποτυπώνει τον τρόπο συνένωσης δυο διαφορετικών σωμάτων κειμένου ώστε να δημιουργηθεί ένα συναισθηματικά εμπλουτισμένο σώμα κειμένου:

$$S_{E.C} = S_{BNC} + \lambda.S_{Whissel} \tag{3.3}$$

Όπου S_{BNC} είναι το πλήθος των προτάσεων του Εθνικού Βρετανικού Σώματος Κειμένου, το $S_{Whissel}$ αναφέρεται στο πλήθος των προτάσεων του Whissell Σώματος Κειμένου, το λ είναι ο συντελεστής επανάληψης των προτάσεων του Whissell Σώματος Κειμένου και ο συνολικός αριθμός προτάσεων του παραγόμενου σώματος κειμένου, $S_{E.C}$.

Τα κυριότερα βήματα του αλγορίθμου συνοψίζονται παρακάτω:

1. Εξαγωγή από το Εθνικό Βρετανικό Σώμα Κειμένου, προτάσεων που όλες οι λέξεις τους ανήκουν στο λεξιλογίο Whissell (Whissell Σώμα Κειμένου)..
2. Το Whissell Σώμα Κειμένου αφού επαναληφθεί λ φορές προσαρτάται στο Εθνικό Βρετανικό Σώμα Κειμένου για την δημιουργία του συναισθηματικά εμπλουτισμένου σώματος κειμένου.
3. Ο συντελεστής επανάληψης λ προσδιορίζεται πειραματικά με την χρήση διαφορετικών τιμών ώστε να μεγιστοποιηθεί η απόδοση του συστήματος της αναγνώρισης φωνής.
4. Το παραγόμενο σώμα κειμένου χρησιμοποιείται για την εκπαίδευση του συναισθηματικά εμπλουτισμένου γλωσσικού μοντέλου.



Σχήμα 3.5 Η σχηματική αναπαράσταση του εμπλουτισμού του σώματος κειμένου με συναισθηματικές προτάσεις που εξάγονται από το Εθνικό Βρετανικό Σώμα Κειμένου σύμφωνα με το λεξιλόγιο Whissell στο πλαίσιο της κατασκευής του ενισχυμένου γλωσσικού μοντέλου για την αναγνώριση συναισθηματικού λόγου.

Στην ενότητα 3.6.2, αποδεικνύεται πειραματικά ότι η καλύτερη απόδοση του συστήματος αναγνώρισης φωνής επιτυγχάνεται για λ=10. Ο Πίνακας 3.5 παρουσιάζει μερικά σημαντικά χαρακτηριστικά των διαφορετικών σωμάτων κειμένων. Το Εθνικό Βρετανικό Σώμα Κειμένου περιλαμβάνει 6.25 M προτάσεις, ενώ το Whissell Σώμα Κειμένου έχει 0.3 M προτάσεις. Το παραγόμενο σώμα κειμένου με συναισθηματικά χαρακτηριστικά έχει 9.24 M προτάσεις, το αποτέλεσμα αυτό εξάγεται από την χρήση της σχέσης 3.3.

Κείμενο	**# προτάσεων**	**# λέξεων**	**Μέγεθος κειμένου (Bytes)**
Εθνικό Βρετανικό Σώμα Κειμένου	$6^{1/4}$ M	100 M	550 Mbytes
Whissell σώμα κείμενο	0.3 M	2 M	10 Mbytes
Συναισθηματικά εμπλουτισμένο σώμα κειμένου	$6^{1/4}$+10*0.3 =$9^{1/4}$M	120 M	550+10*10=650Mbytes

Πίνακας 3.5 Ο αριθμός των προτάσεων, λέξεων και το μέγεθος των σωμάτων κειμένου σε Megabytes, για το Εθνικό Βρετανικό Σώμα Κειμένου, Whissell Σώμα Κειμένου, και το συναισθηματικά εμπλουτισμένο σώμα κειμένου.

3.6 Αξιολόγηση του συστήματος αναγνώρισης φωνής με την χρήση του εμπλουτισμένου γλωσσικού μοντέλου

3.6.1 Πειραματικά δεδομένα

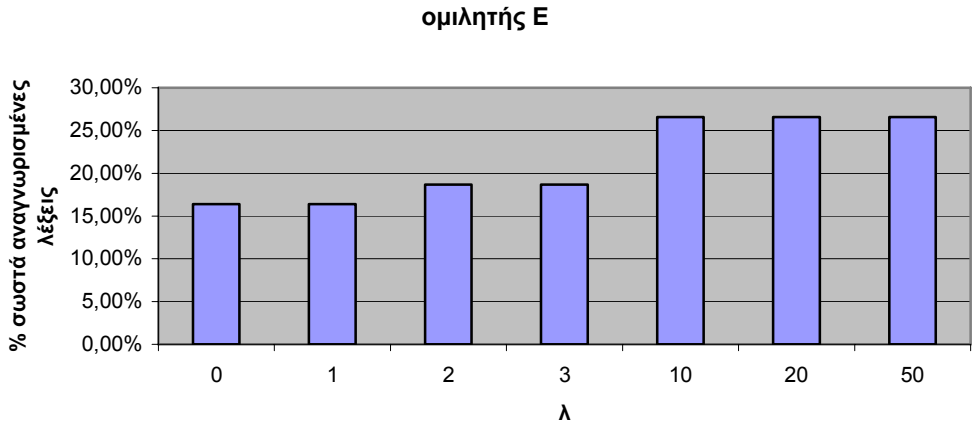
Το πειραματικό μέρος της εργασίας αυτής περιλαμβάνει το σύστημα αναγνώρισης φωνής με το βασικό και ενισχυμένο γλωσσικό μοντέλο και ένα σύνολο από ηχητικά αρχεία με προτάσεις που παρουσιάζουν συναισθηματική έμφαση. Τα ηχητικά αρχεία προέρχονται από την βάση δεδομένων με ελεύθερο συναισθηματικά χρωματισμένο λόγο, "SALAS". Τα ηχητικά αρχεία παρήχθησαν από συνομιλίες που διεξήχθησαν μεταξύ ανθρώπων και του συστήματος, SAL. Το σύστημα αυτό ανταποκρίνεται σε ομιλητές με προ-ηχογραφημένες συναισθηματικές φράσεις. Για την επιτυχή χρήση του συγκεκριμένου συστήματος απαιτείται μια εξοικείωση του χρήστη με αυτό το σύστημα. Από την στιγμή που ο χρήστης εξοικειωθεί με το σύστημα, αυτό ωθεί τον

χρήστη να εκφραστεί περισσότερο συναισθηματικά και να χρησιμοποιήσει διαφορετικές συναισθηματικές εκφράσεις για να διατυπώσει την σκέψη και τις ιδέες του. Ακολουθώντας αυτήν την μεθοδολογία παράχθηκε μια βάση δεδομένων με ελεύθερο συναισθηματικά φορτισμένο λόγο που περιλαμβάνει 4 ομιλητές, Βρετανοί πολίτες, (δυο άνδρες (I, R) και δυο γυναίκες (E,L)), και οι ηχογραφήσεις έχουν γίνει ακολουθώντας όλες τις προδιαγραφές για την βέλτιστη απόδοση του συστήματος αναγνώρισης φωνής. Η βάση αυτή έχει διάρκεια 160 λεπτά περιλαμβάνει 764 νοηματικές φράσεις-τμήματα που είναι ισο-κατανεμημένες στους ομιλητές. Η βάση αυτή καλύπτει όλες τις συναισθηματικές καταστάσεις που ανήκουν και στα τέσσερα τεταρτημόρια του τροχού συναισθημάτων Whissel.

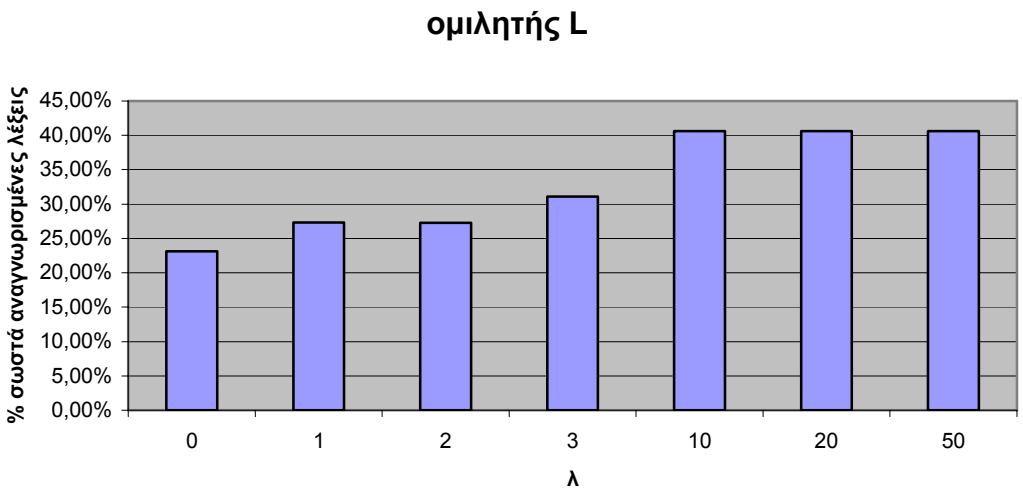
3.6.2 Αποτελέσματα με την χρήση του εμπλουτισμένου γλωσσικού μοντέλου

Τα παρακάτω σχήματα συνοψίζουν τα αποτελέσματα ανά ομιλητή και δείχνουν την απόδοση του συστήματος της αναγνώρισης συναισθηματικού λόγου σε συνάρτηση με το λ. Κάθε σχήμα παρουσιάζει το ποσοστό των σωστά αναγνωρισμένων λέξεων (y-άξονας) χρησιμοποιώντας διαφορετικές τιμές λ,(x-άξονας), ξεκινώντας από 0 και καταλήγοντας σε 50. Στην περίπτωση όπου το λ=0 συνεπάγεται ότι για τα πειράματα χρησιμοποιούμε μόνο το γλωσσικό μοντέλο που προέρχεται από το Εθνικό Βρετανικό Σώμα Κειμένου.

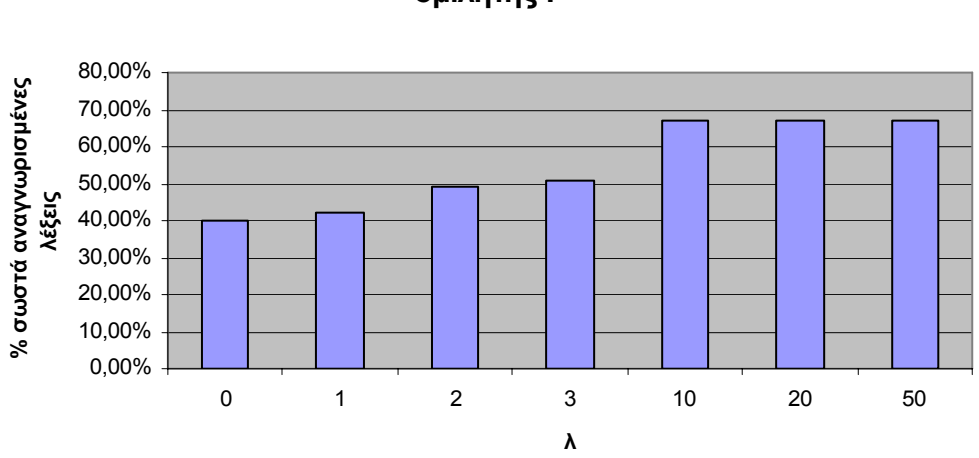
Το Σχήμα 3.6, αποτυπώνει τα αποτελέσματα για τον ομιλητή Ε. Στην περίπτωση του όπου το λ=0 και λ=1, το ποσοστό των σωστά αναγνωρισμένων λέξεων είναι 16,40% ενώ για λ=2 και 3, το αντίστοιχο μέγεθος είναι 18,69%. Για τιμές του λ μεγαλύτερες και ίσες με 10, το ποσοστό των σωστά αναγνωρισμένων λέξεων είναι 26,58%. Τα αντίστοιχα ποσοστά επιτυχίας για τους υπόλοιπους ομιλητές αποτυπώνονται στα σχήματα Σχήμα 3.7-3.9.



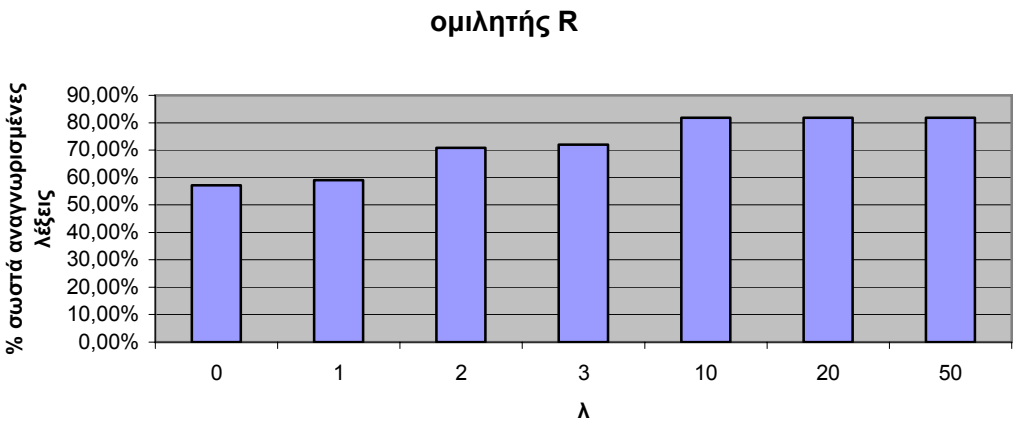
Σχήμα 3.6 Το ποσοστό των σωστά αναγνωρισμένων λέξεων για διαφορετικές τιμές του λ στην περίπτωση του ομιλητή Ε.



Σχήμα 3.7 Το ποσοστό των σωστά αναγνωρισμένων λέξεων για διαφορετικές τιμές του λ στην περίπτωση του ομιλητή L.

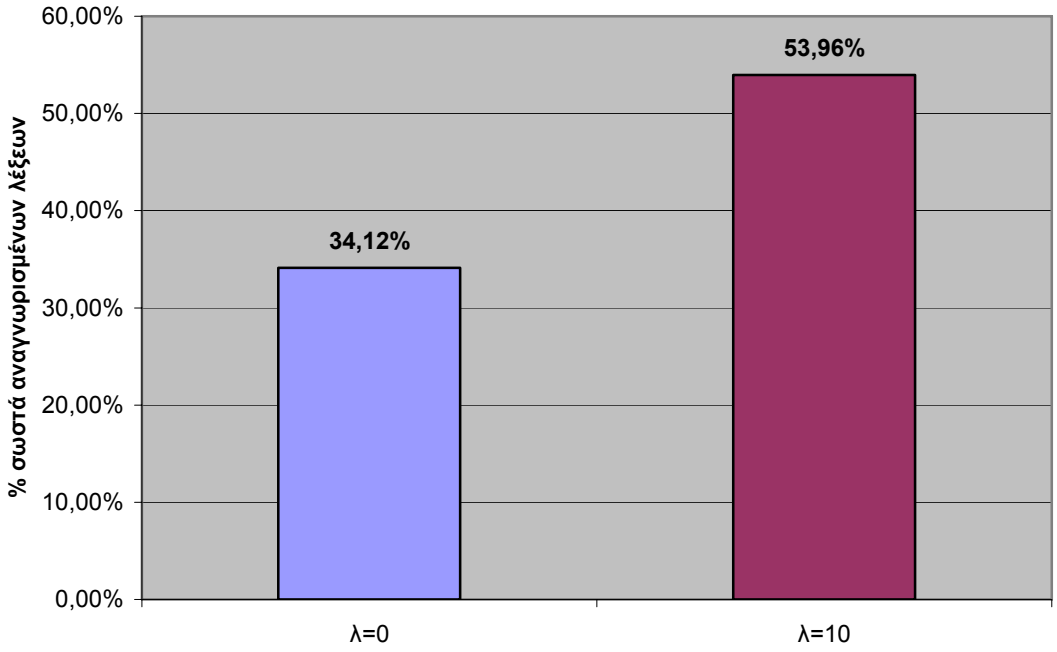


Σχήμα 3.8 Το ποσοστό των σωστά αναγνωρισμένων λέξεων για διαφορετικές τιμές του λ στην περίπτωση του ομιλητή I.



Σχήμα 3.9 Το ποσοστό των σωστά αναγνωρισμένων λέξεων για διαφορετικές τιμές του λ στην περίπτωση του ομιλητή R.

Σε γενικές γραμμές, διαπιστώνεται ότι αυξάνοντας τον παράγοντα λ μέχρι την τιμή 10, το ποσοστό των ορθών αναγνωρίσεων αυξάνει ανεξάρτητα από τον ομιλητή. Από κει και πέρα η αύξηση του συντελεστή λ δεν επηρεάζει το ποσοστό αναγνωρισμένων λέξεων για κανένα ομιλητή. Εξετάζοντας τα αποτελέσματα για δυο τιμές του λ, όταν το λ είναι ίσο με 0 που σημαίνει ότι χρησιμοποιείται μόνο το βασικό γλωσσικό μοντέλο που στηρίζεται στα δεδομένα του Εθνικού Βρετανικού Σώματος Κειμένου, και όταν αυτό παίρνει την τιμή 10 παρατηρούμε ότι το κατώτερο ποσοστό αναγνωρισμένων λέξεων παρατηρείται για τον ομιλητή Ε (26,58%), ενώ το υψηλότερο ποσοστό αναγνωρισμένων λέξεων για τον ομιλητή R με 81,84% όταν το λ=10. Αντίθετα, χωρίς την χρήση του εμπλουτισμένου γλωσσικού μοντέλου δηλαδή όταν $\lambda = 0$, τα αντίστοιχα αποτελέσματα διαμορφώνονται ως εξής: το κατώτερο ποσοστό αναγνωρισμένων λέξεων παρατηρείται για τον ομιλητή Ε (16,40%), ενώ το υψηλότερο ποσοστό αναγνωρισμένων λέξεων για τον ομιλητή R με 57,10%. Τα αποτελέσματα δείχνουν ότι η μεγαλύτερη βελτίωση με τον εμπλουτισμό του γλωσσικού μοντέλου, παρουσιάζεται για τον ομιλητή Ι με ποσοστό 26,94%, ενώ η μικρότερη για τον ομιλητή Ε, με 10,18%. Το παρακάτω Σχήμα 3.10 δείχνει την βελτίωση που επιφέρει η χρήση του συναισθηματικά εμπλουτισμένου γλωσσικού μοντέλου στην απόδοση του συστήματος. Τα αποτελέσματα αναφέρονται στο Μέσο Όρο (Μ.Ο) του ποσοστού των σωστά αναγνωρισμένων λέξεων για όλους τους ομιλητές. Η βελτίωση είναι της τάξης του 20%.



Σχήμα 3.10 Τα συγκριτικά αποτελέσματα για $\lambda = 0$ και $\lambda = 10$.

3.7 Εφαρμογή της αναγνώρισης φωνής με εμπλουτισμένο γλωσσικό μοντέλο σε συστήματα φυσικής επικοινωνίας ανθρώπου-μηχανής

3.7.1 Αρχιτεκτονική συστήματος επικοινωνίας ανθρώπου-μηχανής που λαμβάνει υπόψη το συναίσθημα

Οι φωνητικές διεπαφές με τους υπολογιστές είναι ένας τομέας που διεγείρει και εντυπωσιάζει όλους όσους ασχολούνται με την τεχνολογία φωνής εδώ και δεκαετίες. Για πολλούς, η δυνατότητα ενός χρήστη να συνομιλεί ελεύθερα με μια μηχανή είναι η κορυφή της πυραμίδας που στην βάση της βρίσκεται η κατανόηση της διαδικασίας παραγωγής και αντίληψης του λόγου που κατέχει εξέχουσα θέση στην επικοινωνία των ανθρώπων. Στην σημερινή εποχή τέτοιου είδους διεπαφές αποτελούν αναγκαιότητα και όχι πολυτέλεια. Τα διαδραστικά δίκτυα παρέχουν την δυνατότητα της εύκολης και άμεσης προσπέλασης ενός μεγάλου όγκου πληροφοριών και υπηρεσιών, επηρεάζοντας την ποιότητα της εργασίας και των καθημερινών υποθέσεων. Οι εξελίξεις στην τεχνολογία της ανθρώπινης γλώσσας είναι αναγκαίες για τον

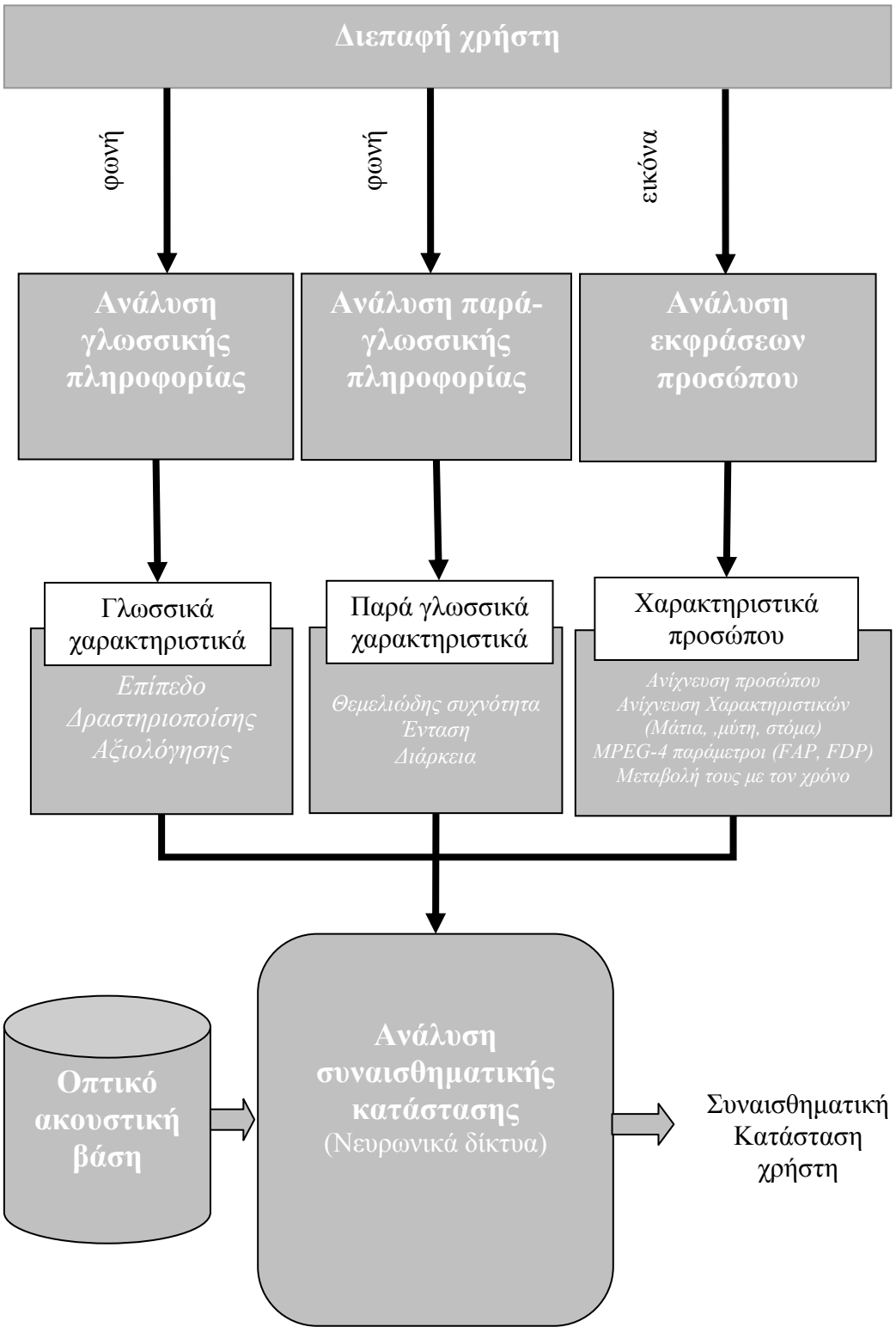
μέσο πολίτη στο να επικοινωνήσει με συσκευές όπως το τηλέφωνο και η τηλεόραση. Χωρίς σημαντική πρόοδο στον τομέα των ανθρωποκεντρικών διεπαφών, ένα μεγάλο κομμάτι της κοινωνίας θα αποτραπεί από το να συμμετέχει σε αυτό που ονομάζουμε κοινωνία της πληροφορίας με αποτέλεσμα την ολοένα και μεγαλύτερη απομόνωση τους και την παράλληλη απώλεια της δημιουργικότητας τους.

Το συναίσθημα στον λόγο είναι ένας τομέας που λαμβάνει όλο και μεγαλύτερο ενδιαφέρον τα τελευταία χρόνια, τόσο σε επίπεδο σύνθεσης φωνής όσο και σε επίπεδο της αυτόματης αναγνώρισης φωνής. Ο στόχος είναι να σχεδιαστούν συστήματα που αντιλαμβάνονται και ανταποκρίνονται σε ανθρώπινα συναισθήματα. Είναι προφανές ότι ίδιες λέξεις μπορούν να χρησιμοποιηθούν σαν αστείο ή σαν μια γνήσια ερώτηση που ζητάει απάντηση η σαν μια επιθετική πρόκληση. Η πιο διαδεδομένη μέθοδος προσέγγισης για την δημιουργία αυτόματου συστήματος ανάλυσης συναισθηματικού λόγου είναι η χρήση μιας βάσης δεδομένων με συναισθηματικό λόγο η οποία έχει σχολιαστεί με συναισθηματικές ετικέτες από ένα σύνολο ακροατών. Το επόμενο βήμα είναι να πραγματοποιηθεί η ακουστική ανάλυση των δεδομένων αυτών ώστε να συσχετιστούν συγκεκριμένα ακουστικά χαρακτηριστικά, όπως η θεμελιώδης συχνότητα με τις συναισθηματικές ετικέτες. Στο τελευταίο βήμα οι εκτιμήσεις των παραμέτρων πιστοποιούνται και προσαρμόζονται με την αξιολόγηση της απόδοσης του συστήματος σε διάφορες δοκιμές.

Ο κύριος στόχος ενός ανάλογου συστήματος διεπαφής ανθρώπου μηχανής που λαμβάνει υπόψη την συναισθηματική κατάσταση του ομιλητή, είναι να μπορεί να ερμηνεύει την συναισθηματική του κατάσταση π.χ την ανία , την πλήξη, την χαρά, την λύπη, και τον θυμό, με βάση τον προφορικό λόγο και τις κινήσεις του προσώπου του. Οι τεχνολογίες που χρησιμοποιούνται αφορούν την γλωσσική και την παραγλωσσική ανάλυση, την αναγνώριση φωνής, την ανάλυση των εκφράσεων του προσώπου και τον συνδυασμό όλων αυτών με την χρήση υβριδικών και fuzzy τεχνικών για την εξαγωγή ενός συμπεράσματος αναφορικά με την συναισθηματική κατάσταση του χρήστη (Η αρχιτεκτονική του συστήματος και των επιμέρους στοιχείων του απεικονίζεται στο Σχήμα 3.11).

Συνοψίζοντας, τα στοιχεία του συστήματος που θα χρησιμοποιηθούν για να επιτευχθεί ο στόχος είναι τα εξής:

- υποσύστημα ανάλυσης γλωσσικής πληροφορίας
- υποσύστημα ανάλυσης παρά-γλωσσικής πληροφορίας
- υποσύστημα ανάλυσης προσώπου
- και ο πυρήνας της αναγνώρισης συναισθήματος.



Σχήμα 3.11 Η αρχιτεκτονική του συστήματος (Ευρωπαϊκό Πρόγραμμα ERMIS / IST-2000-29319).

Όπως φαίνεται μέσω της κατάλληλης διεπαφής συλλέγονται σήματα φωνής και video, ώστε η εικόνα και η φωνή του χρήστη να χρησιμοποιούνται σαν είσοδο στα διαφορετικά υποσυστήματα ανάλυσης.

Όσον αφορά την ανάλυση φωνής συντελείται σε δυο μέρη ώστε να επεξεργαστούν και οι δυο πληροφορίες που φέρει η φωνή, αυτή της γλωσσικής πληροφορίας, πληροφορία σχετική με την έννοια των λέξεων, την δομή και τους κανόνες της γλώσσας, και από την άλλη η παραγλωσσική πληροφορία που σχετίζεται με την προσωδία του λόγου. Στον τομέα της ανάλυσης της γλωσσικής πληροφορίας χρησιμοποιείται ένα αυτόματο σύστημα αναγνώρισης φωνής που βασίζεται στην τεχνική των κρυφών Μαρκοβιανών Μοντέλων και στο επαυξημένο γλωσσικό μοντέλο.

Η ανάλυση της παρα-γλωσσικής πληροφορίας περιλαμβάνει την ανάλυση των μεταβολών του pitch, της έντασης που δεν έχουν γλωσσική λειτουργία και της ποιότητας της φωνής που σχετίζεται με τα φασματικά χαρακτηριστικά και όχι με την ταυτότητα των λέξεων. Η ανάλυση της παρα-γλωσσικής πληροφορίας αφορά κυρίως τον τρόπο με τον οποίον εκφέρονται οι λέξεις.

Το σύστημα ανάλυσης του προσώπου θα έχει σαν στόχο πρώτον να ανιχνεύει σε μια εικόνα το πρόσωπο και έπειτα τα χαρακτηριστικά του προσώπου ώστε να εξάγει πληροφορία σχετική με την κίνηση των χαρακτηριστικών του προσώπου. Έτσι αρχικά θα γίνεται μια διαδικασία ανίχνευσης του τι είναι πρόσωπο και το τι όχι βάσει του χρώματος του δέρματος. Κατόπιν θα ακολουθεί ο εντοπισμός της θέσης και του σχήματος του στόματος, των χειλιών, των ματιών και των φρυδιών. Για αυτόν τον λόγο θα χρησιμοποιηθούν στοιχεία από το MPEG-4 όπως της κίνησης του προσώπου (FAP) και του ορισμού του προσώπου (FDP).

Το υποσύστημα της αναγνώρισης συναισθήματος θα συνδυάσει τα στοιχεία που θα εξαχθούν από τις επιμέρους αναλύσεις των ακουστικών και video σημάτων, ώστε να παρέχει ενδείξεις σχετικά με την συμπεριφορά και την συναισθηματική κατάσταση του χρήστη. Τα διάφορα χαρακτηριστικά της φωνής και προσώπου θα επεξεργαστούν ξεχωριστά αλλά και από κοινού με την χρήση εξειδικευμένων τεχνικών, μαζί με υβριδικές τεχνικές που μπορούν να κωδικοποιήσουν και να εμπλουτίσουν την εκ των προτέρων γνώση σχετικά με την ανάλυση εκφράσεων και συναισθήματος.

3.8 Συμπεράσματα

Στο κεφάλαιο 3 παρουσιάσθηκε ο αλγόριθμος εφαρμογής των N-grams στην αναγνώριση συναισθηματικού λόγου. Αναλύοντας τα αποτελέσματα της μεθόδου ενίσχυσης του γλωσσικού μοντέλου με συναισθηματικά δεδομένα μπορούμε να πούμε ότι η μέθοδος λειτουργεί αποτελεσματικά σε κάθε ομιλητή Διαπιστώνεται ότι αυξάνοντας τον παράγοντα λ μέχρι να πάρει την τιμή 10, το ποσοστό των ορθών αναγνωρίσεων αυξάνει ανεξάρτητα του ομιλητή και

από εκεί και πέρα δεν μεταβάλλεται. Τα αποτελέσματα δείχνουν ότι με την χρήση του εμπλουτισμένου γλωσσικού μοντέλου η μικρότερη βελτίωση εμφανίζεται στην περίπτωση του ομιλητή E (10,18%), και στην περίπτωση του ομιλητή L, με ποσοστό 17,48%. Από την άλλη πλευρά, φαίνεται ότι η επίδραση του εμπλουτισμένου γλωσσικού μοντέλου επιφέρει σημαντική βελτίωση στα αποτελέσματα του ομιλητή I με ποσοστό 26,94% και του ομιλητή R με ποσοστό 24,74%. Ομαδοποιώντας τα αποτελέσματα βάσει του φύλου, μπορούμε να επισημάνουμε ότι για τις γυναίκες ομιλητές το ποσοστό αυξάνεται από 19,76% (λ =0) σε 33,59% (λ =10), ενώ με τους άνδρες ομιλητές το ποσοστό αυξάνεται από 48,49% (λ =0) σε 74,33% (λ =10).

