
**ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ**

**ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ**
**ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ**

**ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ**


**ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ**

# Μηχανική Μάθηση στην Επεξεργασία Φυσικής Γλώσσας


**Γεώργιος Πέτρου Πετάσης**



**ΑΘΗΝΑ**

**ΙΟΥΛΙΟΣ 2011**

**ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ**

Μηχανική Μάθηση στην Επεξεργασία Φυσικής Γλώσσας

**Γεώργιος Πέτρου Πετάσης**

**ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: Κωνσταντίνος Χαλάτσης,** Καθηγητής ΕΚΠΑ

**ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ ΠΑΡΑΚΟΛΟΥΘΗΣΗΣ:**
**Κωνσταντίνος Χαλάτσης,** Καθηγητής ΕΚΠΑ
**Παναγιώτης Σταματόπουλος,** Επίκουρος Καθηγητής ΕΚΠΑ
**Κωνσταντίνος Σπυρόπουλος,** Διευθυντής Έρευνας ΕΚΕΦΕ Δημόκριτος

**ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ**

(Υπογραφή)

**Κωνσταντίνος Χαλάτσης,**
**Καθηγητής ΕΚΠΑ**

(Υπογραφή)

**Κωνσταντίνος Σπυρόπουλος,**
**Διευθυντής Έρευνας ΕΚΕΦΕ Δημόκριτος**

(Υπογραφή)

**Παναγιώτης Σταματόπουλος,**
**Επίκουρος Καθηγητής ΕΚΠΑ**

(Υπογραφή)

**Μαρία Γρηγοριάδου,**
**Καθηγήτρια ΕΚΠΑ**

(Υπογραφή)

**Νικόλαος Φακωτάκης,**
**Καθηγητής Πανεπιστημίου Πατρών**

(Υπογραφή)

**Ευάγγελος Καρκαλέτσης,**
**Διευθυντής Έρευνας ΕΚΕΦΕ Δημόκριτος**

(Υπογραφή)

**Ίων Ανδρουτσόπουλος,**
**Επίκουρος Καθηγητής ΟΠΑ**

**Ημερομηνία εξέτασης 01/07/2011**

# ΠΕΡΙΛΗΨΗ

Η παρούσα διατριβή εξετάζει την χρήση τεχνικών μηχανικής μάθησης σε διάφορα στάδια της επεξεργασίας φυσικής γλώσσας, κυρίως για σκοπούς εξαγωγής πληροφορίας από κείμενα. Στόχος είναι τόσο η βελτίωση της προσαρμοστικότητας των συστημάτων εξαγωγής πληροφορίας σε νέες θεματικές περιοχές (ή ακόμα και γλώσσες), όσο και η επίτευξη καλύτερης απόδοσης χρησιμοποιώντας όσο το δυνατό λιγότερους πόρους (τόσο γλωσσικούς όσο και ανθρώπινους). Η διατριβή κινείται σε δύο κύριους άξονες: α) την έρευνα και αποτίμηση υπαρχόντων αλγορίθμων μηχανικής μάθησης κυρίως στα στάδια της προ-επεξεργασίας (όπως η αναγνώριση μερών του λόγου) και της αναγνώρισης ονομάτων οντοτήτων, και β) τη δημιουργία ενός νέου αλγορίθμου μηχανικής μάθησης και αποτίμησής του, τόσο σε συνθετικά δεδομένα, όσο και σε πραγματικά δεδομένα από το στάδιο της εξαγωγής σχέσεων μεταξύ ονομάτων οντοτήτων.

Η παρούσα διδακτορική διατριβή ερευνά τη δυνατότητα αξιοποίησης τεχνικών μηχανικής μάθησης στην περιοχή της επεξεργασίας φυσικής γλώσσας, με σκοπό την αντιμετώπιση του προβλήματος της αναβάθμισης καθώς και της προσαρμογής συστημάτων επεξεργασίας φυσικής γλώσσας σε νέες θεματικές περιοχές ή γλώσσες. Η έρευνα οριοθετείται σε τρεις σημαντικούς άξονες ενός συστήματος εξαγωγής πληροφορίας:

- Αναγνώριση μερών του λόγου για την Ελληνική γλώσσα.
- Αναγνώριση ονομάτων οντοτήτων.
- Αναγνώριση σχέσεων ανάμεσα σε αναγνωρισμένα ονόματα οντοτήτων.