Στο σημείο αυτό θα πρέπει να τονιστεί ότι η χρήση του συναισθητικά προσανατολισμένου γλωσσικού μοντέλου δίνει τα ίδια αποτελέσματα με αυτά του βασικού γλωσσικού μοντέλου όταν τα πειραματικά δεδομένα προέρχονται από φωνητικές βάσεις υπαγόρευσης, όπως π.χ της φωνητικής βάσης του WSJ (Robinson, 1995).

Κεφάλαιο 4

4 Διόρθωση προτάσεων με αναδιάταξη των λέξεων χρησιμοποιώντας στατιστικές μεθόδους

4.1 Εισαγωγή

Ο όρος «σύνταξη» χρησιμοποιείται για να περιγράψει τις σχέσεις των στοιχείων-λέξεων μεταξύ τους, σε μια πρόταση, σε μια γλώσσα. Αυτό που ισχύει σε όλες τις γλώσσες είναι ότι οι λέξεις δεν μπορούν να τοποθετηθούν σε οποιαδήποτε θέση μέσα σε μια πρόταση, αλλά υπάρχουν συγκεκριμένοι κανόνες που ορίζουν την θέση μιας λέξης σε μια πρόταση. Για παράδειγμα στην Αγγλική γλώσσα ο φυσιολογικός τρόπος για να συνταχθεί μια πρόταση με την σωστή σειρά είναι ο ακόλουθος, πρώτα το υποκείμενο, στην συνέχεια το ρήμα, και στο τέλος το αντικείμενο (π.χ «Τα αγόρια συναντούν τα κορίτσια») (Hawkins, 1994).

Δεν πρέπει να παραλείψουμε να αναφερθούμε και σε κάποιες άλλες γλώσσες που παρουσιάζουν μια σχετική ελευθερία στην διάταξη των λέξεων, και μια τέτοια γλώσσα είναι τα Ελληνικά. Γλώσσες όπως η ελληνική παρουσιάζουν σημαντική ελευθερία στον τρόπο με το οποίον τοποθετούνται οι λέξεις σε μια πρόταση. Στην Αγγλική γλώσσα η θέση του ουσιαστικού καθορίζει τον ρόλο που παίζει στην πρόταση. Έτσι, ο κανόνας Υ-Ρ-Α (Υποκείμενο-Ρήμα-Αντικείμενο) δεν εφαρμόζεται πάντα το ίδιο υποχρεωτικά και στην Ελληνική γλώσσα. Η έμφαση που δίνεται σε μια πρόταση εξαρτάται από την σειρά με την οποίαν διατυπώνονται οι λέξεις αυτής της πρότασης, και επίσης η θέση των επιρρημάτων μπορεί να αλλάξει το νόημα των προτάσεων που τα περιλαμβάνουν. Έτσι μπορεί να βγει το συμπέρασμα ότι η σειρά των λέξεων εξαρτάται από το γραφικό στυλ, το οποίο με την σειρά του επηρεάζεται από την συναισθηματική κατάσταση του συγγραφέα.

Το κεφάλαιο αυτό οργανώνεται ως εξής: στην ενότητα 4.2 γίνεται μια επισκόπηση σε τεχνικές αναδιάταξης λέξεων που εφαρμόζονται στην μηχανική μετάφραση και στην διόρθωση προτάσεων. Η αρχιτεκτονική του όλου συστήματος μαζί με τα υποσυστήματα περιγράφονται στην ενότητα 4.3. Τα αποτελέσματα χρήσης των πειραματικών δεδομένων για την Αγγλική

γλώσσα παρουσιάζονται στην ενότητα 4.4. Στην ενότητα 4.5 παρατίθενται πληροφορίες για τα πειραματικά δεδομένα και αποτελέσματα για την Ελληνική γλώσσα. Στην ενότητα 4.6 δίνονται τα κύρια χαρακτηριστικά της μελέτης που έγινε για την αξιολόγηση της προτεινόμενης μεθόδου από ομάδα χρηστών. Τέλος τα συμπεράσματα και οι παρατηρήσεις περιλαμβάνονται στην ενότητα 4.7.

4.2 Τεχνικές αναδιάταξης λέξεων με την χρήση στατιστικών μεθόδων

4.2.1 Εφαρμογή στην μηχανική μετάφραση

Όπως έχουμε αναφέρει στην αρχή αυτού του κεφαλαίου η προσπάθεια εστιάζεται στο να αναδιαταχθούν οι λέξεις μιας πρότασης με τέτοιο τρόπο ώστε να περιοριστεί ο χώρος αναζήτησης της πρότασης που έχει τον μεγαλύτερο αριθμό τριγραμμάτων που ανήκουν στο γλωσσικό μοντέλο. Για αυτόν τον λόγο σε αυτήν την ενότητα θα προσπαθήσουμε να δώσουμε μια εικόνα του τι έχει γίνει μέχρι σήμερα σε επίπεδο αναδιάταξης των λέξεων μιας πρότασης. Τεχνικές για την αναδιατύπωση των λέξεων συναντάμε στον τομέα της μηχανικής μετάφρασης. Ο σκοπός της μηχανικής μετάφρασης είναι να επιτύχει την καλύτερη δυνατή μετάφραση μιας πηγαίας γλώσσα σε μια γλώσσα στόχου. Στο πλαίσιο αυτής της διαδικασίας έχουν αναπτυχθεί αρκετές τεχνικές με πιο αξιόλογες την βασισμένη σε σώμα κειμένου μηχανική μετάφραση (Corpus based Machine Translation) που περιλαμβάνει την στατιστική μηχανική μετάφραση και την βασισμένη σε παραδείγματα μηχανική μετάφραση, και οι δύο αυτές τεχνικές χρησιμοποιούν δίγλωσσα παράλληλα κείμενα για να εκπαιδευτεί το μεταφραστικό μοντέλο. Η στατιστική μηχανική μετάφραση στηρίζεται στην θεωρεία πιθανοτήτων ενώ η βασισμένη σε παραδείγματα μηχανική μετάφραση στηρίζεται στην λογική όπου κάθε νέα μετάφραση υπολογίζεται σε αναλογία με τις ήδη γνωστές μεταφράσεις που έχουν γίνει από το δίγλωσσο σώμα κειμένων (Carl and Way, 2003). Μεταφράζοντας τις λέξεις μια προς μια από την μια γλώσσα στην άλλην είναι κατανοητό ότι λόγω της δομής της κάθε γλώσσας θα υπάρχει μια ασυμφωνία στην πλευρά της γλώσσας στόχου. Σε επίπεδο συντακτικό υπάρχουν τροποποιήσεις στην διατύπωση της σειράς των λέξεων που αντιστοιχούν σε ρήματα, υποκείμενα και αντικείμενα από γλώσσα σε γλώσσα. Παραδείγματος χάριν η Γερμανική, Αγγλική, είναι γλώσσες όπου ισχύει ο κανόνας Y-P-A ενώ σε άλλες όπως τα Ιαπωνικά ισχύει ο κανόνας P-Y-A. Ένα σημαντικό στοιχείο της όλης διαδικασίας της μηχανικής μετάφρασης είναι να βρεθεί ο κατάλληλος τρόπος μετατόπισης των λέξεων από μια θέση σε μια άλλη, ώστε η δομή της εκάστοτε γλώσσας να μην παραβιάζεται. Για αυτόν τον λόγο σε αυτήν την ενότητα θα αναδείξουμε κάποιες τεχνικές όπου ο τρόπος αναδιάταξης των λέξεων βασίζεται σε

γλωσσολογικούς κανόνες ή σε στατιστικές μεθόδους. Οι ερευνητές της IBM ήταν οι πρώτοι που περιέγραψαν την τεχνική στατιστικής μετάφρασης με τους Brown et al. (1991). Έτσι, κατά την εκπαίδευση του μοντέλου με την χρήση των παράλληλων κειμένων κάθε λέξη της πρότασης στόχου που βρίσκεται στην θέση j έχει πιθανότητα $a(j \mid i,V,T)$ να είναι το προϊόν μετάφρασης μιας λέξης της πηγαίας πρότασης που βρίσκεται στην θέση i όταν V είναι το μέγεθος της πρότασης στόχου και T είναι το μέγεθος της πηγαίας πρότασης. Η τελική αξιολόγηση των προτάσεων γίνεται με την βοήθεια των διγραμμάτων του γλωσσικού μοντέλου. Η κριτική που γίνεται στο τρόπο μετάφρασης που εισήγαγε η IBM είναι ότι δεν μοντελοποιεί τις δομικές και συντακτικές πλευρές της γλώσσας. Με αποτέλεσμα όταν γίνεται μετάφραση από γλώσσα προς γλώσσα με την ίδια συντακτική δομή να μην αντιμετωπίζεται κάποιο ιδιαίτερο πρόβλημα ενώ όταν το ζευγάρι των γλωσσών είναι διαφορετικό σε επίπεδο σειράς των λέξεων τότε το αποτέλεσμα της μετάφρασης δεν είναι το αναμενόμενο. Στην προσπάθεια να ενσωματωθεί η συντακτική πληροφορία της γλώσσας στην μηχανική μετάφραση λαμβάνεται υπόψη η πληροφορία που φέρει το συντακτικό δένδρο της πηγαίας πρότασης (Yamada and Knight, 2001). Οι όποιες μετακινήσεις γίνονται σε επίπεδο κόμβων του συντακτικού δένδρου. Πρώτα οι κόμβοι τέκνα σε κάθε εσωτερικό κόμβο αναδιατάσσονται με συνέπεια για Ν τέκνα θα έχουμε ένα αριθμό της τάξης των Ν! αντιμεταθέσεων. Οι πιθανότητες για κάθε μια από τις αντιμεταθέσεις προέρχεται από μια βάση δεδομένων που αποτελείται από N-grams που υπολογίζονται σε tags και όχι σε λέξεις. Συνεπώς το πόσο πιθανή είναι μια αντιμετάθεση από μια άλλη εξαρτάται από την πιθανότητα εμφάνισης των συγκεκριμένων tags με την συγκεκριμένη σειρά. Το ευρωπαϊκό πρόγραμμα METIS II επιδιώκει να υλοποιήσει μια τεχνική μετάφρασης σε επίπεδο πρότασης (Dologlou et al., 2003). Χρησιμοποιώντας ένα δίγλωσσο λεξικό και κάποιους κανόνες αναδιάταξης των λέξεων, είναι δυνατόν να επιτευχθεί η μετάφραση λέξη με λέξη, της πηγαίας πρότασης. Έτσι στο σώμα κειμένων της γλώσσας στόχου γίνεται αναζήτηση της πιο κοντινής πρότασης προς την πηγαία πρόταση, ως προς την δομή της. Το πραγματικό πρόβλημα αυτής της μεθόδου παραμένει το μέγεθος της πρότασης. Για αυτόν τον λόγο χρησιμοποιούνται τμήματα της πρότασης (constituents) με αποτέλεσμα η σύγκριση να γίνεται μεταξύ τμημάτων προτάσεων και όχι προτάσεων. Το σύστημα αυτό θέλοντας να διαχειριστεί τις αλλαγές στη δομή της γλώσσας στόχου χρησιμοποιεί ένας POS tagger για να διακρίνει τις λέξεις περιεχομένου και τις λέξεις γραμματικής στην κάθε πρόταση. Επιπλέον οι μετατοπίσεις διακρίνονται σε τοπικές και σε μη τοπικές. Έτσι σε τοπικό επίπεδο οι μόνες μετατοπίσεις που επιτρέπονται είναι αυτές των λέξεων περιεχομένου όπως ρήματα, επιρρήματα και επίθετα, ενώ μετατοπίσεις των άρθρων δεν γίνονται αποδεκτές. Η ορθότητα της αλλαγής της σειρά των λέξεων ελέγχεται με την χρήση του γλωσσικού μοντέλου της γλώσσας στόχου. Ενώ σε μη τοπικό επίπεδο (constituents) ο έλεγχος της σειράς των λέξεων επιτυγχάνεται μέσω

της χρήσης ενός δεύτερης τάξης γλωσσικού μοντέλου όπου τα μέλη των N-grams που αποτελούνται απο tags. Ο Koenh (2003) παρουσίασε μια νέα μέθοδο για την αναδιάταξη λέξεων μέσα σε μια πρόταση χρησιμοποιώντας όχι τις λέξεις αλλά τις φράσεις, που περιλαμβάνονται σε αυτήν. Με αυτόν τον τρόπο η πηγαία πρόταση μιας γλώσσας τεμαχίζεται σε φράσεις και κάθε φράση μεταφράζεται στην αντίστοιχη φράση της γλώσσας στόχου. Τέλος θα ήταν σκόπιμο να παρουσιάσουμε ακόμη μια τεχνική που εισήγαγε ο Tillmann (2004) όπου στην μετάφραση προτάσεων από την μια γλώσσας την άλλη συμμετέχουν τμήματα λέξεων και όχι λέξεις. Έτσι αυτό που ενδιαφέρει είναι να αναπτυχθεί ένα μοντέλο προσανατολισμού των τμημάτων, με καθένα τμήμα να έχει την δυνατότητα μετατόπισης αριστερά, δεξιά και ουδέτερα. Αυτό δίνει την δυνατότητα να περιοριστεί ο χώρος αναζήτησης για την καταλληλότερη μετάφραση αφού για την αναδιάταξη των τμημάτων, το μόνο που επιτρέπεται είναι η αμοιβαία αλλαγή θέσεων μεταξύ 2 γειτονικών τμημάτων. Άρα η διαδικασία της μετάφρασης περιλαμβάνει τον τεμαχισμό της πρότασης σε τμήματα όπου η πρόταση εισόδου τεμαχίζεται αριστερά προς δεξιά και η πρόταση στόχου από κάτω προς τα πάνω. Το αποτέλεσμα οδηγεί σε μια μονότονη ακολουθία τμημάτων με κάποιες εξαιρέσεις όσον αναφορά την αμοιβαία αλλαγή θέσης μεταξύ 2 γειτονικών τμημάτων.

4.2.2 Εφαρμογή στην διόρθωση προτάσεων

Ο αυτόματος έλεγχος προτάσεων γίνεται με τον παραδοσιακό τρόπο της χειρονακτικής συλλογής γραπτών κανόνων με σκοπό να περιγραφούν διαφορετικοί τύποι λαθών και να εφαρμοσθούν οι κανόνες αυτοί σε νέες υπό εξέταση προτάσεις. Υπάρχουν όμως και άλλες τεχνικές που χρησιμοποιούν στατιστικές μεθόδους για την ανίχνευση γραμματικών λαθών σε διάφορες προτάσεις. Στην ενότητα αυτή θα επιδιώξουμε να κάνουμε μια σύντομη περιγραφή σε αντίστοιχες μεθόδους που χρησιμοποιούνται στον αυτόματο γραμματικό έλεγχο. Ο Atwell (1987) παρουσίασε μια μέθοδο ανίχνευσης γραμματικών λαθών κάνοντας χρήση των πιθανοτήτων ενός στατιστικού POS αναλυτή. Τα λάθη εντοπίζονται από την στιγμή που υπάρχουν ακολουθίες POS με μικρές πιθανότητες. Για τον καθορισμό ενός POS tag μιας λέξης χρησιμοποιείται ένα απλό Markov Model αντί για πιθανοτικές, ανεξάρτητες συμφραζομένων γραμματικές. Στην περίπτωση που έχουμε αμφιλεγόμενα tags λέξεων τότε χρησιμοποιείται εκείνη η ακολουθία από tags που έχει μεγαλύτερη πιθανότητα. Η πιθανότητα υπολογίζεται με την χρήση του γινομένου όλων των δεσμών μεταξύ των tags. Το πλεονέκτημα αυτής της μεθόδου είναι ότι δεν απαιτείται η χρήση συντακτικού αναλυτή, και επίσης μπορεί να εφαρμοσθεί σε οποιαδήποτε γλώσσα αρκεί να υπάρχει ένα λεξικό και ο πίνακας με τις πιθανότητες εμφάνισης των tags κάθε λέξης. Η μέθοδος που πρότειναν οι Chodorow και Lealock (2000) για την ανίχνευση γραμματικών λαθών με την χρήση αρνητικής μαρτυρίας από

γραπτά κείμενα χρησιμοποιήθηκε στην δημιουργία ενός αυτοματοποιημένου συστήματος ALEK (Assessing Lexical Knowledge) με σκοπό την άμεση αναγνώριση μη κατάλληλων λέξεων σε κείμενα εκθέσεων κοιτώντας τα συμφραζόμενα γύρω από κάθε μια από αυτές. Η τεχνική αυτή πραγματοποιεί μια σύγκριση των συμφραζομένων της υπο-εξέτασης λέξης με τα συμφραζόμενα της ίδιας λέξης που απορρέουν από κείμενα τα οποία είναι γραμματικά και συντακτικά σωστά. Για αυτό τον λόγο χρησιμοποιείται ένα μεγάλο σώμα κειμένου με 30 εκατομμύρια λέξεις και περί των 10000 προτάσεων για κάθε λέξη στόχο. Γύρω από την λέξη στόχο χρησιμοποιείται ένα παράθυρο της τάξης των +-2 λέξεων ώστε να εντοπιστούν στα συμφραζόμενα λειτουργικές λέξεις και tags έτσι ώστε να ομαδοποιηθούν οι λέξεις που ανήκουν στην ίδια κατηγορία. Συνδυασμοί λέξεων που συμβαίνουν σπανιότερα από τους συνδυασμούς λέξεων που υφίστανται στο σώμα κειμένου θεωρούνται ως μη αποδεκτοί και υποδηλώνουν εστία γραμματικού λάθους. Οι Bigert και Knutsson (2002) επέλεξαν μια μέθοδο αναγνώρισης λαθών με την οποία ένα σώμα κειμένου γεμίζει με λάθη σε διάφορα σημεία και χρησιμοποιείται σαν δεδομένα εκπαίδευσης. Στα σημεία όπου προστέθηκαν λάθη γίνεται σημείωση με τον όρος λάθος. Έτσι το σύστημα μηχανικής μάθησης εκπαιδεύεται να ανιχνεύει νέα λάθη όμως του ίδιου τύπου σαν και αυτά που προστέθηκαν στο καθαρό κείμενο. Είναι πιθανόν ο αλγόριθμος αυτός να εκπαιδευτεί πάνω και σε άλλου είδους γραμματικά λάθη ώστε να τα αναγνωρίζει, αν και στην εργασία τους έχει γίνει προσπάθεια να αναγνωρίζονται μόνο λάθη που αφορούν την σειρά των λέξεων και ύπαρξη ή μη σύνθετων λέξεων. Ο αλγόριθμος επιδεικνύει και δυνατότητες διόρθωσης αυτών των λαθών έτσι ώστε όταν μια λέξη *w* ακολουθείται λάθος από μια λέξη *u* μπορεί το σύστημα να παρέχει την πληροφορία για το πώς μπορεί το λάθος να διορθωθεί και να αντιστραφεί η σειρά των λέξεων με πρώτο το *u* και να έπεται το *w* . Έτσι στην περίπτωση που υφίσταται λάθος στον διαμελισμό μιας σύνθετης λέξης χρησιμοποιείται ένα σώμα κείμενου με όλες τις σύνθετες λέξεις διασπασμένες και σημειωμένες σαν λάθος. Για την ανίχνευση λάθους στην σειρά των λέξεων μιας πρότασης χρησιμοποιούνται οι λέξεις, POS και επισημάνσεις σωστό/λάθος. Ο Naber (2003) ανέπτυξε έναν διορθωτή προτάσεων ανοικτού κώδικα για την Αγγλική γλώσσα χρησιμοποιώντας ένα σώμα κειμένων με λάθη από λίστες ηλεκτρονικού ταχυδρομείου. Το προς εξέταση κείμενο συγκρίνεται με τους κανόνες των λαθών που έχουν ορισθεί και αν βρεθεί ομοιότητα τότε το κείμενο θεωρητικά έχει κάποιο λάθος στην συγκεκριμένη θέση. Ο Sjobergh (2005) περιέγραψε μια μέθοδο χρήσης ενός chunker για εντοπισμό γραμματικών λαθών. Ο chunker χρησιμοποιείται για να τεμαχίσει την υπό εξέταση πρόταση σε μη αλληλο-καλυπτόμενες φράσεις. Έστω η φράση “the red car is parked on the sidewalk” τότε ο chunker δίνει [NP the red car] [VC is parked] [PP on the sidewalk]. Η απόδοση τέτοιων εργαλείων είναι πολύ ικανοποιητική δίνοντας συνήθως ένα ποσοστό ακρίβειας της τάξης του 90%. Επιπλέον ο chunker χρησιμοποιείται και για να υπολογισθούν τα

στατιστικά στοιχεία εμφάνισης συγκεκριμένων ακολουθιών με chunks σε ένα κείμενο που θα χρησιμοποιηθεί ως βάση. Έτσι λοιπόν με την βοήθεια του chunker αναζητούνται ακολουθίες από chunks στην υπό-εξέταση πρόταση που δεν έχουν εμφανιστεί στην βάση. Τα chunks αυτά επισημαίνονται σαν πιθανά λάθη. Είναι πιθανόν να χρησιμοποιηθεί κάποιο κατώφλι ώστε μια ακολουθία από chunks να θεωρείται αποδεκτή. Επίσης υπάρχει η δυνατότητα η ακολουθία των chunks να τροποποιηθεί και να δώσει σαν έξοδο όχι μια ακολουθία από chunks αλλά μια ακολουθία από φράσεις και chunks όπως η εξής NP-is parked-PP. Με αυτή την επιλογή είναι δυνατή η πρόβλεψη άλλων ειδών λαθών π.χ λάθος χρόνος. Τέλος ο More (2006) εισήγαγε ένα Αγγλικό διορθωτή προτάσεων για άτομα που δεν έχουν ως μητρική την Αγγλική γλώσσα. Το κύριο χαρακτηριστικό αυτού του διορθωτή είναι η χρήση μιας οποιασδήποτε μηχανής αναζήτησης μέσω του διαδικτύου. Καθώς ένας μεγάλος αριθμός ιστοσελίδων γράφονται στα Αγγλικά το προτεινόμενο σύστημα θεωρεί ότι μια ακολουθία λέξεων ή γραμμάτων που δεν συγκαταλέγονται στις σελίδες αυτές εμπεριέχουν πιθανόν λάθος. Με αποτέλεσμα, ο χρήστης να ενημερώνεται για το πιθανό λάθος και η μηχανής αναζήτησης να προτείνει κάποιο παραπλήσιο περιεχόμενο που μπορεί να βοηθήσει τον χρήστη να αποφασίσει αν θα το διορθώσει ή όχι.

4.3 Περιγραφή της μεθόδου διόρθωσης προτάσεων

4.3.1 Η αρχιτεκτονική του συστήματος

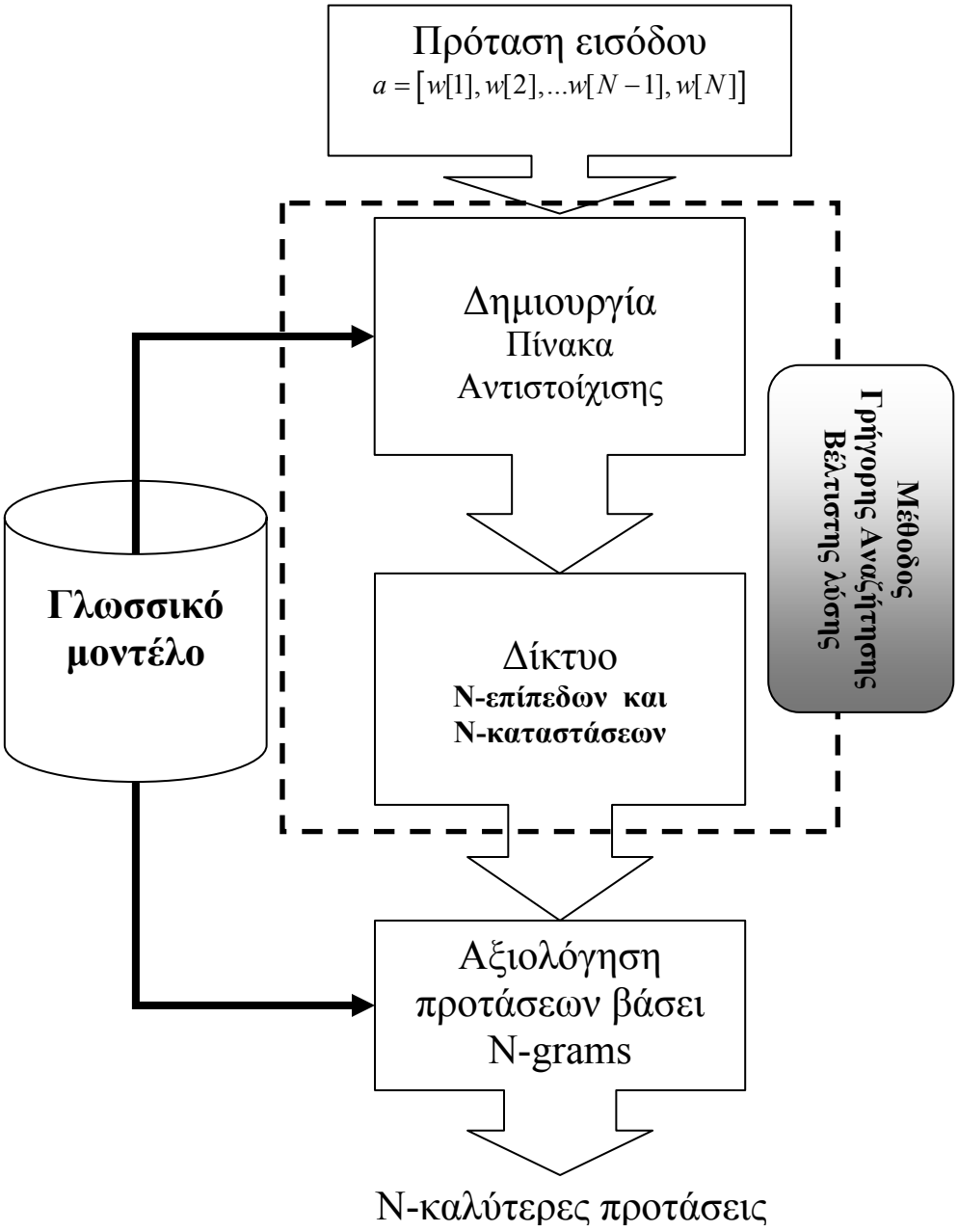
Οι άνθρωποι μερικές φορές κάνουν κάποια λάθη που είναι αντίθετα στους συντακτικούς κανόνες της κάθε γλώσσας, π.χ προτάσεις που έχουν λάθος σειρά λέξεων. Η εργασία αυτή παρουσιάζει μία νέα μέθοδο που έχει σαν σκοπό να ανιχνεύει και να διορθώνει προτάσεις με λάθη στην σειρά των λέξεων. Είναι εύκολα αντιληπτό ότι ένας τρόπος διόρθωσης προτάσεων είναι με λάθη στην σειρά των λέξεων είναι και η αντιμετάθεση όλων των λέξεων της πρότασης. Το ερώτημα που γεννάται είναι με ποιόν τρόπο θα γίνει αυτό την στιγμή που δεν ξέρουμε τι μέρος του λόγου είναι η κάθε λέξη και τι θέση πρέπει να πάρει μέσα στην πρόταση. Πολλές τεχνικές έχουν αναπτυχθεί για αυτόν το λόγο στο παρελθόν χρησιμοποιώντας γραμματικούς αναλυτές και κανόνες. Αντίθετα κάποιος θα μπορούσε να ισχυριστεί ότι μια πιθανή λύση αποτελεί η δημιουργία όλων των πιθανών αντιμεταθέσεων για μια πρόταση μήκους N. Η δυσκολία όμως που παρουσιάζει ένα τέτοιο εγχείρημα είναι το να διαχειριστεί τον μεγάλο όγκο των αντιμεταθέσεων. Ας σημειωθεί ότι για μια πρόταση μήκους N λέξεων, ο αριθμός των αντιμεταθέσεων είναι κατά πολύ μεγαλύτερος, $N!$. Ο αριθμός αυτός φαντάζει πολύ μεγάλος και απαγορευτικός για περαιτέρω επεξεργασία. Η καινοτομία της μεθόδου έγκειται στο γεγονός

ότι χρησιμοποιεί μια νέα μέθοδο γρήγορης αναζήτησης της βέλτιστης λύσης με σκοπό τον περιορισμό του αριθμού των αντιμεταθέσεων. Συνεπώς, για την διόρθωση προτάσεων με λάθη στην σειρά των λέξεων χρησιμοποιούνται τα παρακάτω εργαλεία, το στατιστικό γλωσσικό μοντέλο που παρέχει μια λίστα από διγράμματα και τριγράμματα και η μέθοδος γρήγορης αναζήτησης που βασίζεται στην δημιουργία του πίνακα αντιστοίχισης, όπως φαίνεται και στο Σχήμα 4.1.

Η επόμενη παράγραφος συνοψίζει τα κύρια αλγοριθμικά βήματα της μεθόδου για την ανίχνευση και διόρθωση προτάσεων με λάθη στην σειρά των λέξεων:

1. Δημιουργία του πίνακα αντιστοίχισης για την εξαγωγή των έγκυρων διγραμμάτων από την πρόταση $a = [w[1], w[2], ...w[N-1], w[N]]$ (πρόταση εισόδου).
2. Δημιουργία ενός δικτύου (Ν-επίπεδων και Ν-καταστάσεων) με έγκυρα διγράμματα με σκοπό τον σχηματισμό πιθανών αναδιατεταγμένων προτάσεων μήκους Ν (προτάσεις με διαφορετική σειρά λέξεων).
3. Έλεγχος της κάθε αναδιατεταγμένης πρότασης ως προς τον αριθμό των έγκυρων τριγραμμάτων.
4. Εξαγωγή των Ν-καλύτερων προτάσεων βάσει του αριθμού των έγκυρων τριγραμμάτων.
5. Σύγκριση του αριθμού των έγκυρων τριγραμμάτων της πρότασης που αξιολογείται ως η πρώτη καλύτερη από το σύστημα με τον αριθμό των έγκυρων τριγραμμάτων της πρότασης εισόδου.
6. Στην περίπτωση που η πρόταση εισόδου έχει μικρότερο αριθμό έγκυρων τριγραμμάτων από την προτεινόμενη πρόταση τότε το σύστημα ανιχνεύει λάθος στην σειρά των λέξεων της πρότασης εισόδου και προτείνει σαν καλύτερη εναλλακτική την πρόταση με τα περισσότερα έγκυρα τριγράμματα (πρώτη καλύτερη).

7. Στην λίστα των Ν-καλύτερων μπορεί να υπάρχουν παραπάνω από μια προτάσεις με τον ίδιο αριθμό έγκυρων τριγραμμάτων. Στην περίπτωση αυτή λαμβάνεται υπόψη το γινόμενο των πιθανοτήτων των τριγραμμάτων.



Σχήμα 4.1 Η αρχιτεκτονική του συστήματος.

4.3.2 Μέθοδος γρήγορης αναζήτησης της βέλτιστης λύσης

Ας υποθέσουμε ότι δίνεται μια ακολουθία λέξεων με τυχαία σειρά που μπορεί με την κατάλληλη αντιμετάθεση λέξεων να σχηματίσει μια γραμματικά και συντακτικά αποδεκτή πρόταση. Το ερώτημα που τίθεται είναι πως είναι δυνατόν να βρεθεί κάποιος αυτόματος τρόπος αναδιάταξης των λέξεων. Είναι εύκολο να ισχυριστεί κάποιος ότι δημιουργώντας όλες τις αντιμεταθέσεις των λέξεων μίας πρότασης (αναδιαταγμένες προτάσεις) μια από αυτές τις προτάσεις θα είναι η πρόταση με την αποδεκτή σειρά λέξεων (πρόταση με τις λέξεις στην σωστή σειρά). Το ζητούμενο είναι πόσο εφικτό είναι να διαχειριστούμε τον αριθμό των αναδιαταγμένων προτάσεων, ιδίως για προτάσεις με μεγάλο αριθμό λέξεων. Για αυτόν τον λόγο στην εργασία αυτή προτείνεται μια νέα μέθοδος γρήγορης αναζήτησης της βέλτιστης λύσης που εξυπηρετεί στο περιορισμό των $N!$ αντιμεταθέσεων και του χώρου αναζήτησης της καλύτερης πρότασης. Η μέθοδος γρήγορης αναζήτησης της βέλτιστης λύσης (διαδικασία μείωσης του αριθμού των αντιμεταθέσεων) περιλαμβάνει την δημιουργία ενός πίνακα αντιστοίχισης με μέγεθος $N \times N$ με σκοπό να εξαχθούν όλα τα πιθανά ζεύγη λέξεων σε μια πρόταση. Η δημιουργία του πίνακα αντιστοίχισης έχει σαν αφετηρία την ιδέα ότι όλες οι λέξεις σε μια πρόταση δεν μπορούν να συνδεθούν με όλες τις υπόλοιπες. Στην περίπτωση που γίνει δεκτή η παραδοχή ότι κάθε λέξη μπορεί να συνδεθεί μια άλλη λέξη της πρότασης τότε ο συνολικός αριθμός των ζευγών μια πρότασης είναι $N \times (N-1)$ και ο συνολικός αριθμός αντιμεταθέσεων για μια πρόταση είναι $N!$. Μια τέτοια υπόθεση υποδηλώνει ότι η κάθε λέξη μιας πρότασης μπορεί να βρίσκεται σε οποιαδήποτε θέση. Όμως, με την μέχρι τώρα εμπειρία μας από την χρήση της γλώσσας γίνεται αντιληπτό ότι μια τέτοια υπόθεση δεν ευσταθεί.

Για να γίνει περισσότερο κατανοητή η λειτουργία της διαδικασίας μείωσης του αριθμού των αντιμεταθέσεων μιας πρότασης με Ν λέξεις θα δώσουμε ένα παράδειγμα θεωρώντας ότι οι λέξεις μιας πρότασης αναπαρίστανται με αριθμούς από το 1 έως το 4 και με την ακόλουθη διάταξη (π.χ πρόταση Π: 1 2 3 4) (Σχήμα 4.2). Ο σκοπός του παραδείγματος αυτού είναι να καταδείξει τον ρόλο που έχει ένα δίγραμμα στον σχηματισμό των αντιμεταθέσεων της ακολουθίας 1 2 3 4. Όπως αναφέρθηκε και παραπάνω η όλη διαδικασία στηρίζεται στο γεγονός της χρήσης μόνο των διγραμμάτων που εμφανίζονται στην λίστα του γλωσσικού μοντέλου και όχι όλων των συνδυασμών ανά δυο των λέξεων μιας πρότασης. Στην περίπτωση που έχουμε την πρόταση 4 λέξεων, τότε ο αριθμών όλων των ζευγαριών λέξεων είναι 4*3 και ο αριθμός όλων των αντιμεταθέσεων είναι 4!. Ας υποθέσουμε ότι από την λίστα του γλωσσικού μοντέλου λείπουν τα ζεύγη (2 3) και (1 4). Το ερώτημα που τίθεται είναι πόσο περιορίζεται ο αριθμός των

αντιμεταθέσεων, απαλείφοντας τις αντιμεταθέσεις αυτές στις οποίες συμμετέχουν τα ζεύγη αυτά;

~~1) 1 2 3 4~~	~~13) 3 1 2 4~~
~~2) 1 2 4 3~~	14) 3 1 4 2
3) 1 3 2 4	15) 3 2 1 4
4) 1 3 4 2	16) 3 2 4 1
~~5) 1 4 2 3~~	17) 3 4 2 1
~~6) 1 4 3 2~~	~~18) 3 4 1 2~~
7) 2 1 3 4	~~19) 4 1 2 3~~
8) 2 1 4 3	20) 4 1 3 2
~~9) 2 3 4 1~~	21) 4 2 1 3
~~10) 2 3 1 4~~	~~22) 4 2 3 1~~
11) 2 4 3 1	23) 4 3 2 1
12) 2 4 1 3	24) 4 3 1 2

Σχήμα 4.2 Σχηματισμοί λέξεων με απάλειψη αυτών όπου συμμετέχουν μη εμφανιζόμενα διγράμματα όπως το (2 3) και (1 4), με βάση το γλωσσικό μοντέλο.

Σ' αυτήν την περίπτωση όπου στο γλωσσικό μοντέλο δεν συμπεριλαμβάνονται 2 ζεύγη όπως το (2 3) και το (1 4) μπορούμε να δούμε ότι απαλείφονται 9 σχηματισμοί από τους 24. Τώρα στην περίπτωση όπου δεν συμπεριλαμβάνονται στο γλωσσικό μοντέλο δυο διαφορετικά ζεύγη λέξεων όπως το (1 2) και το (2 3) θα δούμε ότι απαλείφονται διαφορετικοί σχηματισμοί και το σύνολο των σχηματισμών που δεν συμμετέχουν μεγαλώνει από τους 9 στους 11 (Σχήμα 4.3). Από αυτήν την διαπίστωση μπορούμε να συμπεράνουμε ότι στην διαδικασία αυτή παίζει σημαντικό ρόλο ο αριθμός των διγραμμάτων που δεν συμπεριλαμβάνονται στο γλωσσικό μοντέλο και ποια είναι αυτά. Μεταφερόμενοι στην διαδικασία του περιορισμού αντιμεταθέσεων μιας πρότασης λόγω απουσίας κάποιων διγραμμάτων από το γλωσσικό μοντέλο μπορούμε να επισημάνουμε ότι τα διγράμματα που δεν εμφανίζονται είναι ζεύγη λέξεων που κατά κύριο λόγο παραβιάζουν την σειρά των λέξεων και είναι απίθανο η λέξη 2 να έπεται της λέξης 1 (δηλ.

απουσιάζει το δίγραμμα (1 2)). Από όλα αυτά γίνεται κατανοητό ότι υπάρχει το ενδεχόμενο να έχουμε 2 προτάσεις με τον ίδιο αριθμό λέξεων και με τον ίδιο αριθμό μη έγκυρων διγράμματα που όμως να έχουν διαφορετικό αριθμό αντιμεταθέσεων.

~~1) 1 2 3 4~~	13) 3 1 2 4
~~2) 1 2 4 3~~	14) 3 1 4 2
3) 1 3 2 4	15) 3 2 1 4
4) 1 3 4 2	16) 3 2 4 1
~~5) 1 4 2 3~~	17) 3 4 2 1
~~6) 1 4 3 2~~	~~18) 3 4 1 2~~
7) 2 1 3 4	~~19) 4 1 2 3~~
8) 2 1 4 3	20) 4 1 3 2
9) 2 3 4 1	21) 4 2 1 3
~~10) 2 3 1 4~~	~~22) 4 2 3 1~~
11) 2 4 3 1	23) 4 3 2 1
12) 2 4 1 3	~~24) 4 3 1 2~~

Σχήμα 4.3 Σχηματισμοί λέξεων με απάλειψη αυτών όπου συμμετέχουν μη εμφανιζόμενα διγράμματα όπως το (2 3) και (1 2), με βάση το γλωσσικό μοντέλο.

Στην επόμενη παράγραφο θα παρουσιασθεί ο τρόπος υπολογισμού του αριθμού των έγκυρων διγραμμάτων μιας πρότασης (πίνακας αντιστοίχισης) και ο τρόπος αντιμετάθεσης των λέξεων σε μια πρόταση βάσει των έγκυρων διγραμμάτων με σκοπό την μείωση του αριθμού των αντιμεταθέσεων.

Δίνοντας μια πρόταση $a = [w[1], w[2], ...w[N-1], w[N]]$ με N λέξεις, μπορεί να σχηματιστεί ένας πίνακας αντιστοίχισης $A \in R^{NXN}$, βλέπε Πίνακας 4.1,

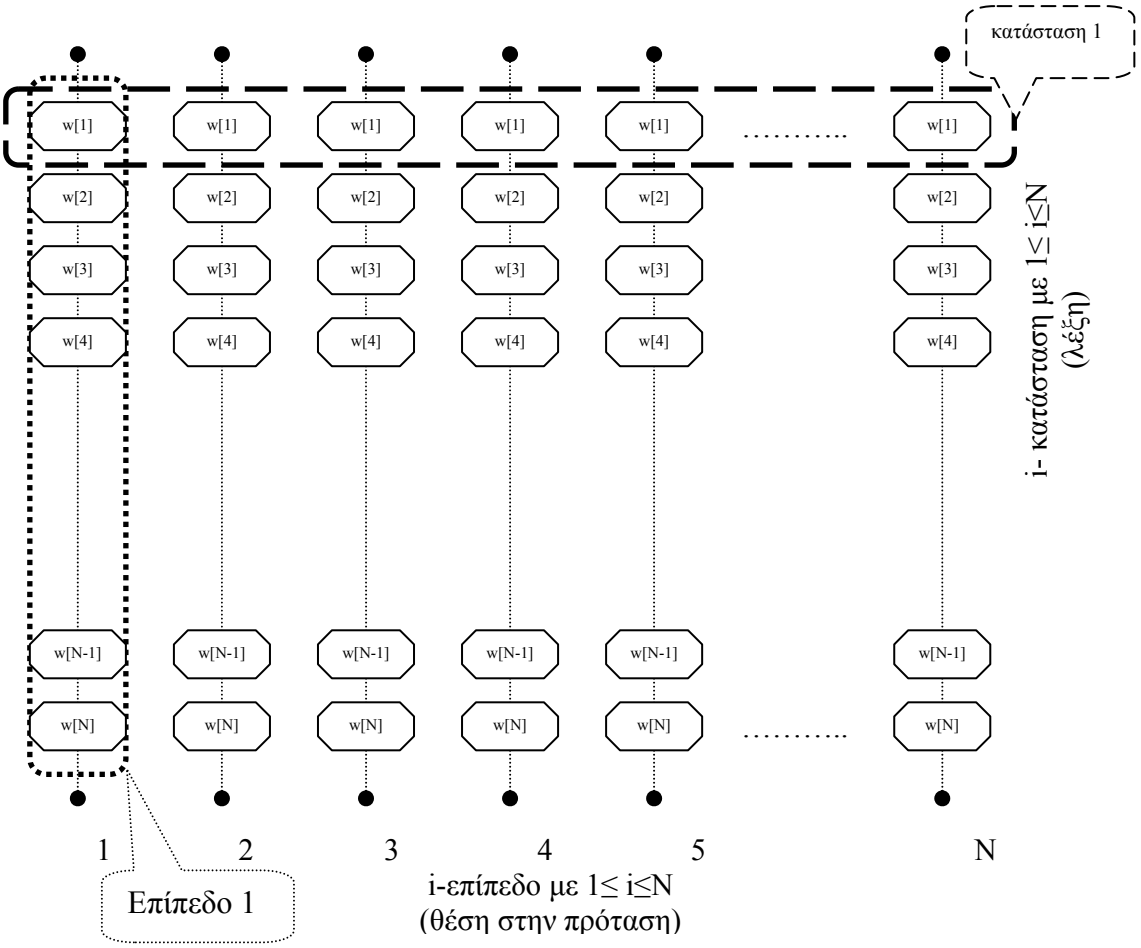
λέξη	**w[1]**	**w[2]**	**w[3]**			**w[N-1]**	**w[N]**
w[1]	■	P[1,2]	P[1,3]			P[1,N-1]	P[1,N]
w[2]	P[2,1]	■	P[2,3]			P[2,N-1]	P[2,N]
w[3]	P[3,1]	P[3,2]	■			P[3,N-1]	P[3,N]
w[N-1]	P[N-1,1]	P[N-1,2]	P[N-1,3]			■	P[N-1,N]
w[N]	P[N,1]	P[N,2]	P[N,3]			P[N,N-1]	■

Πίνακας 4.1 Δημιουργία του πίνακα αντιστοίχισης NXN, δίνοντας την πρόταση α=[w[1],w[2],w[3],…,w[N-2],w[N-1],w[N]]. Ο πίνακας δείχνει τον τρόπο σύνδεσης των λέξεων ώστε να δημιουργηθούν ζεύγη λέξεων και τον έλεγχο εγκυρότητας τους με την χρήση του γλωσσικού μοντέλου.

Το μέγεθος του πίνακα εξαρτάται από το μήκος της πρότασης. Ο σκοπός της δημιουργίας του πίνακα αντιστοίχισης είναι να εξάγουμε τα έγκυρα διγράμματα με την χρήση του γλωσσικού μοντέλου. Το στοιχείο του πίνακα $P[i,j]$ υποδηλώνει την εμφάνιση ή όχι του κάθε ζεύγους από λέξεις $(w[i]w[j])$ βάσει της λίστας με τα διγράμματα που περιέχονται στο γλωσσικό μοντέλο. Εάν ένα ζεύγος από 2 λέξεις $(w[i]w[j])$ ανήκει στην λίστα των διγραμμάτων, τότε το $P[i,j]$ λαμβάνεται ίσο με το ένα αλλιώς με μηδέν. Στο εξής διγράμματα που έχουν $P[i,j]$=1, θα καλούνται έγκυρα διγράμματα. Παρατηρώντας τα πειραματικά αποτελέσματα χρήσης της μεθόδου θα αποδειχθεί ότι ο αριθμός των έγκυρων διγραμμάτων M είναι κατά πολύ μικρότερος από το σύνολο των στοιχείων του πίνακα αντιστοίχισης που ισούται με $N\times(N-1)$, από την στιγμή που όλα τα πιθανά ζεύγη λέξεων δεν είναι έγκυρα βάσει του γλωσσικού μοντέλου (βλέπε Ενότητα 4.4 & 4.5).

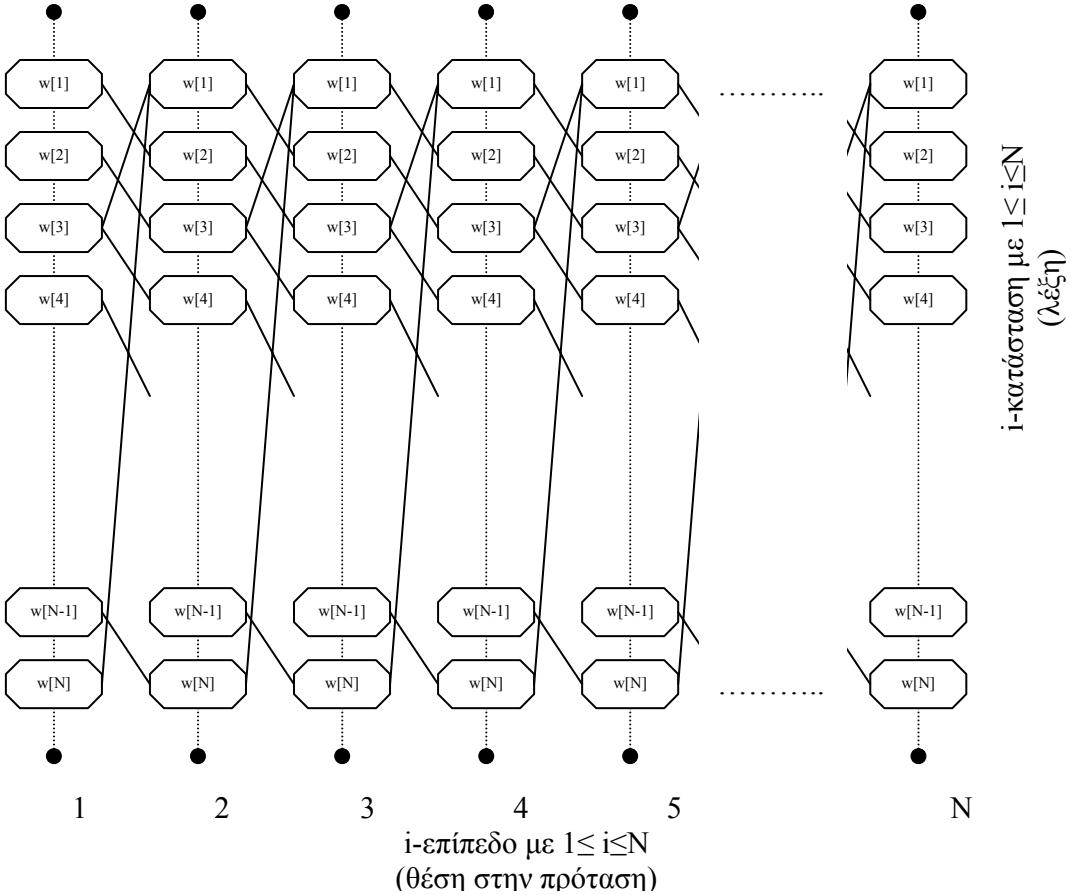
Το ερώτημα είναι πώς τα έγκυρα διγράμματα θα συνδυαστούν για να σχηματίσουν τις πιθανές αναδιατεταγμένες προτάσεις (αντιμεταθέσεις). Όπως γίνεται κατανοητό ένα τέτοιο πρόβλημα αποτελεί πρόβλημα αναζήτησης (search problem) και η επίλυση του αφορά την μέθοδο γρήγορης αναζήτησης της βέλτιστης λύσης. Στην περίπτωση που εξετάζουμε, για να λυθεί το πρόβλημα αναζήτησης θα χρησιμοποιηθεί μια μέθοδος επίλυσης από τα αριστερά προς τα δεξιά. Η μέθοδος αυτή θα πρέπει να έχει φορά από αριστερά προς δεξιά γιατί κάθε πρόταση αναπτύσσεται από αριστερά προς δεξιά. Για να γίνει αντιληπτή η διαδικασία του φιλτραρίσματος των αντιμεταθέσεων, ας φανταστούμε ένα δίκτυο με N επίπεδα και N καταστάσεις (Σχήμα 4.4). Ο παράγοντας N σχετίζεται με τον αριθμό των λέξεων της

πρότασης. Κάθε επίπεδο αντιστοιχεί σε μία θέση λέξης μέσα στην πρόταση και κάθε λέξη αντιστοιχεί σε μια κατάσταση. Όλες οι καταστάσεις από το επίπεδο 1 συνδέονται με όλες τις πιθανές καταστάσεις στο δεύτερο επίπεδο και κάθε κατάσταση στο δεύτερο επίπεδο συνδεέται με όλες τις πιθανές καταστάσεις στο τρίτο επίπεδο. Αυτό συνεχίζεται μέχρι να φθάσουμε στο τελικό επίπεδο που δεν είναι άλλο από το σύνολο των λέξεων. Η σύνδεση μιας κατάστασης στο ένα επίπεδο με μια άλλη κατάσταση σε άλλο επίπεδο γίνεται βάσει των έγκυρων διγραμμάτων. Άρα η σύνδεση ανάμεσα σε 2 καταστάσεις (i, j) γειτνιαζόντων επιπέδων $(N-1, N)$ υφίσταται μόνο και μόνο όταν το δίγραμμα $(w[i]w[j])$ είναι έγκυρο. Το δίκτυο λέξεων που δημιουργείται αποδίδει παραστατικά τον αλγόριθμο παραγωγής των πιθανών προτάσεων (Σχήμα 4.5). Ξεκινώντας από οποιοδήποτε κατάσταση στο επίπεδο 1 και κινούμενος προς τα δεξιά μέσα από όλες τις έγκυρες συνδέσεις καταλήγεις στο N-th επίπεδο του δικτύου, σχηματίζοντας με αυτό τον τρόπο όλες τις πιθανές αναδιαταγμένες προτάσεις (Σχήμα 4.6).



Σχήμα 4.4 Αναπαράσταση του δικτύου με N-επίπεδα και N-καταστάσεις. Το i-επίπεδο αναφέρεται στην θέση i στην πρόταση, και η i-κατάσταση στην λέξη i, με 1≤ i≤N.

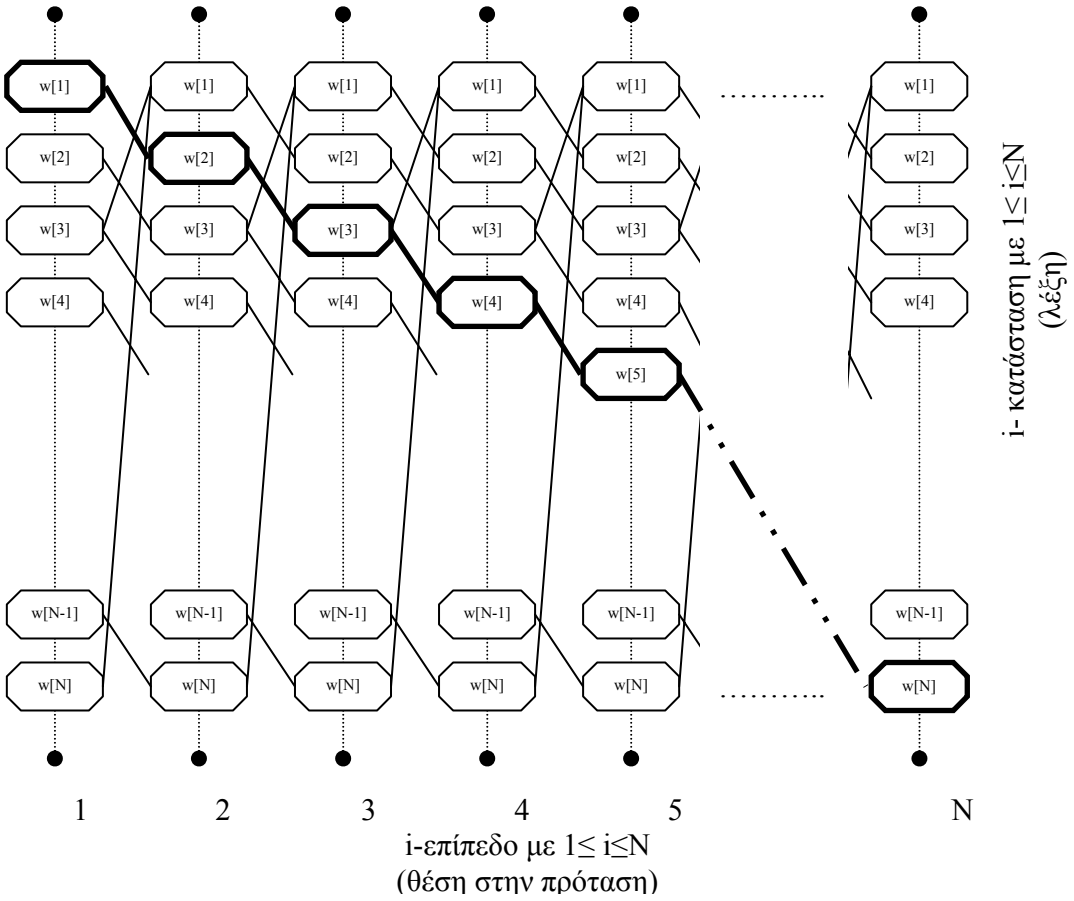
Για περαιτέρω μείωση του αριθμού των αντιμεταθέσεων είναι σκόπιμο να χρησιμοποιήσουμε τις πιθανότητες εμφάνισης των μονογραμμάτων στο πρώτο και στο τελευταίο επίπεδο. Η πιθανότητα του μονογράμματος υπολογίζεται βάσει την συχνότητα εμφάνισης του κάθε μονογράμματος στην πρώτη και στην τελευταία θέση των προτάσεων που συγκαταλέγονται στο σώμα κειμένου εκπαίδευσης. Είναι φανερό ότι κάποιες λέξεις της εκάστοτε πρότασης είναι αδύνατον να τοποθετηθούν στο πρώτο και στο τελευταίο επίπεδο του δικτύου αυτού, από την στιγμή που ουδέποτε τοποθετούνται στην αρχή ή στο τέλος μιας πρότασης. Άρα, κατά μήκος του δικτύου, οι καταστάσεις που ανήκουν στο πρώτο και στο τελευταίο επίπεδο και αντιστοιχούν σε λέξεις με πολύ μικρή πιθανότητα εμφάνισης στην τρέχουσα θέση, δεν θα υπολογίζονται.



Σχήμα 4.5 Αναπαράσταση του δικτύου με N-επίπεδα και N-καταστάσεις. Οι ενώσεις των καταστάσεων γίνεται βάσει των έγκυρων διγραμμάτων του γλωσσικού μοντέλου.

Το πλεονέκτημα της μεθόδου γρήγορης αναζήτησης έγκειται στην αποκλειστική χρήση των έγκυρων διγραμμάτων και όχι στο σύνολο όλων των πιθανών διγραμμάτων. Ο αριθμός των αντιμεταθέσεων της πρότασης εξαρτάται από τον αριθμό των έγκυρων διγραμμάτων. Ας αναλογιστούμε ότι για μία πρόταση με 7 λέξεις ο συνολικός αριθμός των διγραμμάτων είναι

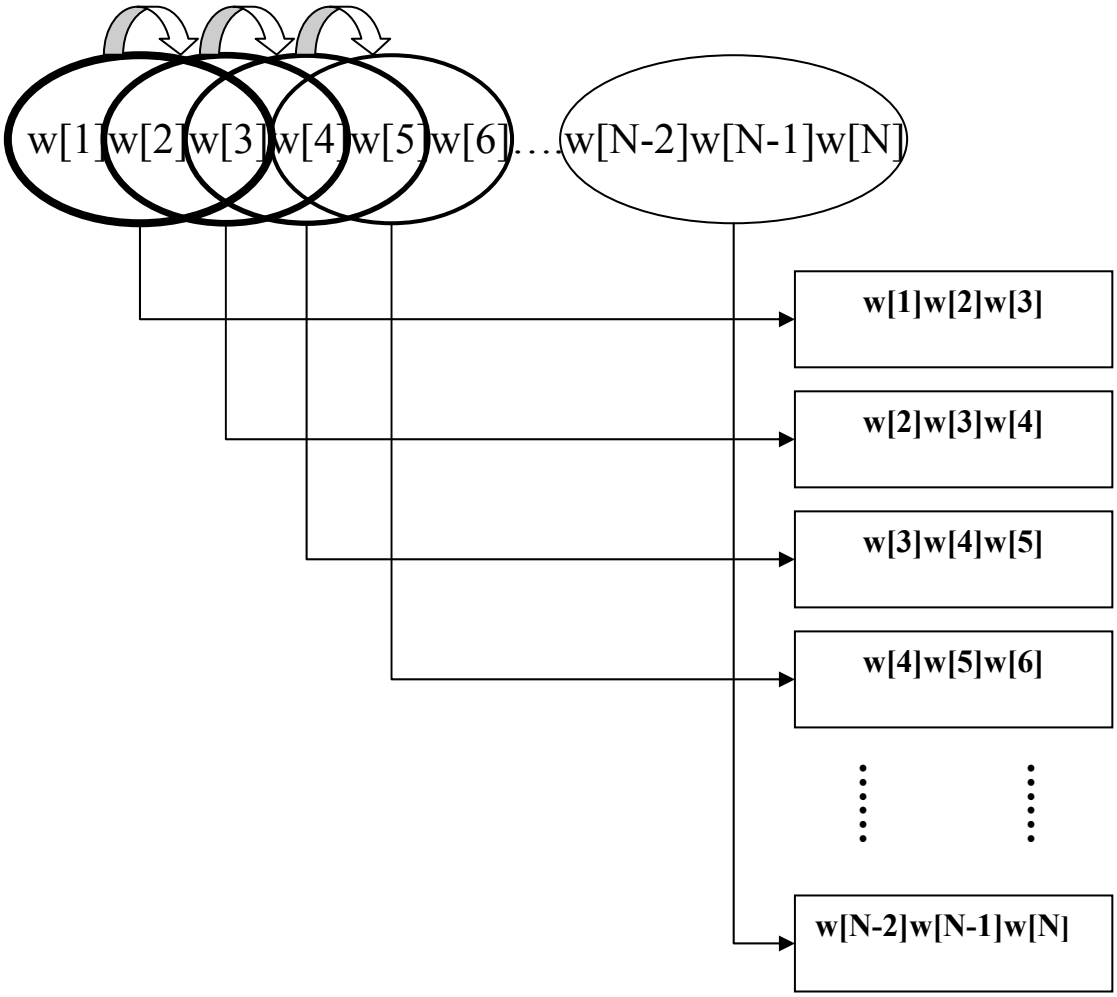
7*6=42 αλλά με την χρήση του γλωσσικού μοντέλου ο αντίστοιχος αριθμός είναι πολύ μικρότερος. Συνεπώς, συνδυάζοντας μόνο τα έγκυρα και όχι τα πιθανά διγράμματα, ο αριθμός των αντιμεταθέσεων μιας πρότασης N λέξεων μειώνεται σημαντικά. Το πόσο μειώνεται ο αριθμός των αντιμεταθέσεων αποδεικνύεται πειραματικά με την χρήση διαφορετικών δεδομενων τόσο στην Αγγλική όσο και στην Ελληνική γλώσσα (βλέπε Ενότητα 4.4 & 4.5).



Σχήμα 4.6 Η δημιουργία των αναδιαταγμένων προτάσεων με την χρήση του δικτύου (N-επιπέδων και N-καταστάσεων) βάσει των έγκυρων διγραμμάτων. Με το έντονο μαύρο χρώμα αποτυπώνεται μια πιθανή διαδρομή από την αρχική λέξη w[1] στην τελική w[N] μέσω των επιτρεπτών κόμβων-λέξεων.

4.3.3 Μέθοδος αξιολόγησης προτάσεων βάσει των N-grams

Η βασική λειτουργία της μεθόδου είναι ο τεμαχισμός της εκάστοτε πρότασης εισόδου σε ένα σύνολο τριγραμμάτων. Αυτό επιτυγχάνεται επιλέγοντας το κατάλληλο παράθυρο επεξεργασίας. Στην περίπτωση που θέλουμε να εξάγουμε τριγράμματα από μία πρόταση χρησιμοποιείται ένα παράθυρο 3 λέξεων με αλληλοκάλυψη 2 λέξεων. Το παράθυρο ξεκινάει από την πρώτη λέξη και καταλήγει στην τρίτη από το τέλος (Σχήμα 4.7).



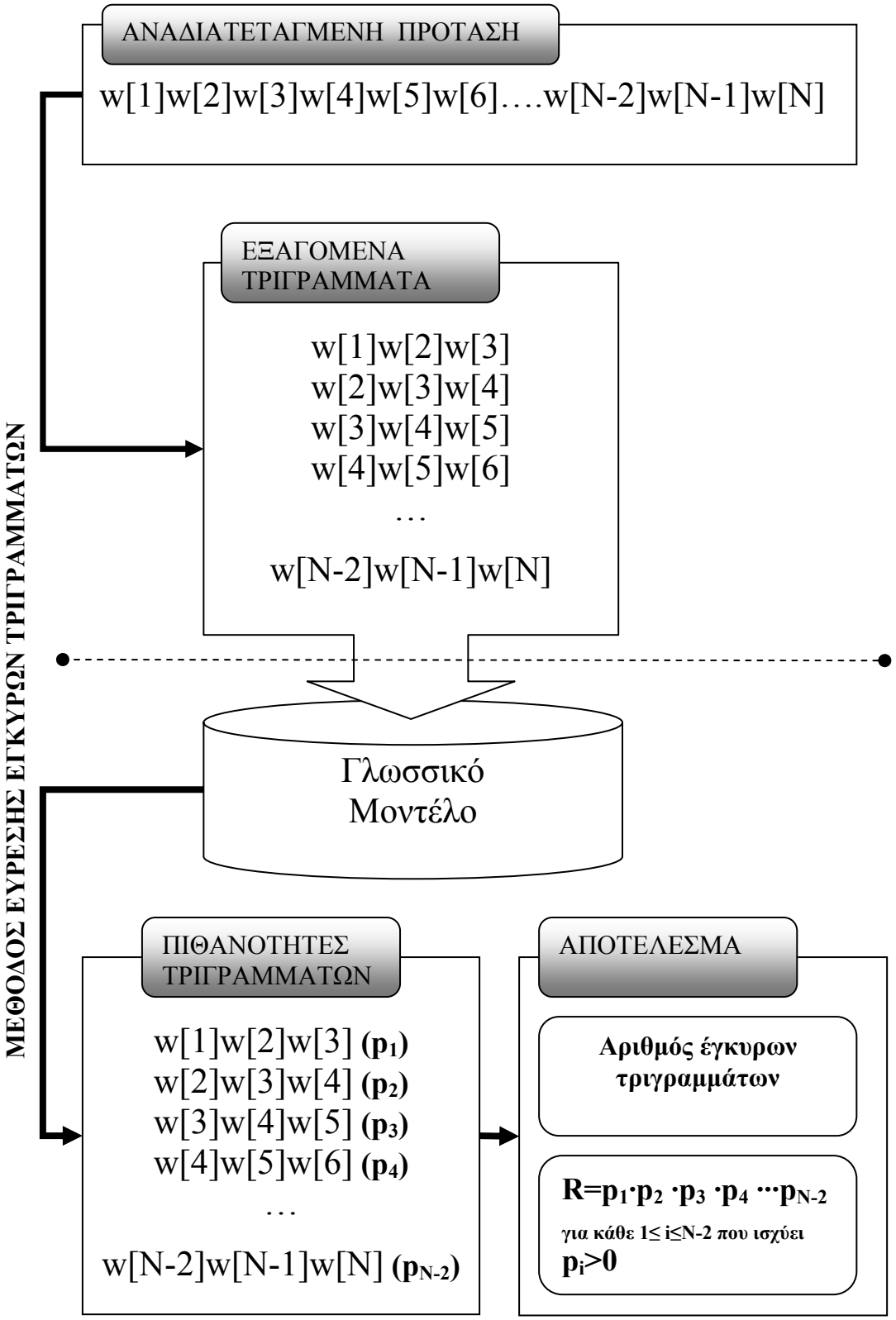
Σχήμα 4.7 Ο τρόπος εξαγωγής τριγραμμάτων από μια πρόταση με την βοήθεια του παράθυρου ολίσθησης.

Έτσι, σε μία πρόταση N λέξεων περιλαμβάνονται $N-2$ τριγράμματα. Είναι φανερό ότι στην προσπάθεια μας να αναζητήσουμε στοιχειώδεις μονάδες μέσα σε μια πρόταση η καλύτερη επιλογή είναι αυτή των τριγραμμάτων πρώτον, γιατί στην βάση δεδομένων του γλωσσικού μοντέλου υπάρχουν μέχρι τριγράμματα και δεύτερον, γιατί με τα υπάρχοντα σώματα κειμένων, τα τριγράμματα είναι περισσότερο ασφαλή για την αναπαράσταση της αλληλοεξάρτησης λέξεων με μικρή απόσταση. Όπως γίνεται κατανοητό θα ήταν χρήσιμο και πιο αποδοτικό για την μέθοδο αυτή να χρησιμοποιούνταν N-grams με N≥3. Θα πρέπει να σημειωθεί όμως ότι η χρήση του αριθμού N εξαρτάται άμεσα από την ποικιλία και το μέγεθος του σώματος εκπαίδευσης του γλωσσικού μοντέλου. Με την μέχρι τώρα εμπειρία πάνω στην εκπαίδευση στατιστικών γλωσσικών μοντέλων βάσει του μεγέθους των σωμάτων εκπαίδευσης μπορεί να ειπωθεί ότι μοντέλα σαν και τα τριγράμματα είναι τα πιο κατάλληλα για να

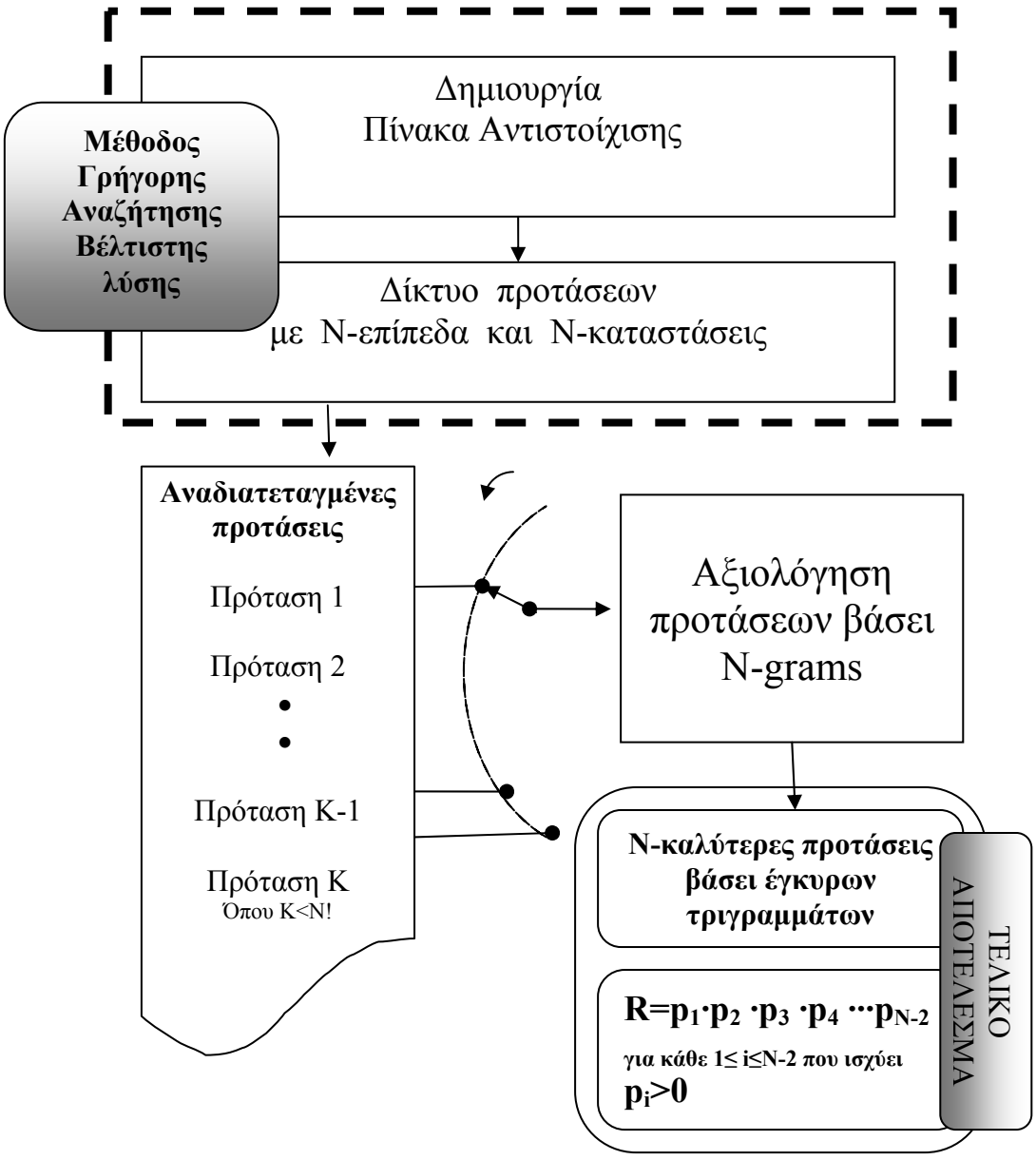
κωδικοποιήσουν τις κοντινές σχέσεις των λέξεων σε μια πρόταση. Η αδυναμία χρήσης μεγαλύτερης τάξης μοντέλων μας οδηγεί στην απόφαση να χρησιμοποιηθεί μια αλληλο-κάλυψη μεταξύ 2 τριγραμμάτων με κοινή περιοχή τις δυο τελευταίες λέξεις του πρώτου τριγράμματος και τις δυο πρώτες του δευτέρου τριγράμματος.

Το δεύτερο βήμα της μεθόδου περιλαμβάνει την εύρεση των έγκυρων τριγραμμάτων για κάθε πρόταση. Κάθε έγκυρο τρίγραμμα συνοδεύεται από μία πιθανότητα που προέρχεται από την συχνότητα εμφάνισης του τριγράμματος μέσα στο ίδιο το σώμα κειμένου. Στο τρίτο βήμα αυτής της μεθόδου υπολογίζεται ο αριθμός των έγκυρων τριγραμμάτων σε κάθε αναδιαταγμένη πρόταση (Σχήμα 4.8). Στην περίπτωση που η πρόταση εισόδου έχει μικρότερο αριθμό τριγραμμάτων από μια άλλη οποιαδήποτε πρόταση, είναι κατανοητό ότι η πρόταση εισόδου είναι υποψήφια για λάθος στην σειρά των λέξεων.

Σε αυτό το σημείο πρέπει να τονιστεί ότι η μέθοδος αυτή έχει να αντιμετωπίσει την σπανιότητα κάποιων δεδομένων. Όπως είναι φυσικό, κάθε σώμα κειμένου δεν μπορεί στις τάξεις του να περιλάβει όλες τις λέξεις και τους συνδυασμούς αυτών, με συνέπεια μερικά τριγράμματα να είναι έγκυρα αλλά να μην περιλαμβάνονται στην λίστα των τριγραμμάτων του γλωσσικού μοντέλου. Όπως αναφέραμε και στο δεύτερο κεφάλαιο όπου γίνεται η επισκόπηση στις κυριότερες μεθόδους στατιστικών γλωσσικών μοντέλων, η κύρια προσπάθεια έγκειται στο γεγονός της εξομάλυνσης των πιθανοτήτων ώστε να μεταφερθεί μάζα πιθανοτήτων από τα εμφανιζόμενα N-grams στα μη εμφανιζόμενα. Έτσι κατά κάποιον τρόπο αντιμετωπίζεται το πρόβλημα της έλλειψης κάποιων καθόλα έγκυρων τριγραμμάτων. Παρόλαυτα το φαινόμενο αυτό αποτελεί την σημαντικότερη τροχοπέδη της μεθόδου αυτής έστω και αν στο μέλλον η συνεχής διεύρυνση των σωμάτων κειμένων θα μειώσει το πρόβλημα. Το κριτήριο με το οποίο αξιολογούνται όλες οι αναδιαταγμένες προτάσεις, είναι ο αριθμός των έγκυρων τριγραμμάτων. Το σύστημα παρέχει ως έξοδο τις Ν-καλύτερες προτάσεις με το μεγαλύτερο αριθμό έγκυρων τριγραμμάτων (Σχήμα 4.9). Στην περίπτωση που παραπάνω από μια προτάσεις έχουν τον ίδιο αριθμό έγκυρων τριγραμμάτων τότε ορίζεται μια νέα μετρική απόστασης. Η μετρική αυτή, ορίζεται με την βοήθεια του γινομένου των πιθανοτήτων των έγκυρων τριγραμμάτων.



Σχήμα 4.8 Ο αλγόριθμος 2 φάσεων που υλοποιείται για την εύρεση έγκυρων τριγραμμάτων από το γλωσσικό μοντέλο. Στην πρώτη φάση όλες οι προτάσεις διασπώνται σε τριγράμματα. Στην δεύτερη φάση όλα τα εξαγώμενα τριγράμματα αναζητούνται στο γλωσσικό μοντέλο. Το αποτέλεσμα αυτής της διαδικασίας είναι μια λίστα από έγκυρα τριγράμματα με τις πιθανότητες τους (p_i).



Σχήμα 4.9 Η αρχιτεκτονική του υποσυστήματος αξιολόγησης αναδιατεταγμένων προτάσεων βάσει του αριθμού έγκυρων τριγραμμάτων σύμφωνα με το γλωσσικό μοντέλο.

Παρακάτω θα δούμε ένα παράδειγμα χρήσης της μεθόδου με μια πρόταση (“I have also campaigned for the government to give AIDS greater recognition”) που επιλέχθηκε τυχαία από το σώμα κειμένου WSJ. Η πρόταση αυτή επιλέχθηκε έτσι ώστε όλα τα τριγράμματα να ανήκουν στην λίστα του γλωσσικού μοντέλου.

Ο στόχος αυτής της διαδικασίας είναι να επιβεβαιώσει ότι η πρόταση εισόδου έχει καλύτερο αποτέλεσμα συγκρινόμενη με όλες τις υπόλοιπες αναδιατεταγμένες προτάσεις (απόρροια της διαδικασίας του φιλτραρίσματος). Οι προτάσεις που παρουσιάζονται στο πίνακα (Πίνακας 4.2) είναι οι πρώτες καλύτερες σύμφωνα με τον αριθμό των έγκυρων τριγραμμάτων. Τα ευρήματα από την πειραματική διαδικασία που ακολουθήσαμε αποδεικνύουν ότι όλες οι προτάσεις έχουν μικρότερο αριθμό έγκυρων τριγραμμάτων, σε σχέση με την πρόταση εισόδου.

	Πρόταση S1
1	**I have also campaigned for the government to give AIDS greater recognition**
2	I also have campaigned for the government to give AIDS greater recognition
3	I have also campaigned for the government to give greater recognition AIDS
4	I have also campaigned to give AIDS greater recognition for the government

Πίνακας 4.2 Ο πίνακας δείχνει τις προτάσεις βάσει των αποτελεσμάτων τους σε φθίνουσα σειρά σύμφωνα με των αριθμό των έγκυρων τριγραμμάτων. Η τονισμένη πρόταση είναι η πρόταση εισόδου και όλες οι υπόλοιπες προτάσεις έχουν εξαχθεί από την διαδικασία του φιλτραρίσματος των αντιμεταθέσεων.

Στον Πίνακας 4.3 αποτυπώνονται τα έγκυρα διγράμματα που έχουν εξαχθεί με την χρήση του πίνακα αντιστοίχισης για την πρόταση εισόδου. Τα διγράμματα αυτά θα χρησιμοποιηθούν για τον σχηματισμό των πιθανών αντιμεταθέσεων των προτάσεων. Αναλογιζόμενοι ότι ο αριθμός των λέξεων μιας πρότασης είναι 12, ο συνολικός αριθμός των διγραμμάτων είναι 12*11, (αυτό ισχύει σε περίπτωση που δεχθούμε ότι όλες οι λέξεις ενώνονται με όλες τις άλλες), ενώ στην περίπτωση που εξετάζουμε ο αριθμός είναι μόλις 61 με την χρήση της διαδικασίας του φιλτραρίσματος.

Ο αριθμός των αναδιαταγμένων προτάσεων που απορρέει από την διαδικασία του φιλτραρίσματος των αντιμεταθέσεων είναι 245,519 ενώ χωρίς την διαδικασία αυτή είναι 479,001,600. Ο Πίνακας 4.3 παρακάτω αποτυπώνει τα έγκυρα διγράμματα. Παρατηρώντας το πίνακα γίνεται αντιληπτό ότι τα στοιχεία της κυρίας διαγώνιου του πίνακα δεν λαμβάνονται υπόψη.

	I	**have**	**also**	**campaigned**	**for**	**the**	**government**	**to**	**give**	**AIDS**	**greater**	**recognition**
I		■	■									
have	■		■				■	■		■		■
also	■	■					■					
campaigned		■	■									
for		■	■	■			■		■	■	■	■
the		■	■		■			■	■	■	■	■
government					■	■		■			■	
to		■	■	■			■		■	■	■	■
give	■		■					■		■		
AIDS		■	■		■	■			■			
greater		■	■		■	■		■	■	■		
recognition					■	■			■		■	

Πίνακας 4.3 Ο πίνακας αντιστοίχισης δείχνει τα έγκυρα διγράμματα για την πρόταση εισόδου "*I have also campaigned for the government to give AIDS greater recognition*" με την χρήση του πίνακα αντιστοίχισης. Το σύμβολο (■) αναφέρεται σε ζεύγη λέξεων που είναι έγκυρα (έγκυρα διγράμματα) βάσει του γλωσσικού μοντέλου. Για το συγκεκριμένο παράδειγμα, ο αριθμός των έγκυρων διγραμμάτων ισούται με 61.

4.4 Πειραματικά αποτελέσματα της μεθόδου για την Αγγλική γλώσσα

Η επόμενη ενότητα περιγράφει τα αποτελέσματα για δυο διαφορετικά πειραματικά δεδομένα με ή χωρίς την χρήση ενός περιορισμού μετακίνησης κάθε λέξης. Ο περιορισμός αυτός επιβάλει σε κάθε λέξη να κινείται μέσα σε συγκεκριμένα όρια. Στην περίπτωση του TOEFL η εκάστοτε λέξη της πρότασης εισόδου δεν μπορεί να μετακινηθεί σε άλλη θέση που απέχει απόσταση μεγαλύτερη των 3 λέξεων. Στην περίπτωση των πειραματικών δεδομένων για την Ελληνική και Αγγλική γλώσσα εξετάζονται διαφορετικές τιμές για το ϕ.

4.4.1 Πειραματικά δεδομένα του TOEFL

Η πειραματική διαδικασία περιλαμβάνει ένα σύνολο 400 προτάσεων με μήκος από 6 ως 12 λέξεις. Οι προτάσεις αυτές έχουν επιλεγεί τυχαία από την ενότητα «Δομή» παλαιοτέρων εξετάσεων του TOEFL (Folse, 1997; Feyton, 2002). Οι εξετάσεις του TOEFL αναφέρονται σε εξετάσεις στην Αγγλική γλώσσα ως αυτή να είναι ξένη γλώσσα. Το πρόγραμμα TOEFL είναι ενδεδειγμένο για την αξιολόγηση της ικανότητας ενός μη-γηγενή ομιλητή να διαβάσει, να

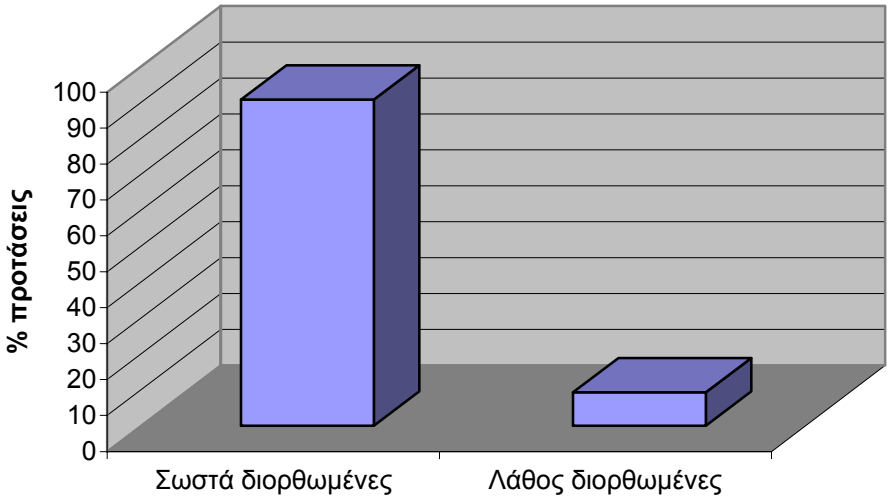
γράψει και να αντιληφθεί την Αγγλική γλώσσα όπως αυτή χρησιμοποιείται σε κολλέγια και πανεπιστήμια της Βόρειας Αμερικής. Η ενότητα «Δομή» επικεντρώνεται στην αναγνώριση λεξιλογίου, γραμματικής, και ορθής χρήσης του γραπτού λόγου στα Αγγλικά. Στην ενότητα αυτή υπάρχουν 2 κατηγορίες από ερωτήσεις. Η μια κατηγορία περιλαμβάνει προτάσεις με κάποιο κενό προς συμπλήρωση. Ο εξεταζόμενος πρέπει να συμπληρώσει το κενό διαλέγοντας από τις 4 υποψηφιότητες την σωστή λέξη ή φράση. Η άλλη κατηγορία ερωτήσεων περιλαμβάνει προτάσεις με 4 υπογραμμισμένες λέξεις. Ο εξεταζόμενος πρέπει να διαλέξει ποια από τις 4 λέξεις περιλαμβάνει λάθος στην γραμματική. Στην περίπτωση μας, το σύνολο των προτάσεων που θα εξετασθούν αναφέρονται στην ενότητα του TOEFL όπου εξετάζεται η σειρά των λέξεων. Ας σημειωθεί ότι οι προτάσεις αυτές δεν ανήκουν στο σύνολο των προτάσεων που χρησιμοποιήθηκε για την εκπαίδευση του γλωσσικού μοντέλου, και οι προτάσεις δεν περιλαμβάνουν όρους όπως “and” και “or”. Ο στόχος της πειραματικής διαδικασίας είναι να επιβεβαιώσει ότι η έξοδος του συστήματος (η πρόταση με το καλύτερο σκορ) είναι και η σωστή απάντηση που προτείνει το TOEFL. Για πειραματικούς λόγους, ο περιορισμός αντιμεταθέσεων τίθεται ίσος με τρία, δηλ. $\phi = 3$. Το κριτήριο επιλογής αυτής της τιμής για το ϕ, είναι ότι οι περισσότερες απαντήσεις στο TOEFL έχουν μήκος 4 λέξεις. Ο περιορισμός στην μετακίνηση των λέξεων κατά ϕ, δηλώνει ότι η κάθε λέξη μπορεί να μετακινηθεί σε κατεύθυνση προς τα εμπρός και όπισθεν κατά 3 λέξεις. Συνεπώς με την χρήση ενός τέτοιου περιορισμού οι αντιμεταθέσεις που προκύπτουν και δεν τηρούν την συνθήκη του περιορισμού μετακίνησης των λέξεων εξαιρούνται από το σύστημα και δεν βαθμολογούνται. Το Σχήμα 4.10 αποτελεί ένα παράδειγμα άσκησης στο TOEFL.

ΠΑΡΑΔΕΙΓΜΑ 1
------- explores the nature of guilt and responsibility and builds to a remarkable conclusion. A. The written beautifully novel B. The beautifully written novel C. The novel beautifully written D. The written novel beautifully

Σχήμα 4.10 Ένα παράδειγμα άσκησης στο TOEFL.

4.4.2 Τα πειραματικά αποτελέσματα του TOEFL

Η αξιολόγηση της μεθόδου έγινε με την σύγκριση της εξόδου του συστήματος με την σωστή πρόταση που προτείνεται από το TOEFL. Τα ευρήματα από την πειραματική διαδικασία δείχνουν ότι 363 προτάσεις (90,75% επί του συνόλου) επελέγησαν σωστά με την χρήση της προτεινόμενης μεθόδου (Σωστά διορθωμένες). Στον αντίποδα, για 37 προτάσεις τα αποτελέσματα ήταν λάθος (9,25% επί του συνόλου) (Λάθος διορθωμένες). Στην περίπτωση όπου είχαμε «Λάθος διόρθωση» το σύστημα έδωσε αποτελέσματα διαφορετικά από την σωστή πρόταση του TOEFL.



Σχήμα 4.11 Τα ποσοστό των σωστών και λανθασμένων διορθώσεων.

Είναι προφανές ότι η απόδοση του συστήματος για την διάγνωση και διόρθωση προτάσεων με λάθη στην σειρά των λέξεων εξαρτάται κυρίως από την ποιότητα του σώματος κειμένου. Η υψηλή αξιοπιστία του συστήματος επιτυγχάνεται λόγω των γραμματικά και συντακτικά σωστών προτάσεων των δεδομένων εκπαίδευσης.

4.4.3 Αξιολόγηση της μεθόδου γρήγορης αναζήτησης για την Αγγλική γλώσσα

Η ενότητα αυτή περιγράφει τα πλεονεκτήματα χρήσης της μεθόδου γρήγορης αναζήτησης για ένα σύνολο προτάσεων με διαφορετικό μήκος λέξεων. Όπως φαίνεται και στον Πίνακας 4.3, ο σκοπός της γρήγορης αναζήτησης (μέθοδος φιλτραρίσματος) βασίζεται στην λογική που δεν υπολογίζει όλους τους συνδυασμούς των λέξεων μέσα σε μια πρόταση. Άλλωστε αυτό αποτελεί και μια πραγματικότητα αναλογιζόμενοι ότι στην καθημερινή χρήση του γραπτού και προφορικού λόγου συνήθως ξέρουμε ή φανταζόμαστε ποιες λέξεις έπονται ή προηγούνται

άλλων. Για να μπορέσουμε όμως να απορρίψουμε τον συνδυασμό κάποιων λέξεων ανά δύο πρέπει να έχουμε κάποια απόδειξη μη εμφάνισης όμοιου διγράμματος στην γλώσσα. Έτσι με την βοήθεια του γλωσσικού μοντέλου και ιδιαιτέρως της βάσης των διγραμμάτων μπορούμε χρησιμοποιώντας τις πιθανότητες τους να δεχθούμε κάποια διγράμματα σαν έγκυρα και άλλα σαν άκυρα. Η διαστρωμάτωση των πιθανοτήτων που παρουσιάζουν τα διγράμματα της βάσης αποτελεί σαφή ένδειξη ότι κάποια διγράμματα είναι περισσότερο πιθανά από κάποια άλλα. Ας υποθέσουμε λοιπόν ότι έχουμε την παρακάτω πρόταση τυχαία, που είναι πρόταση χωρίς λάθος στην σειρά των λέξεων, «there is no limit to the number of ways to raise money». Φτιάχνοντας αντίστοιχο πίνακα αντιστοίχισης μπορούμε να ελέγξουμε την εγκυρότητα όλων των διγραμμάτων. Στο σημείο αυτό πρέπει να τονιστεί ότι με τον όρο εγκυρότητα ενός διγράμματος αναφερόμαστε στην εμφάνιση του ή όχι στην βάση των διγραμμάτων του γλωσσικού μοντέλου. Βάσει του γλωσσικού μοντέλου όλα τα διγράμματα του άξονα της πρότασης είναι έγκυρα. Διγράμματα του άξονα είναι τα διγράμματα της πρότασης που απορρέουν από την ολίσθηση ενός παραθύρου 2 λέξεων με επικάλυψη μιας ξεκινώντας από την αρχή και καταλήγοντας στο τέλος της πρότασης. Κάνοντας χρήση όλων των έγκυρων διγραμμάτων (108 έγκυρα διγράμματα) παράγεται ένας αριθμός αντιμεταθέσεων ίσος με 23,378,400. Ας υπενθυμίσουμε ότι στην περίπτωση χρήσης όλων των διγραμμάτων της πρότασης και όχι μόνο των έγκυρων διγραμμάτων ο αριθμός των αντιμεταθέσεων της πρότασης εισόδου θα ήταν ίσο με 479,001,600. Η διαφορά αυτή όπως αποτυπώνεται στο παράδειγμα αυτό είναι τεράστια αν αναλογιστεί κανείς την μείωση του υπολογιστικού φορτίου για την περαιτέρω επεξεργασία των παραγόμενων αντιμεταθέσεων. Παρατηρώντας τον πίνακα αντιστοίχισης φαίνεται ότι πολλά διγράμματα έχουν πιθανότητες που είναι πολύ μικρές. Άρα το ερώτημα που γεννιέται είναι κατά πόσον μπορούμε να μειώσουμε επιπλέον των αριθμό των αντιμεταθέσεων απαλείφοντας τις πιθανότητες αυτές των διγραμμάτων που είναι πολύ μικρές. Αυτό μπορεί να επιτευχθεί με την χρήση ενός κατωφλίου. Πώς μπορούμε όμως να ορίσουμε το βέλτιστο κατώφλι χωρίς να επηρεάζεται η σωστή πρόταση και παράλληλα να μειωθεί ο χώρος αναζήτησης κατά το μέγιστο. Από την παρατήρηση του πίνακα αντιστοίχισης μπορούμε να εξάγουμε το συμπέρασμα ότι μπορεί να χρησιμοποιηθεί σαν κατώφλι η λογαριθμική πιθανότητα ίση με -3,54. Η τιμή αυτή αντιστοιχεί στην μικρότερη λογαριθμική πιθανότητα των διγραμμάτων του άξονα της πρότασης. Όπως γίνεται φανερό η χρήση ενός τέτοιου κατωφλίου μπορεί να περιορίζει στην συγκεκριμένη περίπτωση το χώρο αναζήτησης από τις 23,378,400 σε 1,463,760 αντιμεταθέσεις αλλά σε κάθε άλλη περίπτωση δηλ. διαφορετική πρόταση και με λάθη στην σειρά των λέξεων, αυτό το κατώφλι μπορεί να ακυρώσει και την σωστή πρόταση από τον χώρο αναζήτησης.

Θέλοντας να διαπιστώσουμε πόσες από τις προτάσεις των πειραματικών δεδομένων θα απαλειφθούν λόγω της χρήσης του κατωφλίου χρησιμοποιούμε ένα μεταβλητό κατώφλι με τιμές από -7,50 ως -4,50 με βήμα -1. Ο Πίνακας 4.4 αποτυπώνει τον αριθμό των προτάσεων της βάσης Wall Street Journal (WSJ) (Robinson, 1995) που χρησιμοποιήθηκαν ως πειραματικά δεδομένα για τον προσδιορισμό του ποσοστού βελτίωσης του αριθμού των αντιμεταθέσεων με την χρήση της μεθόδου φιλτραρίσματος. Αρχικά χρησιμοποιήθηκε ένα τυχαίο σύνολο 1,189,085 προτάσεων διαφορετικού μήκους (από 7 ως 12 λέξεις). Η $1^{η}$ στήλη του πίνακα δείχνει τον αριθμό των προτάσεων ανάλογα με το μήκος τους. Η $2^{η}$ στήλη του πίνακα δείχνει τον αριθμό των προτάσεων με την χρήση κατωφλίου , T=-7,50. Η τιμή του κατωφλίου αυτή είναι μικρότερη από την μικρότερη τιμή οποιαδήποτε διγράμματος της βάσης με αποτέλεσμα να συμπεριλαμβάνονται όλες οι προτάσεις. Με την χρήση ενός τέτοιου κατωφλίου υπολογίζεται ο αριθμός αντιμεταθέσεων που παράγονται για διαφορετικές προτάσεις με τον σχηματισμό του πίνακα αντιστοίχισης. Οι επόμενες στήλες αναφέρονται στον αριθμό των προτάσεων αυξάνοντας το κατώφλι. Όπως φαίνεται η αύξηση του κατωφλίου περιορίζει τον αριθμό των αρχικών προτάσεων από την στιγμή που τα διγράμματα με λογαριθμική πιθανότητα μεγαλύτερη από T=-4,50 είναι το 74 % της βάσης διγραμμάτων του γλωσσικού μοντέλου. Από τον πίνακα που απεικονίζει τα ποσοστά των προτάσεων για διαφορετικά κατώφλια φαίνεται ότι ασφαλή συμπεράσματα για την μείωση του αριθμού των αντιμεταθέσεων μπορούν να προκύψουν μόνο από την περίπτωση όπου το κατώφλι έχει τιμή ίση με T=-7,50. Ενδεχομένως μπορεί να χρησιμοποιηθεί και η περίπτωση όπου το κατώφλι έχει τιμή ίση με T=-6,50 από την στιγμή πού μόνο το 0,5% των προτάσεων απαλείφεται, και άρα ο χώρος αναζήτησης θα συμπεριλαμβάνει την σωστή πρόταση κατά 99,5%.

		Αριθμός Προτάσεων	**Αριθμός προτάσεων μετά την εφαρμογή κατωφλίου**			
Αριθμός λέξεων ανά πρόταση	**N**		**T= -7.5**	**T= -6.5**	**T= -5.5**	**T= -4.5**
	7	269,323	269,323	268,573	254,913	185,786
	8	234,717	234,717	233,995	220,301	152,560
	9	205,216	205,216	204,513	191,447	126,349
	10	180,356	180,356	179,684	167,328	105,494
	11	158,693	158,693	158,020	145,922	87,529
	12	140,780	140,780	140,151	128,424	73,378

Πίνακας 4.4 Ο πίνακας δείχνει τον σύνολο των προτάσεων για διαφορετικό αριθμό λέξεων, που χρησιμοποιήθηκαν στα πειραματικά δεδομένα. Με την χρήση ενός κατωφλίου ο αριθμός των προτάσεων

σε κάθε κατηγορία μειώνεται λόγω της ύπαρξης διγραμμάτων αυτών των προτάσεων που έχουν τιμές μικρότερες από την τιμή του κατωφλίου.

Ο Πίνακας 4.5 αποτυπώνει τον Μ.Ο των αντιμεταθέσεων για προτάσεις μεταβλητού μήκους και δυο διαφορετικές τιμές κατωφλίων. Όπως φαίνεται η μείωση του αριθμού των αντιμεταθέσεων σε προτάσεις μήκους 12 λέξεων είναι καταλυτική αν αναλογιστούμε ότι χωρίς φιλτράρισμα απαιτείται να επεξεργαστούν 479,001,600 αντιμεταθέσεις ενώ ακόμη και στην περίπτωση του κατωφλίου ίσο με Τ=-7,50 ο αριθμός των αντιμεταθέσεων είναι μόλις 11,378,400. Το συμπέρασμα είναι ότι με χρήση ή χωρίς κατωφλίου ο αριθμός των αντιμεταθέσεων κινείται πολύ πιο χαμηλά από ότι σε κάθε άλλη περίπτωση.

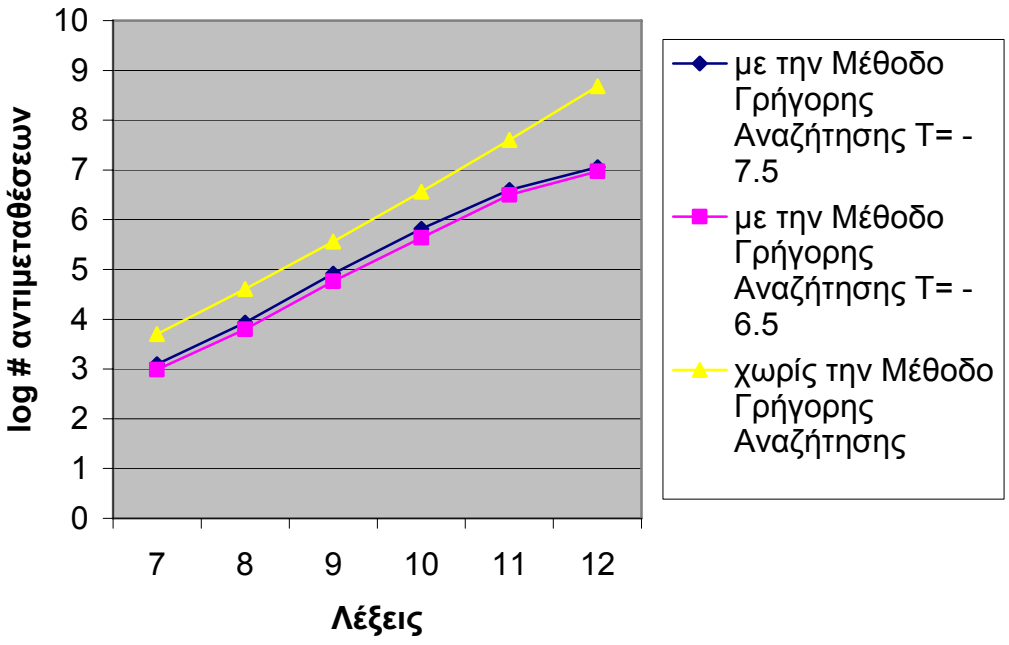
Λέξεις	**Μέθοδος Γρήγορης Αναζήτησης**		**Ν! αντιμεταθέσεις**
Ν	**Τ= -7.5**	**Τ= -6.5**	
7	1,252	968	5,040
8	8,560	6,293	40,320
9	82,012	57,890	362,880
10	652,602	429,600	3,628,800
11	3,963,401	3,127,840	39,916,800
12	11,378,400	9,378,400	479,001,600

Πίνακας 4.5 Ο Μ.Ο των αντιμεταθέσεων για προτάσεις μήκους από 7 έως 12 λέξεων με την μέθοδο γρήγορης αναζήτησης για δυο διαφορετικά κατώφλια.

Ο Πίνακας 4.6 αποτυπώνει το κέρδος σε αριθμό αντιμεταθέσεων για προτάσεις μεταβλητού μήκους που απορρέει από την χρήση της μεθόδου γρήγορης αναζήτησης. Έτσι στην περίπτωση όπου είχαμε προτάσεις με 7 λέξεις το κέρδος της χρήσης της μεθόδου είναι περίπου 4 φορές λιγότερες αντιμεταθέσεις, ενώ στην αντίστοιχη περίπτωση όπου έχουμε προτάσεις με 12 λέξεις το κέρδος φθάνει στο 42.

Αριθμός Λέξεων	Κέρδος
7	4,03
8	4,71
9	4,42
10	5,56
11	10,07
12	42,10

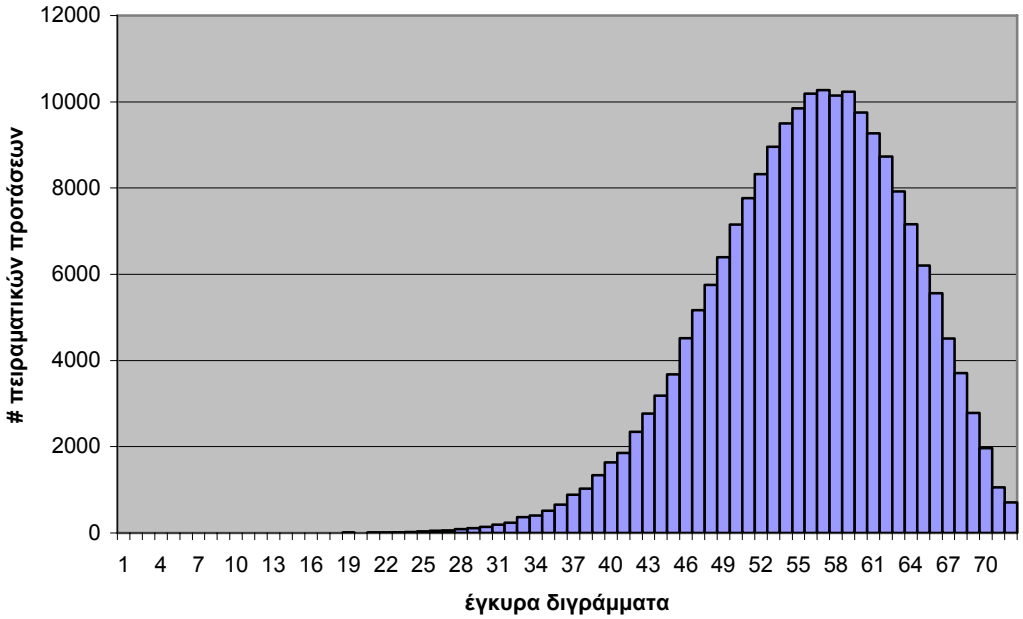
Πίνακας 4.6 Συντελεστής κέρδους πολυπλοκότητας για την Αγγλική γλώσσα.



Σχήμα 4.12 Ο Μ.Ο των αντιμεταθέσεων με την μέθοδο γρήγορης αναζήτησης σε λογαριθμική κλίμακα για προτάσεις από 7 έως 12 λέξεις και για διαφορετικά κατώφλια.

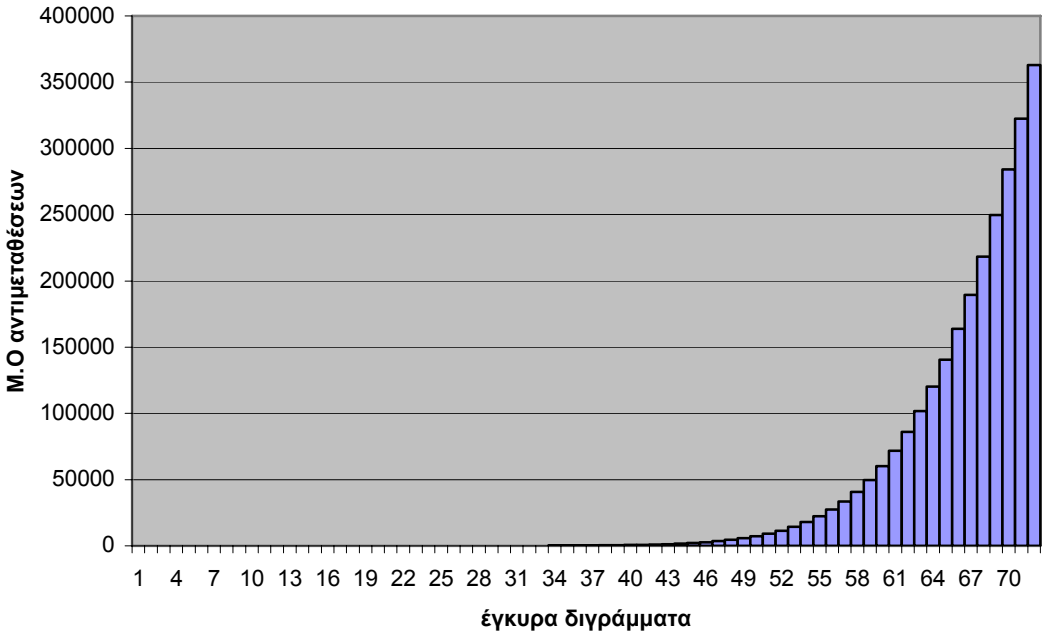
Το Σχήμα 4.13 απεικονίζει τον αριθμό προτάσεων ανά έγκυρα διγράμματα με την χρήση 205,216 προτάσεων με 9 λέξεις. Ο κυριότερος όγκος προτάσεων για T= -7,50 είναι γύρω από

την περιοχή των 55 έγκυρων διγραμμάτων την στιγμή που υπάρχουν 72 πιθανά διγράμματα.



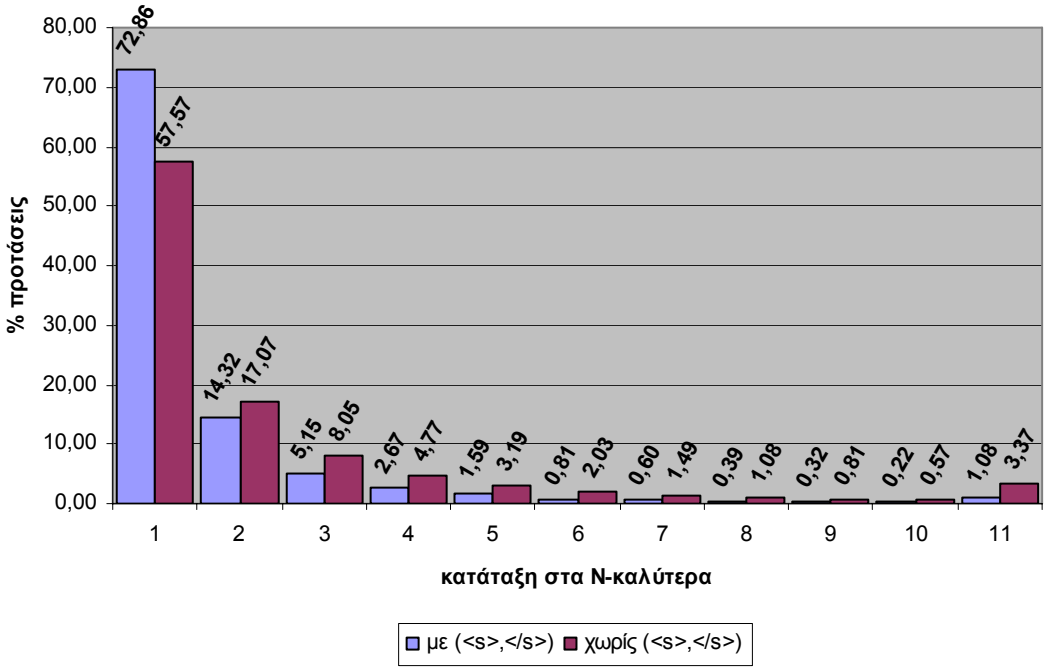
Σχήμα 4.13 Ο αριθμός των προτάσεων για διαφορετικούς αριθμούς διγραμμάτων για 205,216 προτάσεις με 9 λέξεις και κατώφλι ίσο με -7,50. Ας σημειωθεί ότι σε περίπτωση που θεωρούσαμε σαν έγκυρα όλα τα διγράμματα τότε ο αριθμός τους θα ήταν ίσος με 72.

Όπως φαίνεται στο Σχήμα 4.14 ο μέσος όρος των αντιμεταθέσεων ανά έγκυρα διγράμματα παρουσιάζει μεγάλη διακύμανση από την στιγμή που υπάρχουν κάποιες προτάσεις που έχουν έγκυρα διγράμματα όλα τα πιθανά διγράμματα με συνέπεια να αναπτύσσονται όλες οι αντιμεταθέσεις χωρίς καμία μείωση του χώρου αναζήτησης. Αυτό έχει άμεση εξάρτηση από την ποιότητα του σώματος κειμένου που χρησιμοποιείται για την κατασκευή του γλωσσικού μοντέλου.



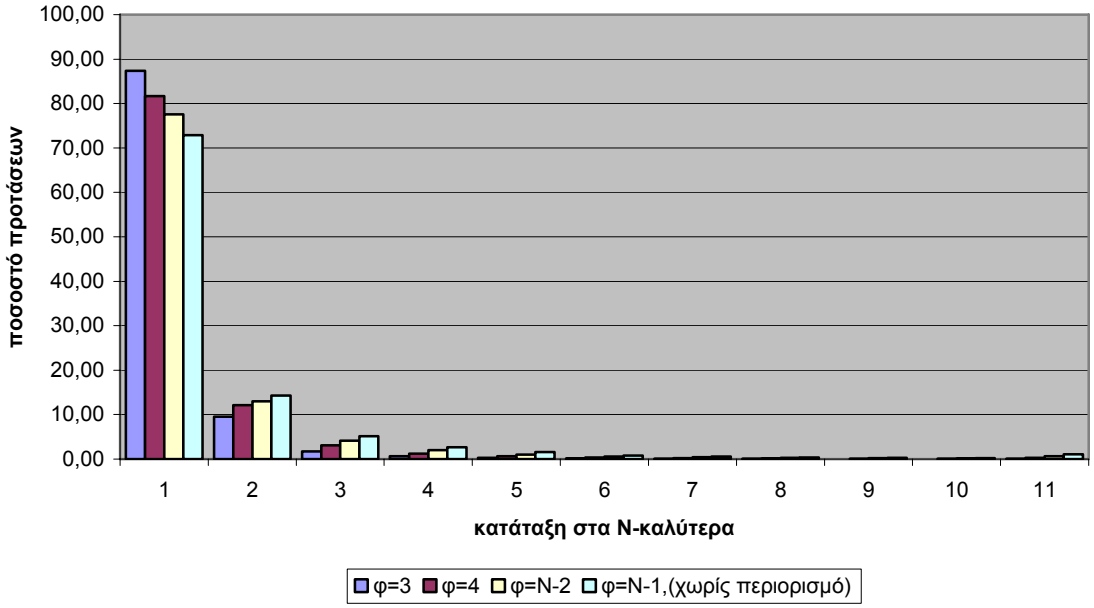
Σχήμα 4.14 Ο μέσος όρος των αντιμεταθέσεων για διαφορετικούς αριθμούς διγραμμάτων για 205,216 προτάσεις με 9 λέξεις και επιτρεπτό όριο λογαριθμικής πιθανότητας το 7,50.

Όπως αναφέραμε και παραπάνω, η απόδοση του συστήματος μπορεί να βελτιωθεί με τον υπολογισμό των πιθανοτήτων των μονογραμμάτων που βρίσκονται στην αρχική και τελική θέση μέσα σε μια πρόταση. Όπως είναι κατανοητό υπάρχουν λέξεις οι οποίες δεν μπορούν να αποτελούν την αφετηρία αλλά και την κατάληξη των προτάσεων. Άρα η απόρριψη λέξεων που δεν μπορούν να βρίσκονται σε ανάλογες θέσεις βοηθάει το σύστημα στην ανάδειξη της σωστής πρότασης στις πρώτες θέσεις των 10-καλύτερων. Όπως φαίνεται από το Σχήμα 4.15 η πειραματική διαδικασία αποδεικνύει ότι το 72,86% των προτάσεων βρέθηκαν στην πρώτη θέση ενώ το 1,08% βρίσκεται εκτός των 10 καλύτερων, με την χρήση των ενδείξεων αρχής και τέλους πρότασης (<s>,</s>). Από την άλλη πλευρά, όπου δεν γίνεται χρήση των ενδείξεων αρχής και τέλους προτάσεων τα ποσοστά διαφοροποιούνται με αποτέλεσμα το 57,57 % των προτάσεων βρέθηκαν στην πρώτη θέση ενώ το 3,37% βρίσκεται εκτός των 10-καλύτερων. Ας σημειωθεί ότι τα πειραματικά αποτελέσματα αφορούν την περίπτωση όπου το, $\phi = N - 1$.



Σχήμα 4.15 Το σχήμα αυτό δείχνει το ποσοστό των σωστών προτάσεων που βρίσκονται στην ανάλογη θέση ανάμεσα στα 10 καλύτερα με την ενσωμάτωση και των πιθανοτήτων που έχουν οι λέξεις στην αρχή και στο τέλος της κάθε πρότασης. Η στήλη 11 αντιστοιχεί σε ποσοστό προτάσεων που δεν συγκαταλέγονται στις 10 καλύτερες.

Για την αξιολόγηση της απόδοσης του συστήματος με την χρήση της μεθόδου γρήγορης αναζήτησης της βέλτιστης λύσης χρησιμοποιήθηκαν όλες οι προτάσεις μήκους από 7 ως 12 λέξεις που παρήχθησαν με την χρήση του κατωφλίου Τ-7,50, και για διαφορετικές τιμές του ϕ. Επιπλέον πρέπει να τονιστεί ότι έχει συνυπολογισθεί και η πιθανότητα εμφάνισης των μονογραμμάτων στην αρχή και στο τέλος των προτάσεων του σώματος εκπαίδευσης. Ο στόχος του πειράματος είναι να αποδειχθεί ότι η σωστή πρόταση συγκαταλέγεται στις 10 καλύτερες προτάσεις που απορρέουν με την χρήση του συστήματος. Έτσι όταν η πρόταση εισόδου που είναι και παράλληλα σωστή βρεθεί στην πρώτη θέση των 10 καλύτερων σημαίνει ότι το σύστημα αξιολογεί αυτή την πρόταση ως την πρόταση με τα περισσότερα τριγράμματα και τη μεγαλύτερη βαθμολογία. Όπως φαίνεται από την παρακάτω σχηματική παράσταση (Σχήμα 4.16) η πειραματική διαδικασία αποδεικνύει ότι όσο το ϕ μεγαλώνει το ποσοστό των προτάσεων που βρίσκονται στην πρώτη θέση ελαττώνεται ενώ αντίστροφα μεγαλώνει ελαφρά το ποσοστό των προτάσεων που κατατάσσονται εκτός των 10 καλύτερων. Η απόδοση του συστήματος δείχνει να φθάνει στο 88% όταν χρησιμοποιείται ο παράγοντας ϕ ίσος με 3.



Σχήμα 4.16 Το σχήμα αυτό δείχνει το ποσοστό των προτάσεων που βρίσκονται στην ανάλογη θέση ανάμεσα στα 10 καλύτερα με την χρήση διαφορετικών τιμών (φ). Η στήλη 11 αντιστοιχεί σε ποσοστό προτάσεων που δεν συγκαταλέγονται στις 10 καλύτερες.

4.4.4 Σύγκριση με υπάρχοντα συστήματα διόρθωσης κειμένων

Για την διόρθωση και ανίχνευση γραμματικών λαθών έχουνε προταθεί όπως είδαμε πολλές μέθοδοι και πολλές από αυτές έχουν ενσωματωθεί σε υπάρχοντα εμπορικά προϊόντα. Για την αξιολόγηση τους, δεν έχει γίνει συστηματική έρευνα από την στιγμή που δεν διορθώνουν τα ίδια είδη γραμματικών λαθών και επιπλέον δεν υπάρχει κοινό σύνολο από πειραματικά δεδομένα για δοκιμή επιδόσεων. Με την παρούσα εργασία επιδιώκουμε να αξιολογήσουμε τις επιδόσεις ενός πολύ γνωστού κειμενογράφου όπως είναι το Word του Microsoft Office ®. Για τον λόγο αυτό χρησιμοποιούμε την έκδοση λογισμικού Word 2003 και 2007 και ένα σύνολο 1,189,085 προτάσεων (7-12 λέξεων) από το σώμα κειμένου της WSJ. Οι προτάσεις αυτές είναι οι ίδιες που χρησιμοποιήθηκαν για την αξιολόγηση της προτεινόμενης μεθόδου (Ενότητα 4.4.3). Για πειραματικούς λόγους θα χρησιμοποιηθεί μόνο μια εκδοχή των N! αντιμεταθέσεων για την κάθε πρόταση εισόδου. Αυτή προέρχεται από την χρήση συντελεστή (ϕ) ίσου με τρία και με δυνατότητα μετατόπισης μέσα στην πρόταση N-2 λέξεων όταν το N κινείται από 7 ως 12. Η επιλογή της μοναδικής εκδοχής ανάμεσα στις N! αντιμεταθέσεις γίνεται με τυχαίο τρόπο. Αξιολογώντας τις επιδόσεις των κειμενογράφων Word 2003 και 2007 ως προς τη ανίχνευση λαθών αναφορικά με την θέση των λέξεων μέσα στην πρόταση, αποδείχθηκε ότι στην πρώτη περίπτωση, μόνο στο 35,25% των προτάσεων εντοπίσθηκε λάθος χωρίς να

προτείνεται κάποια καλύτερη εκδοχή και χωρίς να υποδεικνύεται η περιοχή του λάθους, ενώ στην δεύτερη περίπτωση μόνο στο 45,73% των προτάσεων.

4.5 Πειραματικά αποτελέσματα της μεθόδου για την Ελληνική γλώσσα

Ο Πίνακας 4.7 αποτυπώνει τον αριθμό των πειραματικών προτάσεων που επιλεχθήκαν τυχαία από δημοσιογραφικά κείμενα με στόχο την αξιολόγηση της απόδοσης της μεθόδου γρήγορης αναζήτησης και συνολικά του συστήματος. Αρχικά χρησιμοποιήθηκε ένα τυχαίο σύνολο 592,819 προτάσεων διαφορετικού μήκους (από 7 ως 12 λέξεις). Όπως αναφέρθηκε και προηγουμένως στην ενότητα των πειραματικών αποτελεσμάτων για την Αγγλική γλώσσα, πρέπει να εξεταστεί το κατά πόσον όλες οι πειραματικές προτάσεις περιλαμβάνουν διγράμματα που ανήκουν στο γλωσσικό μοντέλο. Η $3^{η}$ στήλη του πίνακα δείχνει ότι ο αριθμός των προτάσεων παραμένει ο ίδιος ύστερα και από την χρήση του κατωφλιού με τιμή το T=-7,50. Με την χρήση κατωφλίου (T=-7,50) υπολογίζεται ο αριθμός αντιμεταθέσεων που παράγονται για διαφορετικές προτάσεις.

Αριθμός Προτάσεων			**Αριθμός προτάσεων μετά την εφαρμογή κατωφλίου**
Αριθμός λέξεων ανά πρόταση	**N**	**Σύνολο**	**T= -7.5**
	7	135,352	135,352
	8	117,901	117,901
	9	103,050	103,050
	10	89,943	89,943
	11	78,287	78,287
	12	68,286	68,286

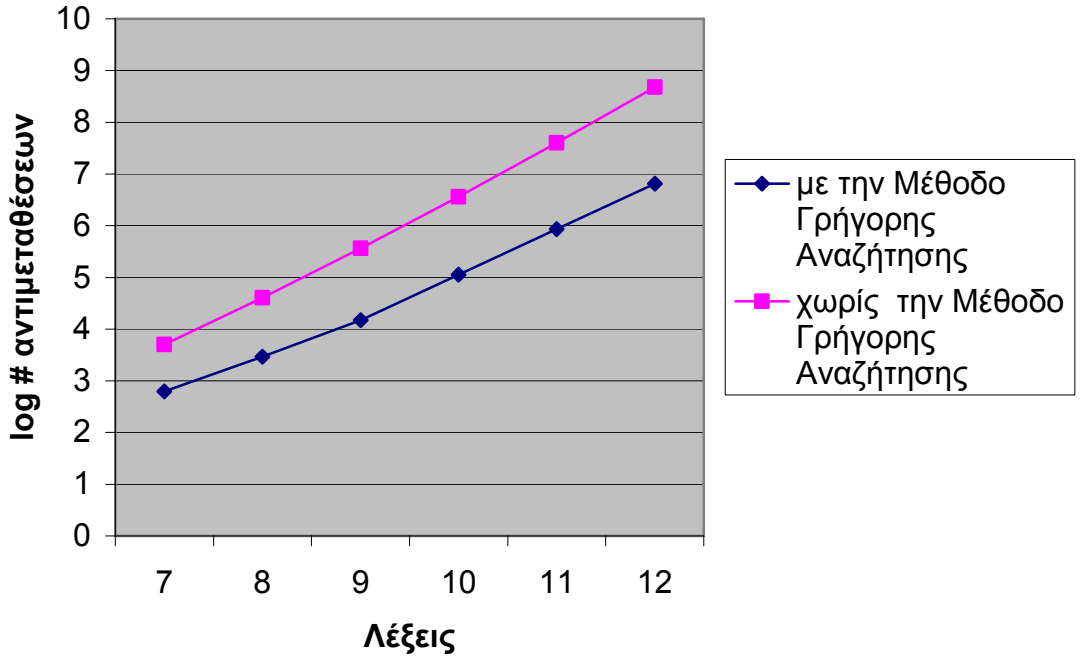
Πίνακας 4.7 Η πρώτη στήλη του πίνακα δείχνει τον σύνολο των προτάσεων με διαφορετικό αριθμό λέξεων, που χρησιμοποιήθηκαν σαν αρχικά πειραματικά δεδομένα.

Ο Πίνακας 4.8 αποτυπώνει τον Μ.Ο των αντιμεταθέσεων για προτάσεις μεταβλητού μήκους για τιμή κατωφλίου ίση με T=-7,50. Όπως φαίνεται η μείωση του αριθμού των αντιμεταθέσεων σε προτάσεις μήκους 12 λέξεων είναι καταλυτική αν αναλογιστούμε ότι χωρίς φιλτράρισμα απαιτούνται 479,001,600 αντιμεταθέσεις ενώ στην περίπτωση χρήσης της

μεθόδου γρήγορης αναζήτησης ο αριθμός των αντιμεταθέσεων είναι μόλις 6,502,661. Στο Σχήμα 4.17 φαίνεται ο Μ.Ο αντιμεταθέσεων σε λογαριθμική κλίμακα.

Ν Λέξεις	**Μέθοδος Γρήγορης Αναζήτησης**	**Ν! αντιμεταθέσεις**
7	623	5,040
8	2941	40,320
9	14,794	362,880
10	113,456	3,628,800
11	852,678	39,916,800
12	6,502,661	479,001,600

Πίνακας 4.8 Ο Μ.Ο των αντιμεταθέσεων για προτάσεις μήκους από 7 έως 12 λέξεων στις περιπτώσεις όπου γίνεται χρήση της μεθόδου γρήγορης αναζήτησης της βέλτιστης λύσης.



Σχήμα 4.17 Ο Μ.Ο των αντιμεταθέσεων με ή χωρίς την χρήση της μεθόδου γρήγορης αναζήτησης σε λογαριθμική κλίμακα για προτάσεις από 7 έως 12 λέξεις.

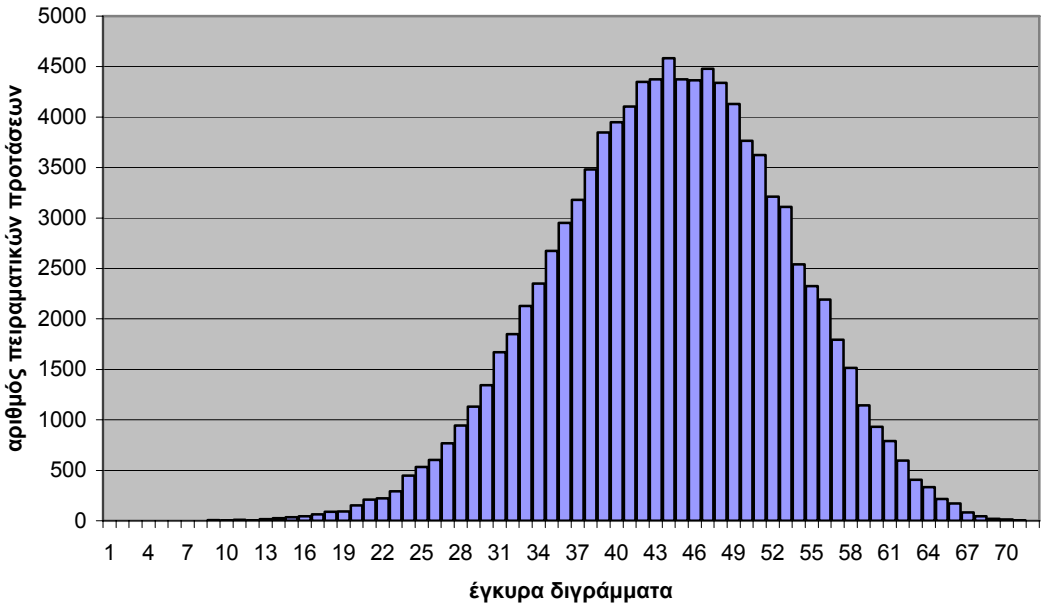
Ο Πίνακας 4.9 αποτυπώνει το κέρδος σε αριθμό αντιμεταθέσεων για προτάσεις μεταβλητού μήκους που απορρέει από την χρήση της μεθόδου γρήγορης αναζήτησης. Έτσι στην περίπτωση όπου είχαμε προτάσεις με 7 λέξεις το κέρδος της χρήσης της μεθόδου είναι περίπου 8 φορές

λιγότερες αντιμεταθέσεις, ενώ στην αντίστοιχη περίπτωση όπου έχουμε προτάσεις με 12 λέξεις το κέρδος φθάνει στο 73.

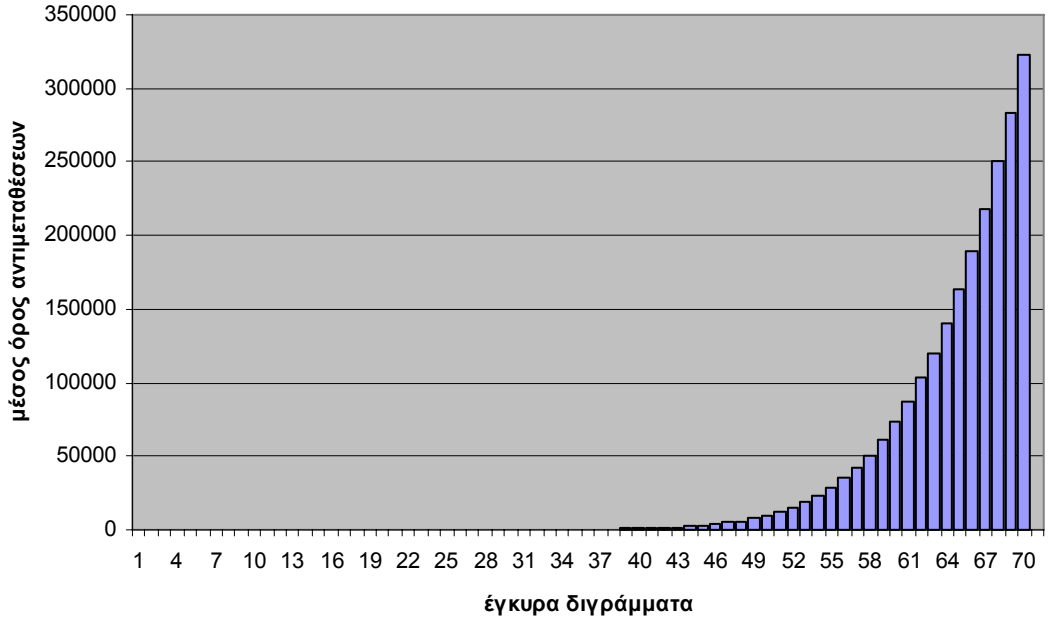
Αριθμός Λέξεων	**Κέρδος**
7	8,09
8	13,71
9	24,53
10	31,98
11	46,81
12	73,66

Πίνακας 4.9 Συντελεστής κέρδους πολυπλοκότητας για την Ελληνική γλώσσα .

Το Σχήμα 4.18 απεικονίζει τον αριθμό των πειραματικών προτάσεων ανά έγκυρα διγράμματα με την χρήση 103,050 προτάσεων με 9 λέξεις. Ο κυριότερος όγκος προτάσεων για T= -7,50 είναι γύρω από την περιοχή των 44-49 έγκυρων διγραμμάτων την στιγμή που υπάρχουν 72 πιθανά διγράμματα. Στο Σχήμα 4.19 φαίνεται ο Μ.Ο των αντιμεταθέσεων ανά αριθμό έγκυρων διγραμμάτων.

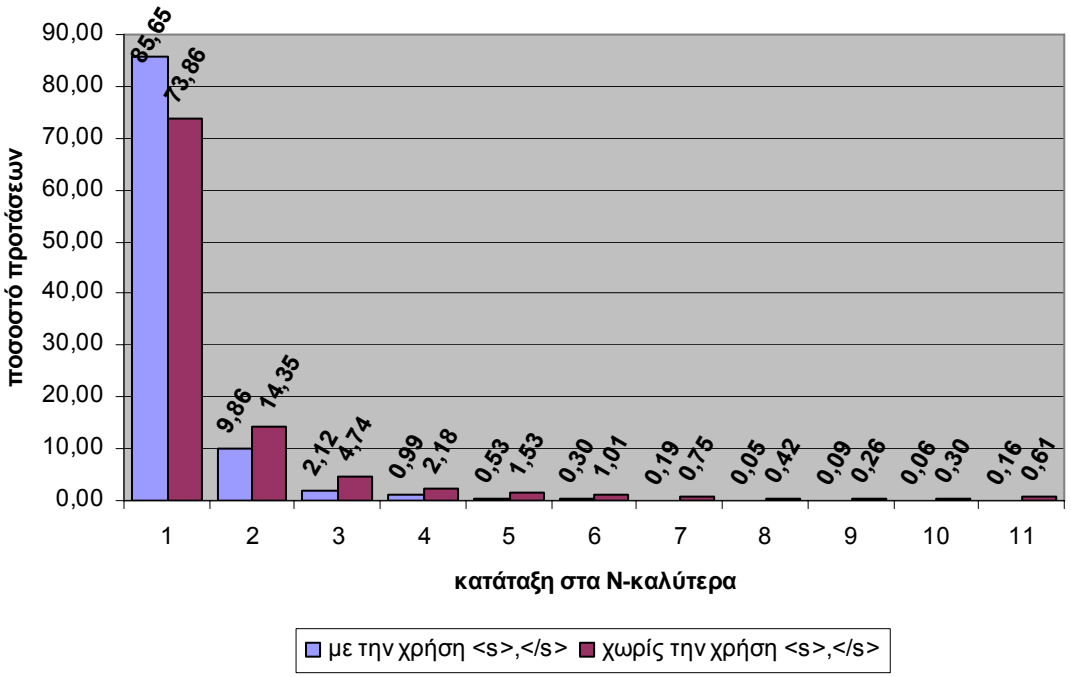


Σχήμα 4.18 Ο αριθμός των προτάσεων για διαφορετικούς αριθμούς διγραμμάτων για 103,050 προτάσεις με 9 λέξεις και και κατώφλι ίσο με T=-7,50. Ας σημειωθεί ότι στην περίπτωση που θεωρούσαμε σαν έγκυρα όλα τα διγράμματα τότε ο αριθμός τους θα ήταν ίσος με 72.



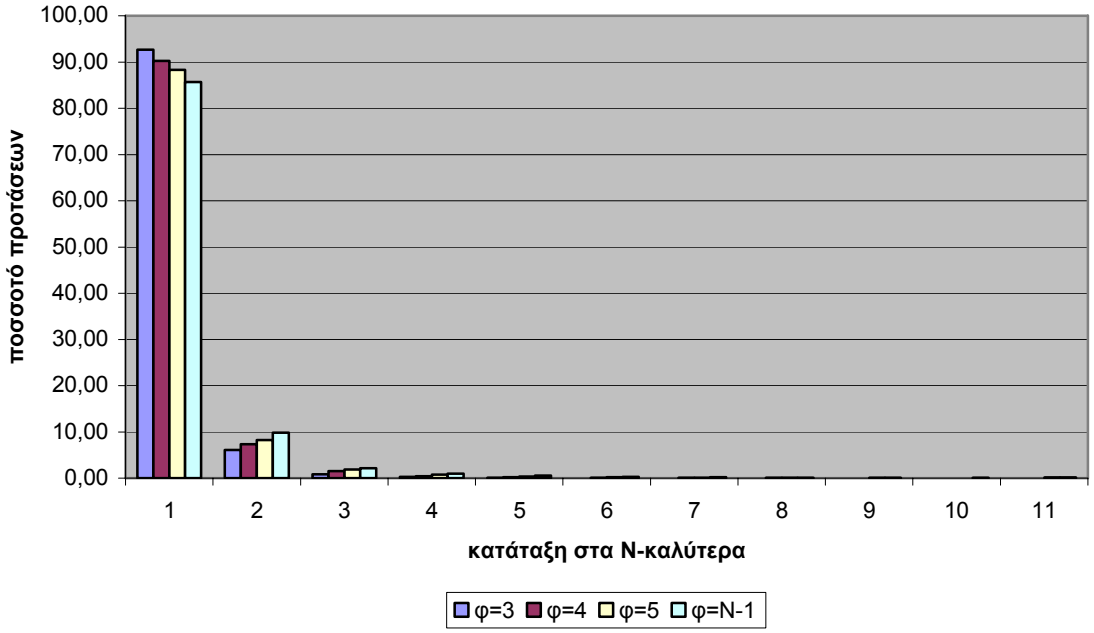
Σχήμα 4.19 Ο μέσος όρος των αντιμεταθέσεων για διαφορετικούς αριθμούς διγραμμάτων για 103,050 προτάσεις με 9 λέξεις και επιτρεπτό όριο λογαριθμικής πιθανότητας το T=-7,50.

Όπως φαίνεται από την παρακάτω σχηματική παράσταση (Σχήμα 4.20) η πειραματική διαδικασία αποδεικνύει ότι το 73,86% των προτάσεων βρέθηκαν στην πρώτη θέση ενώ το 0,61% βρίσκεται εκτός των 10 καλύτερων, για $\phi = N - 1$ και χωρίς την χρήση της πληροφορίας που φέρει η πιθανότητα των μονογραμμάτων να βρίσκονται στην πρώτη και στην τελευταία θέση των υποψήφιων προτάσεων (<s>,</s>). Αντίθετα, όταν γίνεται χρήση της αντίστοιχης πληροφορίας το ποσοστό των προτάσεων που καταλαμβάνουν την πρώτη θέση ανέρχεται στο 85,65% ενώ το ποσοστό των προτάσεων που βαθμολογούνται σε θέση εκτός των 10 καλύτερων μειώνεται στο ποσοστό του 0,16%. Με αυτόν τον τρόπο αποδεικνύεται ότι η προτεινόμενη μέθοδος μπορεί να ανιχνεύσει και διορθώσει το 85,65% των προτάσεων που θα χρησιμοποιηθούν σαν είσοδο στο σύστημα. Ας σημειωθεί ότι τα πειραματικά αποτελέσματα αφορούν την περίπτωση όπου $\phi = N - 1$.



Σχήμα 4.20 Το ποσοστό των προτάσεων που καταλαμβάνουν διαφορετικές θέσεις ανάμεσα στις 10 καλύτερες με ή χωρίς την χρήση των πιθανοτήτων των μονογραμμάτων.

Για την αξιολόγηση της απόδοσης του συστήματος με την χρήση της μεθόδου γρήγορης αναζήτησης χρησιμοποιήθηκαν όλες οι προτάσεις μήκους από 7 ως 12 λέξεις που παρήχθησαν με την χρήση του κατωφλίου Τ=-7,50 και για διαφορετικά ϕ (Σχήμα 4.21). Όπως ήδη έχει αναφερθεί το ϕ παίρνει τιμές από 3 ως N-1. Επιπλέον πρέπει να τονιστεί ότι έχει συνυπολογισθεί και η πιθανότητα εμφάνισης των μονογραμμάτων στην αρχή και στο τέλος των προτάσεων του σώματος εκπαίδευσης. Ο στόχος του πειράματος είναι να αποδειχθεί ότι η σωστή πρόταση συγκαταλέγεται στις 10 καλύτερες προτάσεις που απορρέουν με την χρήση του συστήματος. Έτσι όταν η πρόταση εισόδου που είναι και παράλληλα σωστή βρεθεί στην πρώτη θέση των 10 καλύτερων σημαίνει ότι το σύστημα αξιολογεί αυτή την πρόταση ως την πρόταση με τα περισσότερα τριγράμματα και την μεγαλύτερη βαθμολογία. Παράλληλα πρέπει να τονιστεί ότι για μεγαλύτερα ϕ το ποσοστό των προτάσεων που καταλαμβάνουν την πρώτη θέση ανάμεσα στα 10 καλύτερα μειώνεται ενώ το ποσοστό των προτάσεων που βρίσκονται στις υπόλοιπες θέσεις αυξάνει. Το ποσοστό των προτάσεων που απαλείφονται με την χρήση του παράγοντα $\phi = 3$ είναι το 0,36% των συνολικών προτάσεων που θα υφίσταντο για $\phi = N - 1$.



Σχήμα 4.21 Τα αποτελέσματα κατάταξης των πειραματικών προτάσεων ανάμεσα στις 10 καλύτερες για διαφορετικές τιμές του φ.

4.6 Αξιολόγηση της μεθόδου αναδιάταξης λέξεων από ομάδα χρηστών

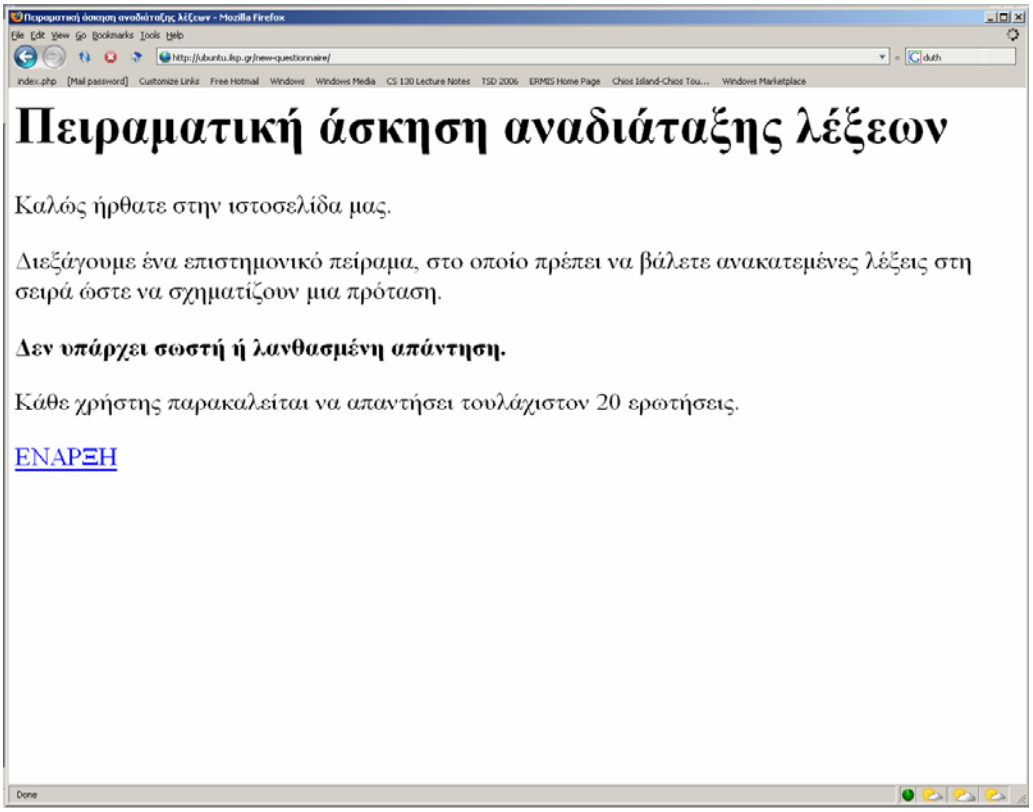
4.6.1 Σκοπός της πειραματικής άσκησης

Μέχρι στιγμής στις δυο παραπάνω πειραματικές διαδικασίες έχουν χρησιμοποιηθεί προτάσεις από σώματα κειμένων στην Αγγλική και στην Ελληνική γλώσσα ώστε να ελεγχθεί η απόδοση του συστήματος. Αυτό που αναμένουμε είναι ότι με την χρήση της μεθόδου οι προτάσεις εισόδου θα έχουν ένα σκορ μεγαλύτερο από οποιαδήποτε άλλη πρόταση που έχει προκύψει από την τεχνική των αντιμεταθέσεων. Για την Αγγλική γλώσσα χρησιμοποιώντας το σώμα κειμένου της Wall Street Journal αποδείχθηκε ότι το 87,40% των προτάσεων μπορούν να διορθωθούν με την χρήση της προτεινόμενης μεθόδου, ενώ για την Ελληνική γλώσσα το αντίστοιχο ποσοστό είναι πολύ υψηλότερο και φθάνει το 92,63%. Αυτό όμως που μπορεί να παρατηρηθεί είναι ότι ένα σημαντικό ποσοστό αρχικών προτάσεων δεν καταλαμβάνουν την πρώτη θέση αλλά την δεύτερη και πολλές φορές την τρίτη θέση. Το ερώτημα λοιπόν που τίθεται είναι κατά πόσον οι προτάσεις αυτές που ανήκουν στις πρώτες θέσεις των 10 καλύτερων, είναι αποδεκτές από έναν αναγνώστη; Αυτό προκύπτει από το γεγονός ότι πολλές προτάσεις έχουν διπλή και τριπλή εγγραφή χωρίς να αλλοιώνετε το νόημα τους. Για όλες τις γλώσσες ισχύει ότι σε αρκετές

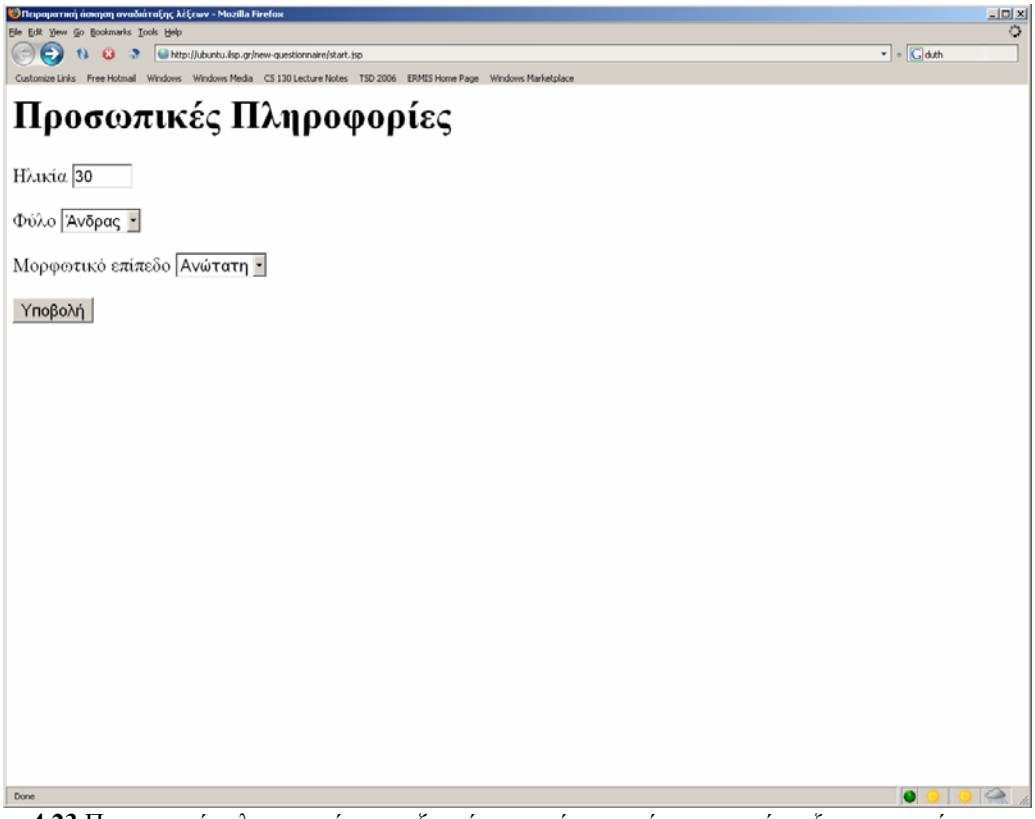
περιπτώσεις οι λέξεις μιας πρότασης μπορούν να τοποθετηθούν σε διαφορετική θέση από την κανονική για λόγους έμφασης. Για αυτόν τον λόγο θα γίνει χρήση μιας πειραματικής άσκησης που θα υλοποιηθεί από χρήστες μέσω του διαδικτύου. Ο στόχος της άσκησης αυτής είναι να συλλεχθούν δεδομένα από διαφορετικούς χρήστες για τον τρόπο σύνταξης 150 προτάσεων. Η πειραματική άσκηση αυτή έχει να κάνει με την διαφορετικότητα με την οποία οι χρήστες διατάσσουν τις λέξεις ώστε να προκύψουν νοηματικά σωστές προτάσεις. Κίνητρο όλης αυτής της διαδικτυακής άσκησης είναι να δείξουμε ποια σχέση υπάρχει μεταξύ των επιλογών των χρηστών και των προτάσεων που προτείνονται από το σύστημα.

4.6.2 Οργάνωση της πειραματικής άσκησης

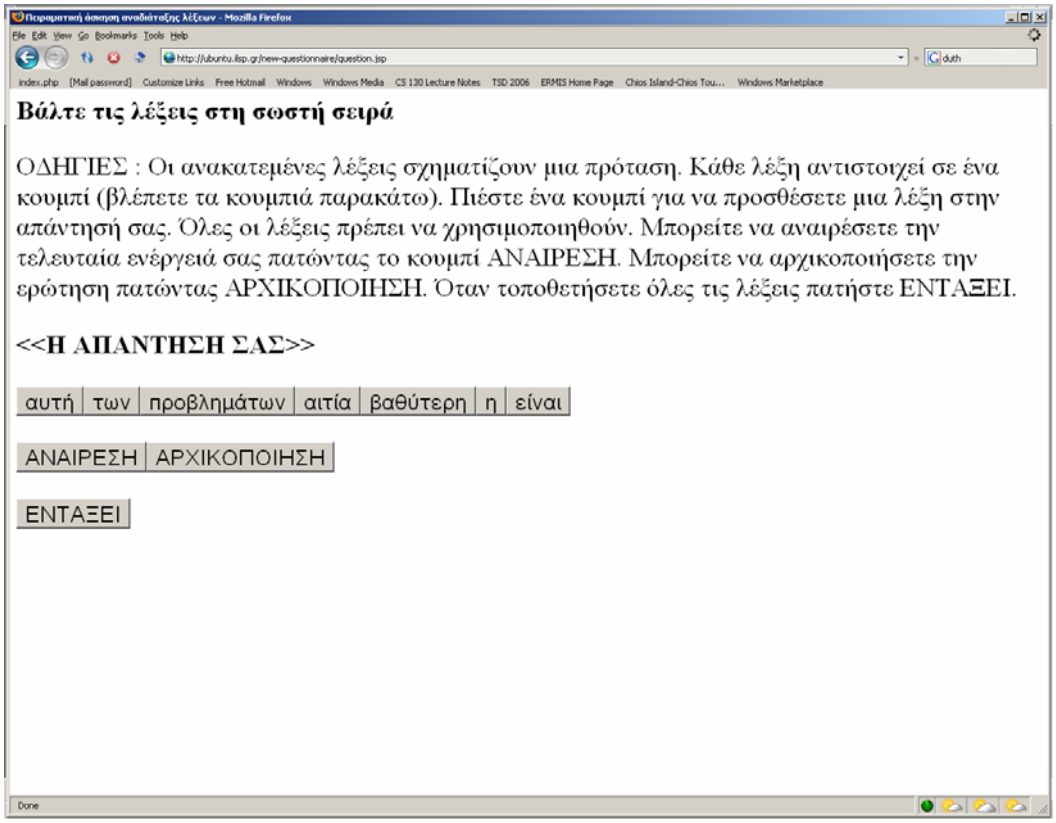
Ο διαδικτυακός κόμβος της εφαρμογής φιλοξενήθηκε στην ιστοσελίδα http://ubuntu.ilsp.gr/new-questionnaire. Η πειραματική άσκηση οργανώνεται σε 50 προτάσεις των επτά, οκτώ και εννιά λέξεων. Η επιλογή του μήκους των προτάσεων έγινε με το κριτήριο της μη αλλοίωσης του νοήματος μιας πρότασης. Προτάσεις με περισσότερες από 9 λέξεις μπορούν να αναδιαταχθούν με διαφορετικούς τρόπους και να έχουν λογικό περιεχόμενο. Από κάθε χρήστη ζητήθηκε να απαντήσει σε τουλάχιστον 20 προτάσεις. Οι προτάσεις που παρατίθενται έχουν τυχαία σειρά λέξεων χωρίς να δίνεται καμιά πληροφορία για το πώς πρέπει να αναδιαταχθούν. Η επιλογή των προτάσεων έγινε με τυχαίο τρόπο χωρίς να λαμβάνεται υπόψη η βαθμολόγηση τους από το σύστημα. Οι προτάσεις αυτές δεν περιλαμβάνουν σημεία στίξης και διπλές λέξεις. Παράλληλα δεν χρησιμοποιηθήκαν προτάσεις που περιλαμβάνουν συνδετικούς όρους και κύρια ονόματα. Επιπλέον κατά την αρχική πρόσβαση του κάθε χρήστη συμπληρώνονται πληροφορίες σχετικά με την ηλικία, φύλο και μορφωτικό επίπεδο. Δίνεται η δυνατότητα στον χρήστη να βλέπει το ιστορικό των απαντήσεων του και να επαναλαμβάνει την πειραματική διαδικασία από εκεί που σταμάτησε. Ο εκάστοτε χρήστης δεν έχει την δυνατότητα να επισκεφτεί το ιστορικό άλλων χρηστών ή να δει πόσες απαντήσεις υπάρχουν στην κάθε πρόταση. Το ανώτατο όριο προτάσεων που μπορεί να εκπονήσει είναι 150 προτάσεις και για κάθε μια πρόταση μπορεί να δώσει μια μόνο απάντηση, χωρίς να μπορεί να παρακάμψει προτάσεις. Η εφαρμογή δεν παρέχει την δυνατότητα στον χρήστη να αλλάξει την απάντηση που έδωσε για μια πρόταση όταν αυτή προστεθεί στο ιστορικό. Η ταξινόμηση των απαντήσεων θα γίνει στο σύνολο των χρηστών βάσει της ομοιότητας των απαντήσεων. Τα παρακάτω σχήματα αποτυπώνουν την μορφή και την λειτουργικότητα της διαδικτυακής εφαρμογής.



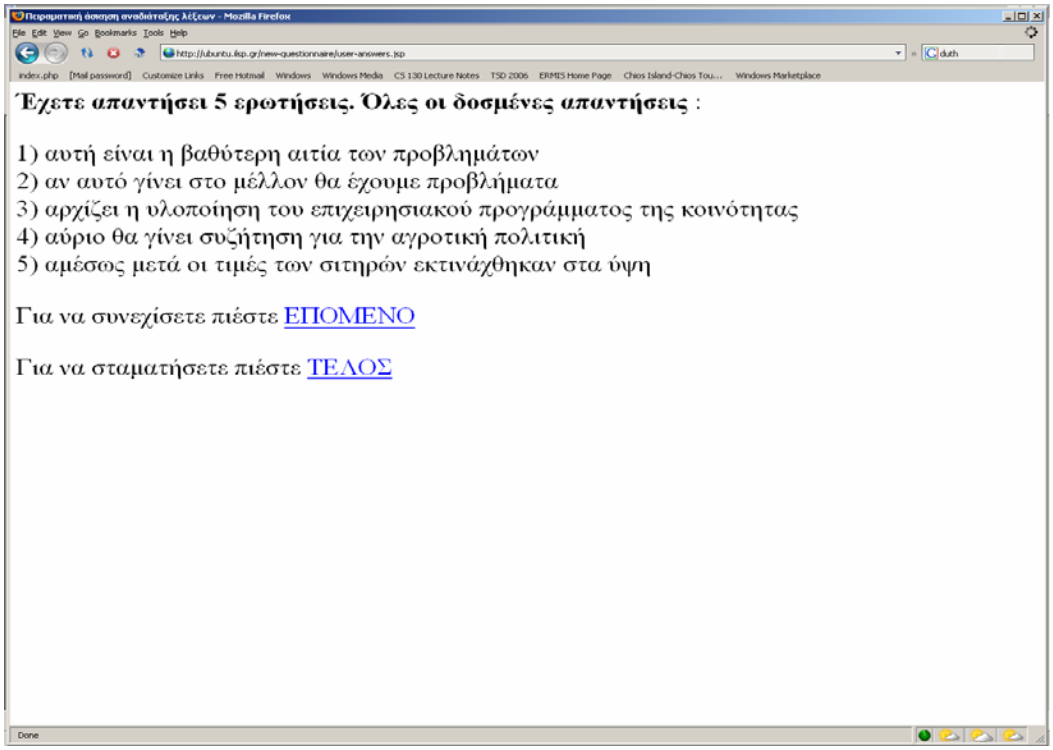
Σχήμα 4.22 Η διεπαφή της εφαρμογής «Πειραματική άσκηση αναδιάταξης λέξεων» που χρησιμοποιήθηκε στο πλαίσιο της συλλογής αποτελεσμάτων για την αξιολόγηση της μεθόδου από ομάδα χρηστών.



Σχήμα 4.23 Προσωπικές πληροφορίες που ζητούνται από τον χρήστη στην έναρξη του πειράματος.



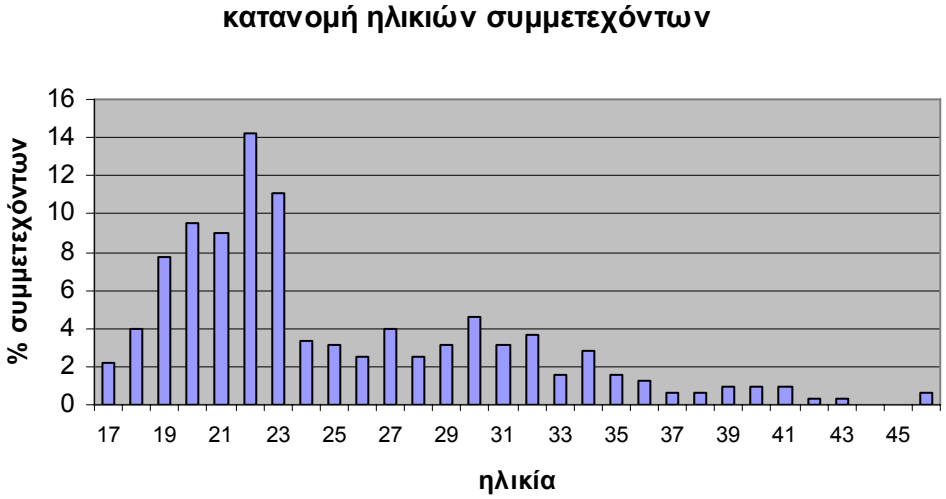
Σχήμα 4.24 Μια από τις πειραματικές ασκήσεις με στόχο την αναδιάταξη των λέξεων μιας προτάσης που εμφανίζονται σε τυχαία σειρά.



Σχήμα 4.25 Η συγκεντρωτική λίστα «ιστορικό» με τις προτάσεις που έχουν ήδη απαντηθεί από τον χρήστη.

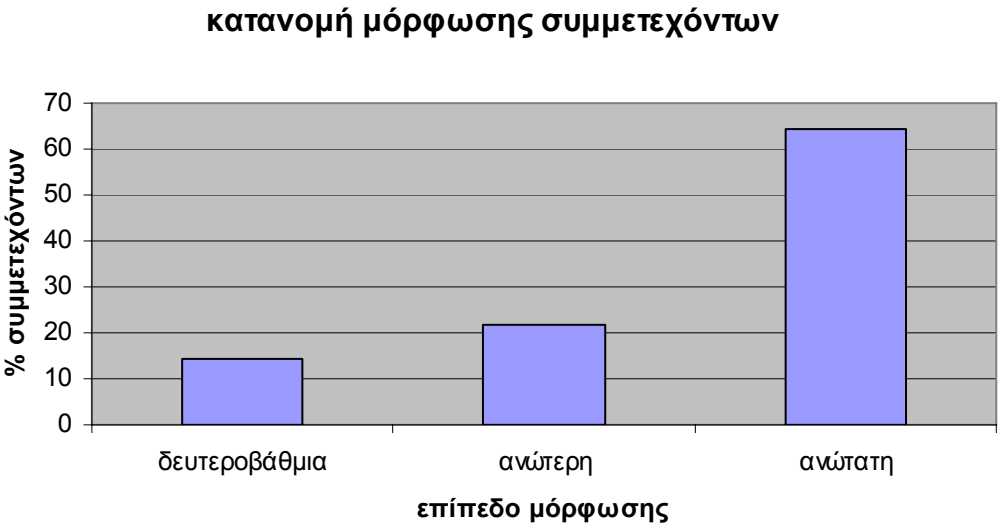
4.6.3 Ανάλυση πειραματικών δεδομένων

Στην διαδικτυακή πειραματική άσκηση συμμετείχαν 324 άτομα με μικρότερη ηλικία τα 17 έτη και μεγαλύτερη τα 46 έτη. Η κατανομή των ηλικιών των συμμετεχόντων φαίνεται στο παρακάτω ιστόγραμμα.



Σχήμα 4.26 Η κατανομή των ηλικιών των συμμετεχόντων στην πειραματική διαδικασία.

Από τους συμμετέχοντες ζητήθηκε επίσης να δηλώσουν το φύλο τους και το μορφωτικό τους επίπεδο. Στην διαδικασία συμμετείχαν 65% άνδρες και 35 % γυναίκες. Για το μορφωτικό επίπεδο δόθηκαν τρεις επιλογές : δευτεροβάθμια, ανώτερη, ανώτατη εκπαίδευση. Στο παρακάτω σχήμα αποτυπώνονται τα ποσοστά των ατόμων με την ανάλογη μόρφωση.



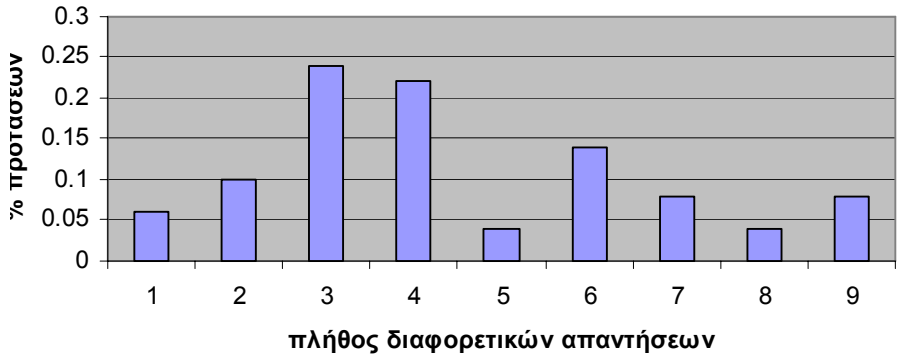
Σχήμα 4.27 Η κατανομή του μορφωτικού επιπέδου των συμμετεχόντων στην πειραματική διαδικασία.

Όπως αναφέρθηκε και προηγουμένως, στο πλαίσιο της πειραματικής διαδικασίας χρησιμοποιήθηκαν 150 προτάσεις. Από κάθε συμμετέχοντα ζητήθηκε να απαντήσει σε 20 τουλάχιστον από αυτές. Όλες οι απαντήσεις που δόθηκαν – (1 εως 150) – λαμβάνονται υπ' όψιν στην επεξεργασία των πειραματικών δεδομένων. Το διαδικτυακό σύστημα προσπαθεί να κρατήσει το πλήθος των απαντήσεων που δίνονται σε κάθε πρόταση σταθερό : όταν κάποιος χρήστης ζητά να συνεχίσει, το σύστημα ελέγχει ποιες προτάσεις έχει απαντήσει και του προτείνονται αυτές που έχουν τις λιγότερες συνολικές απαντήσεις. Στο Σχήμα 4.27 φαίνεται η κατανομή του πλήθους των απαντήσεων που δόθηκαν ανά συμμετέχοντα.

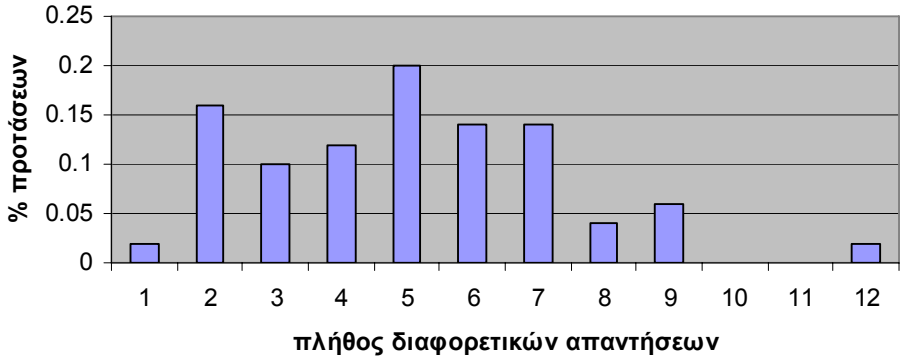
Σχήμα 4.27 Η κατανομή των απαντήσεων αν συμμετέχοντα.

Όπως φαίνεται από το παραπάνω σχήμα, ένα μεγάλο ποσοστό των συμμετεχόντων (το 8.95%, δηλαδή 29 συμμετέχοντες) δεν έδωσε καμία απάντηση. Τριάντα δυο συμμετέχοντες (το 9.88% περίπου) έδωσαν μόνο μία απάντηση. Είναι εμφανές ότι μεγάλο μέρος των συμμετεχόντων (οι 49, δηλαδή το 15.12%) απάντησε ακριβώς σε 20 ερωτήσεις.

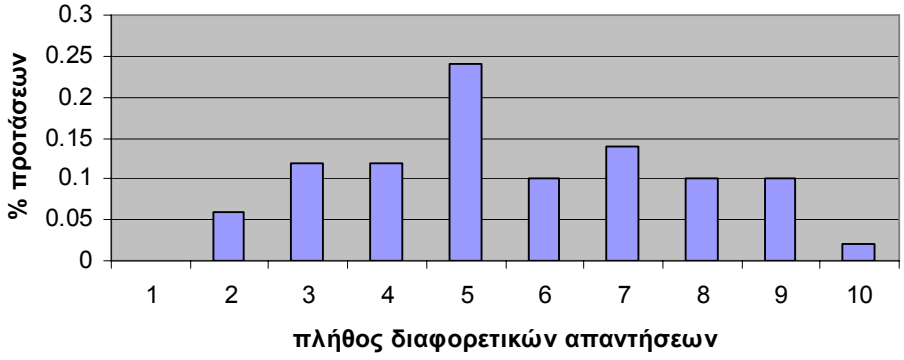
Συνολικά συλλέχθησαν 5034 απαντήσεις. Άρα κάθε πρόταση ανασχηματίστηκε κατά μέσο όρο 5034/150 = 33.56 φορές. Το ερώτημα που τίθεται είναι πόσες διαφορετικές απαντήσεις έχουν δώσει οι συμμετέχοντες για κάθε μια ερώτηση-πρόταση. Για παράδειγμα η πρώτη πρόταση έχει απαντηθεί 34 φορές και έχουν ληφθεί 4 διαφορετικές απαντήσεις (η πρώτη απάντηση έχει δοθεί από 25 χρήστες, η δεύτερη από 4, η τρίτη από 3 και η τέταρτη από 2). Έχει ενδιαφέρον να διερευνήσουμε την κατανομή (ποσοστό προτάσεων) – (πλήθος διαφορετικών απαντήσεων ανά πρόταση) κυρίως σε σχέση με το πλήθος των λέξεων που περιέχει η κάθε πρόταση.



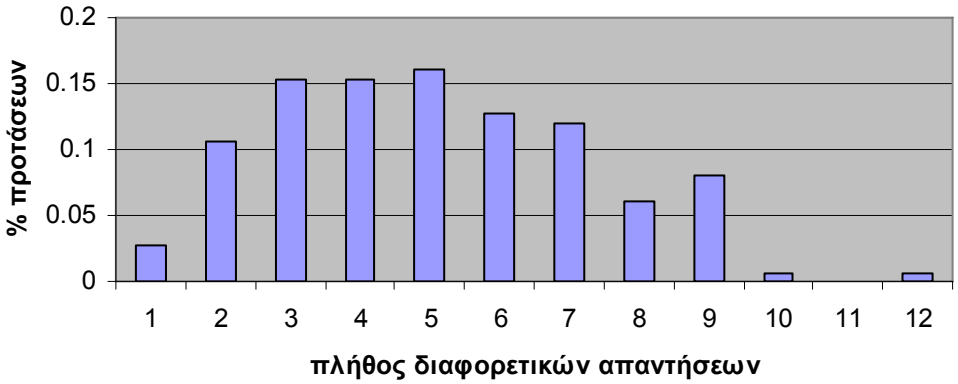
Σχήμα 4.28 Κατανομή πλήθους διαφορετικών απαντήσεων για τις προτάσεις με 7 λέξεις.



Σχήμα 4.29 Κατανομή πλήθους διαφορετικών απαντήσεων για τις προτάσεις με 8 λέξεις.

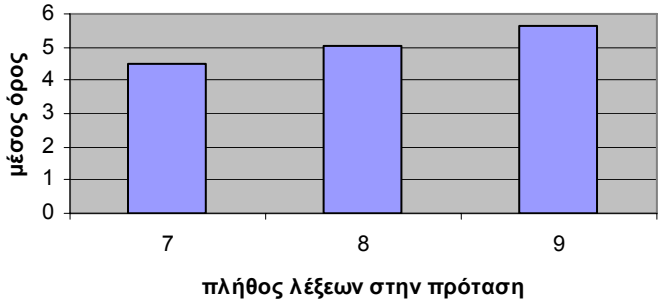


Σχήμα 4.30 Κατανομή πλήθους διαφορετικών απαντήσεων για τις προτάσεις με 9 λέξεις.



Σχήμα 4.31 Κατανομή πλήθους διαφορετικών απαντήσεων για όλες τις προτάσεις.

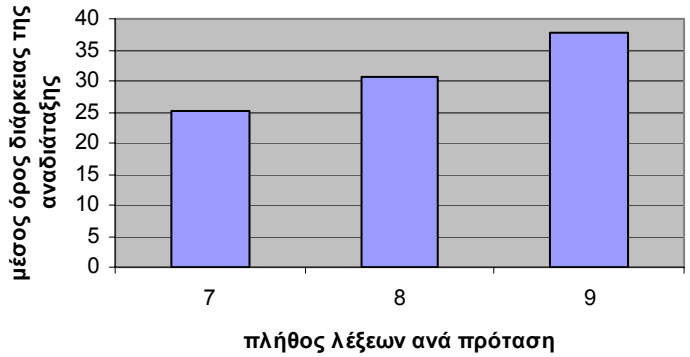
Αυτό που διαπιστώνεται είναι ότι υπάρχει εξάρτηση μεταξύ του πλήθος των λέξεων και του πλήθους των διαφορετικών απαντήσεων. Αυτό φαίνεται πιο καθαρά βλέποντας τον μέσο όρο του πλήθους διαφορετικών απαντήσεων συναρτήσει του πλήθους των λέξεων που περιέχει η πρόταση (βλέπε Σχήμα 4.32).



Σχήμα 4.32 Μ.Ο διαφορετικών απαντήσεων ανά μήκος πρότασεων.

Συνολικά και για τις 150 πρότασεις έχουμε έναν μέσο όρο 5.06 διαφορετικών απαντήσεων ανά πρόταση. Αυτό είναι ένδειξη του ότι σε μία πρόταση ορισμένες λέξεις ή σύνολά λέξεων μπορούν να αλλάξουν θέση χωρίς η πρόταση να στερείται νοήματος, αλλά αλλάζοντας ελαφρά το ύφος ή τον όρο στον οποίο δίδεται έμφαση.

Κατά τη διάρκεια του πειράματος χρονομετρήθηκε επίσης η διάρκεια της ολοκλήρωσης της αναδιάταξης των λέξεων. Υπολογίστηκαν οι μέσοι όροι (σε secs) για προτάσεις 7, 8 και 9 λέξεων και προέκυψε όπως αναμένονταν, ότι ο χρόνος απάντησης μιας πρότασης εξαρτάται από το πλήθος των λέξεων της, όπως φαίνεται και από το παρακάτω Σχήμα 4.33.



Σχήμα 4.33 Ο Μ.Ο χρόνου αναδιάταξης λέξεων για τους συμμετέχοντες (σε secs).

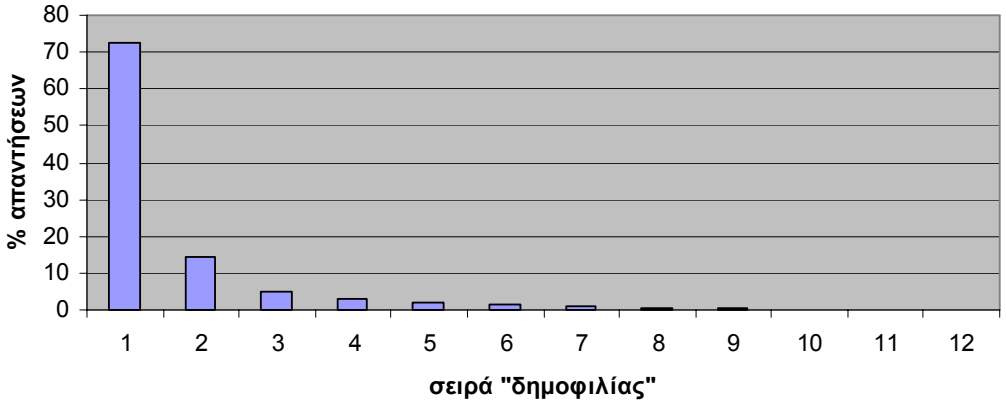
Αυτό που φαίνεται είναι ότι με την αύξηση του πλήθους των λέξεων ανά πρόταση αυξάνεται και ο χρόνος που χρειάζεται ένας άνθρωπος για να τις αναδιατάξει.

4.6.4 Αποτελέσματα αξιολόγησης του συστήματος από τους χρήστες

Στην περίπτωση μιας πρότασης εισόδου, η έξοδος του συστήματος αποτελείται από μία λίστα προτάσεων, που συνοδεύονται από ένα σκορ σύμφωνα με το γλωσσικό μοντέλο. Για την ανάλυση στο πείραμα αυτό χρησιμοποιήθηκαν οι 10 πιο πιθανές αναδιατάξεις της πρότασης εισόδου με το σκεπτικό ότι 10 αναδιατάξεις γενικώς καλύπτουν ένα μεγάλο μέρος των αποδεκτών διαφοροποιήσεων προτάσεων μήκους από 7 μέχρι 9 λέξεις. Η πρώτη πρόταση είναι η πιο πιθανή επιλογή για το σύστημα και ακολουθούν με σειρά φθίνουσας πιθανότητας οι υπόλοιπες.

Το ερώτημα που γεννάται είναι κατά πόσον οι χρήστες αυτής της πειραματικής άσκησης συμφωνούν με τις 10 καλύτερες προτάσεις του συστήματος. Ποιες από τις 10 καλύτερες προτάσεις βρίσκονται στις προτιμήσεις των χρηστών; Στην παρούσα περίπτωση δεν χρησιμοποιήθηκαν οι 10 καλύτερες προτάσεις αυτές απευθείας για αξιολόγηση από τους χρήστες αλλά οι χρήστες αφεθεισαν ελεύθεροι να δημιουργήσουν προτάσεις με συγκεκριμένες λέξεις. Ο τελικός στόχος είναι να δείξουμε την συσχέτιση των αποτελεσμάτων των χρηστών και του συστήματος. Περισσότερες κοινές προτάσεις μεταξύ του συστήματος και των χρηστών περισσότερο ικανό σύστημα να παράγει νοηματικά ορθές προτάσεις, περισσότερες κοινές προτάσεις στην πρώτη προτίμηση των χρηστών περισσότερο αξιόπιστο το σύστημα.

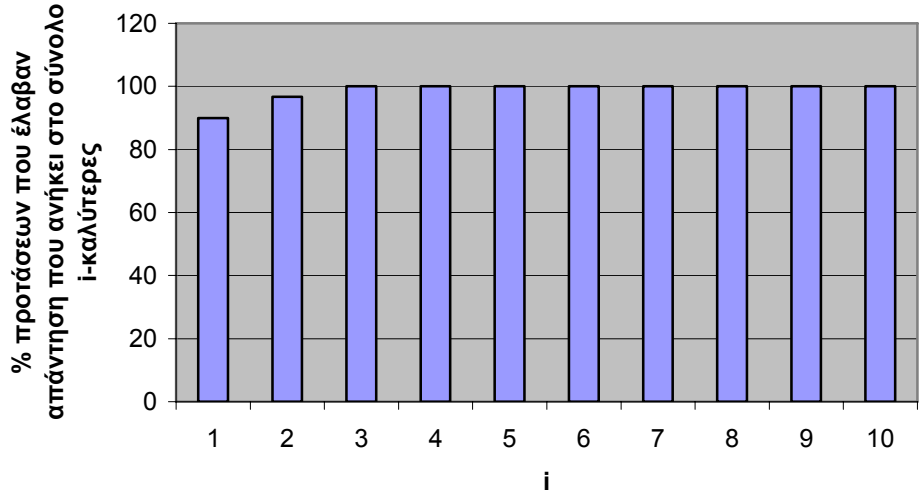
Στο σημείο αυτό πρέπει να εστιάσουμε το ενδιαφέρον μας στο κατά πόσο οι απαντήσεις των συμμετεχόντων συγκεντρώνονται σε συγκεκριμένες ομάδες απαντήσεων και ποια η βαρύτητα τους. Επεξεργαζόμενοι τις 5034 διαφορετικές απαντήσεις, αυτές χωρίστηκαν σε 12 ομάδες φθίνουσας «δημοφιλίας»:



Σχήμα 4.34 Το ποσοστό των απαντήσεων σε σχέση με τις ομάδες δημοφιλίας.

Από το παραπάνω φαίνεται καθαρά ότι η πιο «δημοφιλής» ομάδα απαντήσεων συγκεντρώνει την συντριπτική πλειοψηφία των απαντήσεων (το 72.7%, δηλαδή 3661 απαντήσεις).

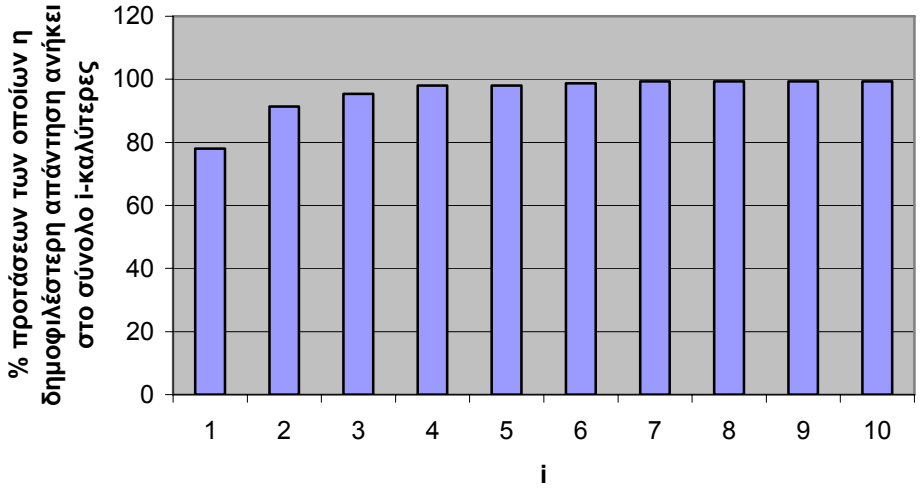
Στο επόμενο σχήμα αποτυπώνεται το διάγραμμα που δείχνει το ποσοστό των προτάσεων από τις 150 για τις οποίες έχει δοθεί απάντηση που ανήκει στο σύνολο των i-καλύτερων, όπου το i κινείται από το 1 μέχρι και το 10.



Σχήμα 4.35 Το ποσοστό των προτάσεων-ερωτήσεων, όπου οι συμμετέχοντες περιλαμβάνουν τις i-καλύτερες στις απαντήσεις τους.

Φαίνεται ότι το 90% των προτάσεων-ερωτήσεων έχουν λάβει σαν απάντηση την πρόταση που το σύστημα προτείνει ως καλύτερη, και το 100% των προτάσεων έχουν λάβει απάντηση που ανήκει στο σύνολο των τριών πρώτων καλύτερων σύμφωνα με το σύστημα.

Όπως φάνηκε πιο πάνω, η δημοφιλέστερη απάντηση αναφορικά με τους συμμετέχοντες αντιπροσωπεύει μεγάλο τμήμα των απαντήσεων. Άρα είναι ενδιαφέρον να δείξουμε πώς το σύστημα αξιολογεί την δημοφιλέστερη πρόταση δηλ. το ποσοστό των 150 προτάσεων στο οποίο η δημοφιλέστερη απάντηση ανήκει στο σύνολο των i-καλύτερων, όπου το i κινείται από το 1 μέχρι και το 10.



Σχήμα 4.36 Το ποσοστό των προτάσεων-ερωτήσεων, όπου οι δημοφιλέστερες ομάδες περιλαμβάνουν τις i-καλύτερες στις απαντήσεις τους.

Για το 80% περίπου των προτάσεων η δημοφιλέστερη απάντηση ταυτίζεται με την πρώτη καλύτερη εναλλακτική του συστήματος και ήδη για i = 4 έχουμε ότι για το 98% των προτάσεων εισόδου η δημοφιλέστερη απάντηση ανήκει στις 4 πρώτες καλύτερες σύμφωνα με το σύστημα. Για ένα μικρό ποσοστό της τάξης του 0,6% των αρχικών προτάσεων, η δημοφιλέστερη απάντηση δεν συμπεριλαμβάνεται στις 10 καλύτερες που προτείνονται από το σύστημα.

4.7 Συμπεράσματα

Στο κεφάλαιο 4 παρουσιάσθηκε ο αλγόριθμος ανίχνευσης και διόρθωσης προτάσεων με λάθη στην σειρά των λέξεων με την χρήση των N-grams. Το πλεονέκτημα χρήσης του αλγορίθμου αυτού είναι η μέθοδος γρήγορης αναζήτησης της βέλτιστης λύσης. Αναλύοντας τα πειραματικά δεδομένα φαίνεται ότι η μέθοδος στο σύνολο της αλλά και ειδικότερα η μέθοδος της γρήγορης

αναζήτησης λειτουργεί αποδοτικά για όλα τα πειραματικά δεδομένα, τόσο για την Αγγλική όσο και για την Ελληνική γλώσσα. Αποδείχθηκε ότι η μέθοδος γρήγορης αναζήτησης μπορεί να μειώσει τον μέσο όρο των αντιμεταθέσεων από 479,001,600 σε 11,378,400 (συντελεστής κέρδους πολυπλοκότητας 42,10) για τα πειραματικά δεδομένα της Αγγλικής γλώσσας και από 479,001,600 σε 6,502,661, (συντελεστής κέρδους πολυπλοκότητας 73,66) στην αντίστοιχη περίπτωση όπου χρησιμοποιούνται δεδομένα για την Ελληνική γλώσσα. Η περίπτωση χρήσης ενός κατωφλίου στον πίνακα αντιστοίχισης, για την περαιτέρω μείωση των αντιμεταθέσεων εξετάστηκε αλλά έδειξε ότι δεν μπορεί να εφαρμοσθεί αφού μειώνεται σημαντικά και ο αριθμός των προτάσεων που μπορούν να διορθωθούν. Για την αξιολόγηση του συστήματος χρησιμοποιήθηκαν 3 διαφορετικά πειραματικά δεδομένα. Στην περίπτωση των πειραματικών δεδομενων του TOEFL, η αξιολόγηση του προτεινόμενου συστήματος διόρθωσης προτάσεων έδειξε ότι 363 προτάσεις (90,75%) επελέγησαν σωστά ενώ σε 37 προτάσεις τα αποτελέσματα δεν ήταν τα αναμενόμενα (9,25%). Για την Αγγλική γλώσσα, η πειραματική διαδικασία αποδεικνύει ότι το 87,40% των προτάσεων βρέθηκαν στην πρώτη θέση της λίστας των 10 καλύτερων ενώ το 0,07% βρίσκεται εκτός των 10 καλύτερων. Στην περίπτωση των πειραματικών δεδομένων για την Ελληνική γλώσσα αποδείχθηκε το 92,63% των προτάσεων βρέθηκαν στην πρώτη θέση ενώ το 0,01% βρίσκεται εκτός των 10 καλύτερων. Στο σημείο αυτό πρέπει να τονίσουμε ότι η προτεινόμενη μέθοδος αποδίδει πολύ καλύτερα σε σχέση με το λογισμικό Microsoft Office ® Word. Αναφορικά με την πειραματική άσκηση αξιολόγησης της μεθόδου από ομάδα χρηστών μέσω διαδικτύου, τα ευρήματα δείχνουν ότι για το 90% των προτάσεων δόθηκε απάντηση από τους χρήστες που το σύστημα προτείνει ως την πρώτη καλύτερη, ενώ για το 100% των προτάσεων δόθηκε απάντηση που ανήκει στο σύνολο των τριών πρώτων καλύτερων σύμφωνα με το σύστημα.

Κεφάλαιο 5

5 Συμπεράσματα και προτάσεις για περαιτέρω έρευνα

Η διατριβή αυτή ασχολήθηκε με την εφαρμογή των N-grams σε θέματα επεξεργασίας φωνής και κειμένου. Ο ρόλος τους είναι πολύ σημαντικός για μια σειρά από εφαρμογές της γλωσσικής τεχνολογίας, όπως η αναγνώριση φωνής, η οπτική αναγνώριση χαρακτήρων, η μηχανική μετάφραση και ακόμη η ορθογραφική διόρθωση. Με την παρούσα εργασία προτάθηκαν δυο νέοι αλγόριθμοι εφαρμογής των N-grams μοντέλων στην αναγνώριση συναισθηματικού λόγου και στην διόρθωση κειμένων. Στον τομέα της αναγνώρισης συναισθηματικού λόγου, χρησιμοποιήθηκε μια νέα τεχνική εμπλουτισμού του γλωσσικού μοντέλου με κείμενα συναισθηματικού περιεχόμενου. Στον τομέα της διόρθωσης κειμένων, προτάθηκε και εξετάσθηκε μια νέα τεχνική αναδιάταξης λέξεων που βασίζεται στον αλγόριθμο γρήγορης αναζήτησης της καλύτερης πρότασης. Συνοψίζοντας τις κυριότερες δράσεις που αναπτύχθηκαν σε αυτήν την εργασία μπορούμε να διακρίνουμε δυο βασικούς άξονες:

- *Ο εμπλουτισμός του γλωσσικού μοντέλου (N-grams) με συναισθηματικά δεδομένα.*

- *Χρήση διγραμμάτων και τριγραμμάτων για την εύρεση της βέλτιστης λύσης.*

Η αναγνώριση της γλωσσικής πληροφορίας του συναισθηματικά χρωματισμένου λόγου αποτελεί μια πρόκληση που έχει ήδη αρχίσει να κεντρίζει το ενδιαφέρον των ερευνητών με σκοπό την φιλικότερη επικοινωνία ανθρώπου-υπολογιστή. Τα ποσοστά επιτυχίας των υπαρχόντων συστημάτων αναγνώρισης φωνής είναι αρκετά χαμηλά για σήματα φωνής που έχουν έντονο συναισθηματικό χρώμα. Όπως είδαμε και παραπάνω, την αναγνώριση φωνής συναισθηματικά χρωματισμένου λόγου επηρεάζουν δυο παράγοντες: ο πρώτος αφορά την προσωδία, και ο δεύτερος παράγοντας που διερευνήθηκε σε αυτή την εργασία αφορά την χρήση

ενός γλωσσικού μοντέλου που προσαρμόζεται περισσότερο στην γλωσσική δομή παρά στα παρα-γλωσσικά φαινόμενα του συναισθηματικού λόγου.

Κατά αυτόν τον τρόπο προτείνεται μια νέα στρατηγική, η οποία βασίζεται στο γεγονός ότι το συναίσθημα δεν επηρεάζει μόνο χαρακτηριστικά της φωνής αλλά και την δομή της γλώσσας. Βάσει αυτής της στρατηγικής ανιχνεύονται σώματα κειμένων τα οποία αντιπροσωπεύουν γλώσσα με συναισθηματικό περιεχόμενο, ώστε να χρησιμοποιηθούν για την εκμάθηση του συναισθηματικά προσανατολισμένου γλωσσικού μοντέλου. Άλλωστε η μεταγραφή συναισθηματικά επηρεασμένου λόγου είναι μια επίπονη και χρονοβόρα διαδικασία. Ο προτεινόμενος αλγόριθμος δημιουργεί σώμα κειμένου με έντονο συναισθηματικό χαρακτήρα από ένα κανονικό σώμα κειμένου, όπως το Εθνικό Βρετανικό Σώμα κειμένου. Στο σημείο αυτό πρέπει να τονίσουμε τον καθοριστικό ρόλο του λεξιλογίου με συναισθηματικούς όρους ώστε να εξαχθούν οι συναισθηματικά επηρεασμένες προτάσεις από το Εθνικό Βρετανικό Σώμα κειμένου. Το συναισθηματικά προσανατολισμένο γλωσσικό μοντέλο υπολογίζεται από ένα νέο σώμα κειμένου που προέρχεται από την προσάρτηση των προτάσεων με συναισθηματικό περιεχόμενο στο αρχικό σώμα κειμένου. Για την προσάρτηση των προτάσεων χρησιμοποιήθηκε συντελεστής επανάληψης λ. Πειραματικά υπολογίσθηκε ότι η βέλτιστη τιμή του είναι το δέκα. Τα πειραματικά αποτελέσματα δείχνουν ότι με την κατάλληλη εκπαίδευση του γλωσσικού μοντέλου με συναισθηματικά δεδομένα μπορεί να βελτιωθεί η απόδοση ενός συμβατικού συστήματος αναγνώρισης φωνής κατά 20%, όταν χρησιμοποιείται σαν είσοδος σήμα με συναισθηματικό λόγο. Τέλος θα πρέπει να επισημανθεί ότι το επαυξημένο γλωσσικό μοντέλο δεν αλλοιώνει τα αποτελέσματα του συστήματος αναγνώρισης φωνής σε περίπτωση χρήσης σημάτων φωνής με υπαγορευτικό λόγο.

Ένα σημαντικό σημείο περαιτέρω επέκτασης αυτής της μεθόδου μπορεί να αποτελέσει ο συνδυασμός των N-grams με τεχνικές σημασιολογικής ανάλυσης. Αυτό που θα ήταν εξαιρετικά ενδιαφέρον να διερευνηθεί είναι το κατά πόσον η απόδοση των συμβατικών μοντέλων εξομάλυνσης των N-grams μπορεί να βοηθηθεί από την χρήση αλγόριθμων σημασιολογικής ανάλυσης. Γίνεται κατανοητό ότι τα λεγόμενα ενός ομιλητή μπορεί να μοντελοποιηθούν καλύτερα όταν γνωρίζουμε και την σημασιολογική ομοιότητα των λέξεων εκτός από τους συνδυασμούς τους.

Αναφορικά με την προτεινόμενη μέθοδο διόρθωσης κειμένων θα πρέπει να επισημανθεί ότι μπορεί να εφαρμοσθεί σε οποιαδήποτε γλώσσα και για οποιοδήποτε σύνολο λέξεων μιας που το σύστημα έχει εκπαιδευτεί πάνω σ' ένα μεγάλο όγκο λέξεων. Αναφορικά με άλλα συστήματα που αναφέρονται στην βιβλιογραφία πρέπει να τονιστεί ότι ενώ η απόδοση τους μεγιστοποιείται για ένα περιορισμένο σύνολο λέξεων, δεν συμβαίνει το ίδιο όταν εφαρμόζονται σε προτάσεις με λέξεις που δεν ανήκουν σε αυτό. Ένα άλλο σημαντικό πλεονέκτημα που

κατέχει η μέθοδος αυτή είναι η ανεξαρτησία της από την χρήση κάποιου συντακτικού αναλυτή. Αυτό υποδηλώνει ότι μπορεί να εφαρμοσθεί σε οποιαδήποτε γλώσσα έστω και αν για αυτήν δεν υπάρχει κάποιος διαθέσιμος συντακτικός αναλυτής. Επίσης η χειρονακτική συλλογή γραπτών κανόνων δεν είναι απαραίτητη, από την στιγμή που αυτοί κωδικοποιούνται με το στατιστικό γλωσσικό μοντέλο. Παράλληλα με το προηγούμενο επιχείρημα θα πρέπει να τονιστεί ότι δεν απαιτείται πολυδάπανη και επίπονη συλλογή ανάλογων λαθών ώστε να δημιουργηθούν αντίστοιχα πρότυπα. Εν τέλει, θα πρέπει να σημειώσουμε ότι η απόδοση της μεθόδου δεν εξαρτάται από συγκεκριμένα γραμματικά πρότυπα που άλλωστε ποικίλουν από γλώσσα σε γλώσσα και γι' αυτό μπορεί να εφαρμοσθεί σε οποιαδήποτε γλώσσα ανεξάρτητα του τρόπου διάταξης των λέξεων. Η εφαρμογή της μεθόδου σε όλες τις γλώσσες εξασφαλίζεται από το γεγονός της χρήσης μόνο του στατιστικού γλωσσικού μοντέλου.

Βάσει των πειραματικών διαδικασιών η προτεινόμενη μέθοδος είναι αποτελεσματική σε ότι αφορά την διόρθωση λανθασμένων προτάσεων. Είναι ενδιαφέρον να επισημάνουμε ότι το σύστημα δεν ανιχνεύει μόνο λάθη αλλά και διορθώνει προτάσεις. Ο αλγόριθμος γρήγορης αναζήτησης μειώνει τον αριθμό των αντιμεταθέσεων σε μεγάλο βαθμό με αποτέλεσμα το σύστημα διόρθωσης προτάσεων να λειτουργεί αξιόπιστα για διαφορετικά δεδομένα δοκιμής τόσο για την Αγγλική όσο και για την Ελληνική γλώσσα. Τα ευρήματα δείχνουν ότι κατά μέσο όρο το 90,10% των προτάσεων μπορούν να διορθωθούν με την χρήση της προτεινόμενης μεθόδου. Σε αυτό το σημείο πρέπει να τονίσουμε ότι ο ρόλος της ποιότητας του γλωσσικού μοντέλου είναι καθοριστικός για την απόδοση του συστήματος και υπάρχει ανάγκη για εκπαίδευση του γλωσσικού μοντέλου από ποικιλία κειμένων. Με την χρήση της μεθόδου γρήγορης αναζήτησης (φιλτράρισμα των αντιμεταθέσεων), το σύστημα επιτυγχάνει ταχύτερη απόκριση κατά 73,66 φορές (Ελληνική γλώσσα), ενώ για την Αγγλική γλώσσα, 42,10 φορές. Επιπλέον φαίνεται ότι η απόδοση του συστήματος αξιολογείται θετικά και από ομάδα 324 χρηστών. Οι απαντήσεις τους συμφωνούν πάντοτε με μια τουλάχιστον από τρεις πρώτες καλύτερες απαντήσεις του συστήματος.

Τα μελλοντικά σχέδια βελτίωσης της μεθόδου αυτής αφορούν την δοκιμή και άλλων στατιστικών γλωσσικών μοντέλων που όμως συμπεριλαμβάνουν πληροφορία σχετική με POS. Η χρήση αυτής της πληροφορίας είναι σίγουρο ότι θα δώσει καλύτερα αποτελέσματα για τον λόγο ότι συνδυασμοί λέξεων που δεν υπάρχουν λόγω της σπανιότητας δεδομένων θα μπορούν να καλυφθούν από την χρήση συνδυασμών κατηγοριών λέξεων. Όπως είδαμε και στην επισκόπηση διαφορετικών ειδών γλωσσικών μοντέλων, μοντέλα παράλειψης και πρόκλησης μπορούν να συνδράμουν στην καλύτερη απόδοση του συστήματος. Αναφορικά με την βελτίωση της απόδοσης της μεθόδου της γρήγορης αναζήτησης της βέλτιστης λύσης μπορεί να εξετασθεί και η χρήση τεχνικών διαμελισμού μιας πρότασης σε υπό-προτάσεις (χρήση

chunker). Ενδιαφέρον παρουσιάζει και η μελέτη του κατά πόσον τα μοντέλα N-grams μπορούν να συνεισφέρουν στην διόρθωση και άλλων γραμματικών λαθών όπως π.χ ασυμφωνία υποκείμενου-ρήματος.

6 ΒΙΒΛΙΟΓΡΑΦΙΑ

ΚΕΦΑΛΑΙΟ 1

1) Chomsky, N. (1964).current issues in linguistic theory. (The Hague: Mouton).
2) Bakamidis, S., Carayannis, G., (1987). 'Phonemia' A Phoneme Transcription System for Speech Synthesis in Modern Greek, Speech Communication 6, pp. 159-169
3) Jurafsky, D., and Martin J. H., (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Prentice-Hall.
4) Manning, C. D., and Schutze, H . (1999). Foundations of statistical natural Lnaguage processing. MIT press. Cambrigde, MA.
5) Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pp. 1-32. Oxford: Philological Society. Reprinted in F.R. Palmer (ed.), *Selected Papers of J.R. Firth 1952-1959*, London: Longman (1968).
6) Μαΐστρος, Γ., και Μαρκαντωνάτου, Σ., Παράσταση και Επεξεργασία Φυσικής Γλώσσας.
7) Μπαμπινιώτης Γ. (1998) Θεωρητική Γλωσσολογία. Αθήνα.

ΚΕΦΑΛΑΙΟ 2

8) ASSOCIATED PRESS http://www.ap.org/
9) Bahl, L., Jelinek, F., and Mercer, R., (1983). (The first LM model in speech) A maximum likelihood approach to continuous speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 5(2):179-190.
10) Baum, L. E., (1972). An inequality and associated maximazation technique in statistical estimation for probabilistic functions of Markov processes. In Shisha, O. (Eds), Inequalities III: Processings of the third Symposium on Inequalities, University of California, Los Angeles, pp. 1-8. Academic Press.
11) Beeferman, D., Berger, A., and Lafferty, J. (1997). A model of lexical attraction and repulsion. In *ACL/EACL '97*.
12) Bellegarda, J., (2000). Exploiting latent semantic information in statistical language modeling, Proc. IEEE, vol. 88, no. 8, 1279--1296,.

13) Brown, P F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin. P. S. (1990) A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

14) Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C., and Mercer, R. L., (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467-479.

15) Brown, P., Lai, J. C., and Mercer, R. (1991). Aligning Sentences in Parallel Corpora. In *Proceedings of ACL-91*, Berkeley CA.

16) Charniak, E., (2001). Immediate-head parsing for language models. In Proc. Assoc. for Computational Linguistics (ACL), 116--123.

17) Chelba, C. and Jelinek, F., (1998). Exploiting syntactic structure for language modelling. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montreal, Canada, August, pp. 225-231.

18) Chen S. F. and Goodman. J.T., (1998). An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group, Harvard University.

19) Chomsky, N. (1957). *Syntactic Structures*. (The Hague: Mouton).

20) Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

21) Church, K. W. and Gale, W. A., (1991). *Concordances for parallel text*. In Proceedings of the Seventh Annual Conference of the UW Center for the New OED and Text Research, pp. 40-62.

22) Church, K. W., Gale, W. A., (1991). A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. In Computer Speech and Language, 19-54.

23) Church, K. W., Gale, W. A., and Kruskal J. B., (1991). Appendix A: the Good-Turing theorem. In Computer Speech and Language (Church and Gale 1991), 19-54.

24) Cover, T. M. and Thomas, J., A. (1991). Elements of information theory. Wiley, New York.

25) Darroch, J. N., and Ratcliff D., Generalized Iterative Scaling for Log-Linear Models. *The Annals of Mathematical Statistics* . 1972. Vol. 43, No. 5, 1470-1480.

26) Dempster, A.P., Laird, N.M., & Rubin, D. *Maximum-likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society, Series B, 39. di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolati, G. (1992).

Understanding motor events: A neurophysiological study. Experimental Brain Research, 91, 176--180.

27) Good, I, J, (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*,40(3 and 4):237–264, 1953.

28) DOW JONES http://www.dowjones.com/

29) Hull, J. (1992). *Combining syntactic knowledge and visual text recognition: A hidden Markov model for part of speech tagging in a word recognition algorithm*. In AAAI Symposium: Probabilistic Approaches to Natural Language, pp 77-83.

30) Iyer, R.M., and Ostendorf, M., (1999). Modeling long distance dependence in language: topic mixtures versus dynamic cache models Speech and Audio Processing, IEEE Transactions on , Volume: 7 Issue: 1 , 30 -39.

31) Jeffreys, H., (1948). Theory of Probability. Clarendon Press, Oxford

32) Jelinek, F., and Mercer, R. L., (1980). Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, pp. 381–397, The Netherlands: North-Holland, Amsterdam.

33) Jelinek, F. (1995). The language modeling summer workshop at Johns Hopkins University. Closing remarks.

34) Jelinek, F., Merialdo, B., Roukos, S., and Strauss, M., (1991). A Dynamic LM for Speech Recognition Proceedings ARPA workshop on Speech and Natural Language, pp. 293-295.

35) Jelinek, F., Statistical Methods for Speech Recognition, MIT Press, 1997.

36) Johnson, W. E., (1932). Probability: deductive and inductive problems. Mind 41:421-423.

37) Katz. S. M., (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3):400–401.

38) Kernighan M. D., Church, K. W., and Gale, W. A. (1990). A spelling correction program base on a noisy channel model. In COLING-90, Helsinki, Vol. II, pp. 205-211.

39) Kuhn, R., (1988). Speech recognition and the frequency of recently used words: A modified markov model for natural language. In *12th International Conference on Computational Linguistics*, pp. 348–350.

40) Kuhn, R., and De Mori R., (1990). A cache-based natural language model for speech reproduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-12(6):570–583.

41) Kuhn, R., and De Mori, R., (1992). Roland Kuhn and Renato De Mori. Correction to: A cache-based natural language model for speech reproduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-14(6):691–692.

42) Kupiec, J., (1989). Probabilistic models of short and long distance word dependencies in running text. In *Proceedings of the DARPA Workshop on Speech and Natural Language*, pages 290–295.

43) Lindstone, G. J., (1920). Note on the general case of the Bayes-Laplace formula for inductive or a priori probabilities. IEEE Transactions on Information theory 38:1842-1845.

44) Miller, G. (1995). WORDNET: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

45) Ney, H., Essen, U., and Kneser, R., (1994). On structuring probabilistic dependences in stochastic language modeling. *Computer Speech and Language*, 8:1–38.

46) Rosenfeld, R., (1994). Adaptive Statistical Language Modelling: A Maximum Entropy Approach. PhD thesis, Carnegie Mellon University.

47) Rosenfeld, R., (1996). A Maximum Entropy Approach to Adaptive Statistical Language Modeling, Computer, Speech and Language, vol. 10, 187—228.

48) Shannon, C. E. (1948). A mathematical theory of communication (parts I and II). *Bell System Technical Journal*, XXVII:379-423.

49) Shannon, C. E. (1951). *Prediction and entropy of printed English*. Bell System Technical Journal, 30(1), 50-64.

50) Srihari, R. K., and Baltus, C. M., (1993). Incorporating Syntactic Constraints in Recognizing Handwritten Sentences, in Proc. of the International Joint Conference on Artificial Intelligence (IJCAI-93), Chambery, France, August 1993, pp. 1262--1267.

51) Witten I. H., and Bell, C. T., (1991). The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression, in IEEE Transactions on Information Theory, Vol 37, No. 4, 1085-1094.

ΚΕΦΑΛΑΙΟ 3

52) André, E., Dybkjaer, L., Minker, W., et al. (2004). (Eds) *Affective Dialogue Systems: Tutorial and Research Workshop, ADS 2004*, Kloster Irsee, Germany, June 14-16, 2004. *Lecture Notes in Computer Science 3068 /2004.* Springer-Verlag, Heidelberg.

53) Ang, J., Dhillon, R., Krupski, A., Shriberg, E. & Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog. *Proc. ICSLP 2002*, Denver, Colorado.

54) Batliner A., Fischer K., Huber R., Spilker J. & Nöth E., (2003). How to find trouble in communication. *Speech Communication* 40, 117-143.

55) Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E. (2003). How to find trouble in communication. *Speech Communication*, 40, 117-143.

56) Batliner, A., Hacker, C., Steidl, S., Nöth, E., Haas, J. (2004). From Emotion to Interaction: Lessons from Real Human-Machine-Dialogues. In E. André, .L. Dybkjaer, W. Minker, et al. (Eds.) *Affective Dialogue Systems: Tutorial and Research Workshop, ADS 2004*, Kloster Irsee, Germany, June 14-16, 2004. *Lecture Notes in Computer Science 3068 /2004*. Springer-Verlag, Heidelberg, pp1-12.

57) Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences* 17, 97-110.

58) Boozer, A., Seneff, S. & Spina, M. (2003). Towards recognition of emotional speech in human-computer dialogues. Abstract. MIT Laboratory for Computer Science. www.csail.mit.edu/research/abstracts/abstracts03/ interfaces-applications/03boozer.pdf

59) Boutri, A., and Stalikas, A.(2004). What is the contribution of positive emotions in psychotherapy? 1st International Conference of the Psychological Society of Northern Greece: Quality of Life and Psychology, Thessaloniki, Greece.

60) Campbell, N. (2000). Databases of emotional speech. *Proc. ISCA ITRW Speech and Emotion*, Newcastle, Ireland, Sept. 2000, 34-39.

61) Cowie R. & Cornelius R. 2003. Describing the Emotional States that are Expressed in Speech. *Speech Communication* 40, 5-32.

62) Cowie, R., & Schröder, M. (2005). Piecing Together the Emotion Jigsaw. In S. Bengio & H. Bourlard (Eds.) MLMI 2004, LNCS 3361, Springer-Verlag Berlin Heidelberg, pp. 305-317.

63) Cowie, R., and Cornelius, R. (2003). Describing the Emotional States that are Expressed in Speech. *Speech Communication,* 40, 5-32.

64) Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M. (2000). 'Feeltrace': An instrument for recording perceived emotion in real time. In E. Douglas-Cowie, R. Cowie & M. Schröder (Eds.) *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, Belfast, pp.19-24.

65) Douglas-Cowie, E, Cowie R., and Romano A., (1999). Changing emotional tone in dialogue and its prosodic correlates In *ESCA Tutorial and Research Workshop on Dialogue and Prosody,* Eindhoven, The Netherlands, pp. 41-46.

66) Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P., (2003a). Emotional Speech: towards a new generation of databases. *Speech Communication,* 40, 33-60.

67) Douglas-Cowie, E., et al., (2003b). Multimodal data in action and interaction: a library of recordings and labelling schemes HUMAINE report D5d http://emotion-research.net/deliverables/

68) Ekman, P. & Friesen, W.,(1978). *The Facial Action Coding System*. Consulting Psychologists' Press, San Francisco, CA.

69) Fernandez, R. & Picard, R. W. (2003). Modeling drivers' speech under stress. *Speech Communication* 40 (1-2), Special Issue following the ISCA Workshop on Speech and Emotion, 145-159.

70) Hansen, J. & Bou-Ghazale, S., (1997). Getting started with SUSAS: A Speech Under Simulated and Actual Stress Database. *Proc. Eurospeech 1997*, Rhodes, Greece, vol. 5, 2387-2390.

71) Hansen, J. H. L. & Womack, B. D., (1996). Feature analysis and neural network-based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing* IV(4), 307- 313.

72) Kucera and Francis, W.N. (1967). *Computational Analysis of Present-Day American English.* Providence: Brown University Press.

73) Kwon, O., Chan, K., Hao, J. & L, T-W. (2003). Emotion recognition by speech signals. *Proc. Eurospeech 2003*, Geneva, 125-128.

74) Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press, Philadelphia.

75) Lee, C. & Narayanan, S. (2003). Emotion recognition using a data-driven fuzzy inference system. *Proc. Eurospeech 2003*, Geneva

76) Mitchell, C., Menezes, C., Williams, J., Pardo, B., Erickson, D. & Fujimura, O. (2000). Changes in syllable and boundary strengths due to irritation. *Proc. ISCA ITRW on Speech and Emotion*, 5-7 September 2000, Textflow, Belfast, 98- 103.

77) MPEG http://www.apple.com/quicktime/technologies/mpeg4/

78) Plutchik, R., (1980). Emotion: A Psychoevolutionary Synthesis, New York: Harper & Row.

79) Plutchik, R. & Conte, H. (1997). *Circumplex models of Personality and Emotions*. Washington: APA.

80) Polzin, S.T., & Waibel, A., (1998). Pronunciation variations in emotional speech. In H. Strik, J. M. Kessens & M. Wester (Eds.) *Modeling Pronunciation Variation for Automatic Speech Recognition*. Proc. of the ESCA Workshop, pp. 103-108.

81) Roach, P., Stibbard, R., Osborne, J., Arnfield, S. & Setter, J.(1998). Transcription of prosodic and paralinguistic features of emotional speech. *Journal of IPA* 28, 83-94.

82) Robinson, T., Fransen, J., Pye, D., Foote, J., & Renals, S., (1995). WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, 81–84.

83) Scherer, K.R., (1999).Appraisal theory. In Dalgleish, T. & Power, M. (eds.) Handbook *of Cognition and Emotion*. New York: John Wiley, 637-663.

84) Scherer, K.R., (2004). HUMAINE Deliverable 3c Preliminary plans for exemplars: Theory.

85) Schröder, M., (2004). Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis. PhD thesis, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University.

86) Steeneken, H.J.M., and Hansen, J.H.L., (1999). Speech Under Stress Conditions: Overview of the Effect of Speech Production and on System Performance. *IEEE ICASSP-99: Inter. Conf. on Acoustics, Speech, and Signal Processing* 4, 2079-2082.

87) Tolkmitt, F.J. & Scherer, K.R., (1986). Effects of experimentally induced stress on vocal parameters. *J. Exp. Psychol.: Human Perception and Performance* 12, 302-313.

88) Weizenbaum, J.(1966). ELIZA- A computer program for the study of natural language communication between man and machine. Communications of the Association for Computing Machinery 9, 36-45.

89) Whissell, C., (1989). The dictionary of affect in language. In R. Plutchnik & H. Kellerman (Eds.) *Emotion: Theory and research.* New York, Harcourt Brace, pp. 113-131.

90) Young, S.J. (1996). Large Vocabulary Continuous Speech Recognition. *IEEE Signal Processing Magazine* 13, (5), 45-57.

91) Γαλανάκης Μ. (2003). Ο Ρόλος του Τύπου της Προσωπικότητας και της Βίωσης των Θετικών Συναισθημάτων στην Ομαδική Αποτελεσματικότητα. Αθήνα, Πάντειο Πανεπιστήμιο Κοινωνικών και Πολιτικών Επιστημών. (Μεταπτυχιακή Διατριβή).

92) Σταλίκας, Α., & Μπούτρη, Α., (2004). Θεμελιώδη θέματα ψυχοθεραπείας: Το συναίσθημα στην ψυχοθεραπεία. Ελληνικά Γράμματα.

ΚΕΦΑΛΑΙΟ 4

93) Atwell, E.S., (1987). How to detect grammatical errors in a text without parsing it. In Proceedings of the 3rd EACL, 38–45.

94) Atwell, E and Elliot, S, (1987b). "*Dealing with ill-formed English text"* both in Garside, R, Sampson, G & Leech, G (editors), The computational analysis of English: a corpus-based approach, London, Longman.

95) Bangalore and Knigth (2000). "Exploiting a probabilistic hierarchical model for generation", In Proceedings of COLING/ACL'00. pp: 42 – 48.

96) Bigert, J., Knutsson. O., (2002). Robust error detection: A hybrid approach combining unsupervised error detection and linguistic knowledge. In Proceedings of Robust Methods in Analysis of Natural language Data, (ROMAND 2002), 10–19.

97) BNC http://www.natcorp.ox.ac.uk/corpus/index.xml

98) Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roossin, (1990), A Statistical Approach to Machine Translation, Computational Linguistics, 16(2).

99) Chodorow M., Leacock C., (2000). An unsupervised method for detecting grammatical errors. In Proceedings of NAACL'00, 140–147.

100) Church, K. W., and Mercer, R., (1993). "Introduction to the special issue on computational linguistics using large corpora". Computational Linguistics, 19(1):1--24.

101) Connors, R. L., Lunsford, A. A., (1988). "Frequency of Formal Errors in Current College Writing, or, Ma and Pa Kettle Do Research." *CCC* 39, 395-409.

102) Dologlou, Y., Markantonatou, S., Tambouratzis, G., Yiannoutsou, O., Fourla, S., Ioannou, N. (2003), Using Monolingual Corpora for Statistical Machine Translation: The METIS System. In Proceedings of the EAMT- CLAW'03 Workshop, Dublin, Ireland, 15-17 May, pp. 61-68.

103) Eastwood, J., (1997). Order of place, time and frequency words (never, often), Oxford Practice Grammar Oxford University Press, Oxford. Unit 89.

104) Faigley, L., "The Brief Penguin Handbook, (2nd Edition)", *University of Texas at Austin* Longman 2003

105) Feyton, C. M. (2002). Teaching ESL/EFL with the internet. Merill Prentice- Hall.

106) Folse, K.S., (1997). Intermediate TOEFL Test Practices (rev. ed.). Ann Arbor, MI: The University of Michigan Press.

107) Golding, A., (1995). A Bayesian hybrid for context-sensitive spelling correction. Proceedings of the 3rd Workshop on Very Large Corpora, 39--53.

108) Hawkins, J. A., (1994). A Performance Theory of Order and Constituency. Cambridge, Cambridge University Press.

109) HEARCOM http://www.hearcom.org

110) Heift, T., (1998). Designed Intelligence: A Language Teacher Model, Unpublished Ph.D. Dissertation, Simon Fraser University.

111) Heift, T., (2001). Intelligent Language Tutoring Systems for Grammar Practice. Zeitschrift für Interkulturellen Fremdsprachenunterricht (Online), 6 (2), 15 pp.

112) Izumi, E., Uchimoto, K., Saiga, T., Supnithi, T., Isahara, H., (2003). Automatic error detection in the Japanese learners English spoken data. In Companion Volume to the Proceedings of ACL '03, 145–148.

113) Jelinek, F., (1976). "Continuous Speech Recognition by Statisical Methods". IEEE Proceedings 64:4: 532-556.

114) Katz, S. M., (1987). "Estimation of probabilities from sparse data for the language model component of a speech recogniser". *IEEE Transactions on Acoustics, Speech and Signal Processing, 35*(3),400-401.

115) Koehn P., (2003). "Noun Phrase Translation", PhD thesis, University of Southern California,

116) Langklide, I., and Knight, K., (1998). "Generation that exploits corpus-based statistical knowledge". In Proceedings of COLING/ACL'98. pp: 704 - 710.

117) More, J., (2006), "A grammar Checker based on Web searching" Digithum [online article]. Iss. 8. UOC.

118) Murphy, R., (1990). Order of several describing words together (adjectives), English Grammar in Use Cambridge University Press, Cambridge, Unit 95.

119) Naber, D., (2003), "A rule based style and grammar checker". Bielefeld University

120) Och, F. J., Ney, H., (2000), Statistical Machine Translation. EAMT Workshop, pp. 39-46, Ljubljana, Slovenia.

121) Park, J. C., Palmer, M., and Washburn, G. (1997). An English grammar checker as a writing aid for students of English as a second language, In Proceedings of Conference on Applied Natural Language Process, New Brunswick, NJ.

122) Sjöbergh, J., (2005). Chunking: an unsupervised method to find errors in text, Proceedings of the 15th Nordic Conference of Computational Linguistics, NODALIDA 2005.

123) Stolcke, A., (2002). "SRILM -- An Extensible Language Modeling Toolkit", in ICSLP, Denver, Colorado, USA.

124) Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., Sawaf, H., (1997). Accelerated DP based search for statistical translation", In *EUROSPEECH-1997*, 2667-2670.

125) Tillmann, C., Vogel, S., Ney, H., and Zubiaga, A., (2004). A DP Based Search Using Monotone Alignments in Statistical Translation, Proceedings of the Conference of the Association for Computational Linguistics (ACL).

126) TOEFL Test of English as a Foreign Language, Educational Testing Service

127) Yamada, K. and Knight, K., (2001). A Syntax-Based Statistical Translation Model, Proc. of the Conference of the Association for Computational Linguistics ACL 2001.

128) Young, S.J., (1996). Large Vocabulary Continuous Speech Recognition, IEEE Signal Processing Magazine 13, (5), 45-57.

129) Way, A., (2003). Translating with Examples: The LFG-DOT Models of Translation. In M. Carl and A. Way (eds.), Recent Advances in Example-Based Machine Translation, Kluwer Academic Publishers.

130) Vogel, S., Ney, H., and Tillman, C., (1996), HMM-Based Word Alignment in Statistical Translation, Proceedings of the International Conference on Computational Linguistics (COLING)

7 ΔΗΜΟΣΙΕΥΣΕΙΣ

7.1 Δημοσιεύσεις σε συνέδρια

1. Fotinea, S-E., Bakamidis, S., Athanaselis, T., Dologlou, I., Carayannis, G., Cowie, R., Douglas-Cowie, E., Fragopanagos, N., Taylor, J.G. (2003). Emotion in Speech: towards an integration of linguistic, paralinguistic and psychological analysis**,** Lecture Notes in Computer Science (LNCS), Eds. O. Kaynak et al., Springer-Verlag Berlin Heidelberg (ISSN: 0302-9743), Vol. 2714 / 2003, 1125-1132.
2. Athanaselis, T., Fotinea, S-E., Bakamidis, S., Dologlou, I., Giannopoulos, G. (2003). Signal Enhancement for Continuous Speech Recognition**,** Lecture Notes in Computer Science (LNCS), Eds. O. Kaynak et al., Springer-Verlag Berlin Heidelberg (ISSN: 0302-9743), Vol. 2714 / 2003, 1117-1124.
3. Athanaselis, T., Bakamidis, S., Fotinea, S-E., Dologlou, I., Fragapanagos, N., Taylor, J.G., Cowie, R., Douglas-Cowie, E. (2003). Impact of Speech Enhancement on ASR Confidence Score, in Proceedings of EUNITE 2003, Oulu, Finland, 548-552.
4. Fragopanagos, N., Taylor, J.G., Cowie, R., Douglas Cowie E., Athanaselis, T., Fotinea, S-E., Bakamidis, S., Dologlou, I. (2003). Detecting Moving Emotion, in Proceedings of EUNITE 2003, Oulu, Finland, 542 – 547.
5. Athanaselis, T., Bakamidis, S., Fotinea, S-E., Dologlou, I. (2004). Impact of Speech Enhancement on ASR Time Stamping, in Proceedings of EUNITE 2004, Aachen, Germany, 452 – 457.
6. Athanaselis, T., Bakamidis, S., Dologlou, I. (2005). Improving speech recognition performance in noisy environments, in Proceedings of the 7th Hellenic-European Conference on Computer Mathematics and its Applications (HERCMA'2005), Athens, Greece.
7. Athanaselis, T., Bakamidis, S., Dologlou, I. (2006). Words Reordering based on Statistical Language Model, in Proceedings of the Transactions on Engineering, Computing and Technology, ICCS'06 Vienna, Austria, Volume 12, March 29-31, 270-273.
8. Athanaselis, T., Bakamidis, S., Dologlou, I. (2006). Automatic Recognition of Emotionally Coloured Speech, in Proceedings of the Transactions on Engineering,

Computing and Technology, ICCS'06 Vienna, Austria, Volume 12, March 29-31, 274-277.

9. Athanaselis, T., Bakamidis, S., Dologlou, I. (2006). An automatic method for revising ill-formed sentences based on N-grams, in Proceedings of the 3rd International conference on Speech Prosody, Dresden, Germany, 370-373.
10. Athanaselis, T., Bakamidis, S., Dologlou, I. (2006). A New Approach for Words Reordering Based On Statistical Language Model, in Proceedings of the 11-th International Conference Speech and Computer (SPECOM), 463-466.
11. Athanaselis, T., Bakamidis, S., Dologlou, I. (2006). Recognising verbal content of emotionally coloured speech, in Proceedings of the 14th European Signal Processing Conference (EUSIPCO).
12. Athanaselis, T., Bakamidis, S., Dologlou, I. (2006). A fast algorithm for words reordering based on language model., Lecture Notes in Computer Science (LNCS), Eds. S. Kollias et al., Springer-Verlag Berlin Heidelberg, (ISSN: 0302-9743), Volume 4132/2006, 943-951.
13. Athanaselis, T., Bakamidis, S., Dologlou, I. (2006). A comparative study of ASR performance in different emotional states, in Proceedings of the XXVIII-th International Congress of Audiology, Innsbruck, Austria, 98-102.
14. Athanaselis, T., Bakamidis, S., Dologlou, I. (2006). Impact of the cocktail party effect on the confidence accuracy of emotional speech recogniser, in Proceedings of the 18th BeNeLux Conference on Artificial Intelligence , (BNAIC), Namour, Belgium.
15. Athanaselis, T., Bakamidis, S., Dologlou, I. (2006). A Statistical Method for Correcting Word Order Errors in Greek Texts, in Proceedings of the 2nd international conference of IASTED on Computational Intelligence (CI), San Francisco, 502-506
16. Athanaselis, T., Bakamidis, S., Dologlou, I. (2006). A novel technique for words reordering based on n-grams, in Proceedings of the International Symposium on Signal Processing and its Applications *in conjunction* with the International Conference on Information Sciences, Signal Processing and its Applications, Sharjah, United Arab Emirates (U.A.E.).

7.2 Δημοσιεύσεις σε περιοδικά με κριτές

1. Athanaselis, T., Bakamidis, S., Dologlou, I., Cowie, R., Douglas-Cowie, E., and Cox, C. (2005). "ASR for emotional speech: clarifying the issues and enhancing performance", Special issue: Emotion and brain, Neural Networks, Elsevier Publications, Volume 18, Issue 4, 437- 444.
2. Athanaselis, T., Bakamidis, S., Dologlou, I., Mamouras, K. (2006). "N-grams: A Tool for Repairing Word Order Errors in ill-formed Texts" International Journal of Signal Processing, Volume 3 Number 2, 123-128.

Βιογραφικό σημείωμα

Ο Θεολόγος Αθανασέλης του Δημητρίου γεννήθηκε στην Μυτιλήνη, Λέσβου στις 28 Ιουνίου 1976. Αποφοίτησε από το 3^{ο} Γυμνάσιο Μυτιλήνης τον Ιούνιο του 1994. Την ίδια χρονιά εισήλθε με τις εισαγωγικές εξετάσεις στο τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Η.Υ, της Πολυτεχνικής σχολής Ξάνθης, του Δημοκριτείου Πανεπιστημίου Θράκης. Το έτος 1999 ολοκλήρωσε τις προπτυχιακές του σπουδές του και το 2000 απέκτησε το μεταπτυχιακό τίτλο ειδίκευσης, MSc in Engineering and Physical Science in Medicine, του τμήματος Βιοϊατρικής, του πανεπιστημίου του Λονδίνου, Imperial College of Science, Technology and Medicine. Το Νοέμβριο του 2001 έγινε δεκτός από την σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Η.Υ του Ε.Μ.Π για εκπόνηση διδακτορικής διατριβής στον τομέα Σημάτων, Ελέγχου και Ρομποτικής. Παράλληλα από το ίδιο έτος εργάζεται στο Ινστιτούτο Επεξεργασίας του Λόγου ως μηχανικός πληροφορικής.

Συμμετοχή σε ερευνητικά-αναπτυξιακά προγράμματα:

- "IST-ERMIS" (Emotionally Rich Man-machine Intelligent System) (2002-2004) συμμετοχή σε θέματα αναγνώρισης φωνής από συναισθηματικά δεδομένα.
- "IST-HEARCOM" (Hearing in the communication society) συμμετοχή σε θέματα που σχετίζονται με την ενσωμάτωση του συστήματος αναγνώρισης φωνής σε περιβάλλοντα εξυπηρετητή-χρήστη για Α.Μ.Ε.Α.
- "IST-AGENT_DYSL" (Accommodative intelliGENT educational environments for DYSLexic learners) συμμετοχή σε θέματα που σχετίζονται με την προσαρμογή του συστήματος αναγνώρισης φωνής σε άτομα με μαθησιακές δυσκολίες.
- "IST-SOPRANO" (Service Oriented PRogrammable smArt enviroNments for Older Europeans) συμμετοχή σε θέματα που σχετίζονται με την ενσωμάτωση του συστήματος αναγνώρισης φωνής σε έξυπνα σπίτια.

- “ΓΓΕΤ-ΕΗΓ-Αυθόρμητο” (Αναγνώριση φωνής αυθόρμητου λόγου) συμμετοχή σε θέματα που σχετίζονται με τον εμπλουτισμό του γλωσσικού μοντέλου.
- “ΓΓΕΤ-ΕΗΓ-SUB4ALL”: (Τεχνολογίες Αυτόματου Υποτιτλισμού) συμμετοχή σε θέματα που σχετίζονται με την αποδελτίωση ραδιοφωνικών και τηλεοπτικών εκπομπών.

Διδακτική εμπειρία:

- Σεμινάρια στο πρόγραμμα ΕΥΤΕΧΝΟΣ με θέμα Υποστηρικτικές Τεχνολογίες Πληροφορικής και Τηλεπικοινωνιών για Άτομα με Αναπηρίες.
    - Εισαγωγή στην αναγνώριση φωνής συναισθηματικού λόγου.
- Με την ιδιότητα του υποψήφιου διδάκτορα, συμμετοχή στην διδασκαλία του μαθήματος «Σήματα και Συστήματα» του τμήματος ΗΜΜΥ του Ε.Μ.Π.

Ξένες Γλώσσες:

- Αγγλικά (Cambridge First Certificate in English)
- Γαλλικά (Certificat de Langue Francais, D.E.L.F A3,A4)