Η διατριβή εξετάζει το πώς μπορούν να αξιοποιηθούν μέθοδοι και τεχνικές μηχανικής μάθησης για την κατασκευή συστημάτων που υποστηρίζουν τις εργασίες αυτές, τα οποία θα προσαρμόζονται ευκολότερα σε νέες θεματικές περιοχές και γλώσσες σε σχέση με τα συμβατικά συστήματα που βασίζονται σε κανόνες, κατασκευασμένους συχνά από ειδικούς. Ειδικότερα, η διατριβή ερευνά τεχνικές μηχανικής μάθησης προς δύο κύριους άξονες:

1. Την εφαρμογή υπαρχουσών τεχνικών (τόσο συμβολικών όσο και στατιστικών) σε επιλεγμένα στάδια της εξαγωγής πληροφορίας. Οι τεχνικές αυτές αποτιμούνται συγκριτικά μεταξύ τους σε κείμενα τόσο στην Ελληνική όσο και την Αγγλική γλώσσα. Όλοι οι υπάρχοντες αλγόριθμοι μηχανικής μάθησης που εξετάστηκαν χρειάζονται σαν είσοδο ένα διάνυσμα σταθερού μήκους. Ωστόσο η μετατροπή της φυσικής γλώσσας σε διάνυσμα σταθερού μήκους δεν είναι πάντα εύκολη χωρίς την χρήση αυθαιρέτων ορίων όσον αφορά τον μέγιστο αριθμό λέξεων. Η παρατήρηση αυτή αποτέλεσε και το βασικό κίνητρο για την δημιουργία ενός νέου αλγορίθμου μηχανικής μάθησης χωρίς τον περιορισμό των διανυσμάτων σταθερού μήκους σαν είσοδο.
2. Την ανάπτυξη ενός νέου αλγορίθμου μηχανικής μάθησης, χωρίς την απαίτηση για είσοδο διανυσμάτων σταθερού μήκους. Ο νέος αυτός αλγόριθμος μαθαίνει *γραμματικές ανεξάρτητες από τα συμφραζόμενα (context free grammars)* από θετικά παραδείγματα, με καθοδήγηση μέσω ευριστικών, όπως το *ελάχιστο μήκος περιγραφής (minimum description length)*.

Όσον αφορά τον πρώτο άξονα, αναπτύχθηκαν και αξιολογήθηκαν συστήματα αναγνώρισης ονομάτων οντοτήτων, βασισμένα σε υπάρχοντες αλγορίθμους μηχανικής μάθησης, όπως δέντρα αποφάσεων και νευρωνικά δίκτυα. Τα συστήματα που αναπτύχθηκαν αφορούν διάφορες θεματικές περιοχές (μετακινήσεις στελεχών επιχειρήσεων, χρηματο-οικονομικές ειδήσεις, δικαστικές αποφάσεις) τόσο στην

Ελληνική όσο και στην Αγγλική γλώσσα. Τα συστήματα αυτά αξιολογήθηκαν σε κείμενα της Ελληνικής γλώσσας, και οδήγησαν στον προσδιορισμό των μειονεκτημάτων και των περιορισμών που επιβάλλουν οι εξετασθέντες αλγόριθμοι όταν εφαρμόζονται σε δεδομένα φυσικής γλώσσας. Από την παραπάνω ανάλυση, προέκυψε ότι ένα από τα σημαντικότερα προβλήματα της εφαρμογής μηχανικής μάθησης είναι η δυσκολία διαχείρισης δεδομένων μεταβλητού μήκους, όπως π.χ. χαρακτηριστικά που αφορούν όλες τις λέξεις μιας πρότασης. Αντίθετα, ένας συντακτικός αναλυτής μπορεί να εξετάσει εύκολα αν μια πρόταση ή μέρος αυτής περιγράφεται από μια δεδομένη γραμματική. Ωστόσο, η χειροκίνητη ανάπτυξη γραμματικών κατάλληλων για μια εργασία είναι μια σύνθετη διαδικασία, ενώ τα αποτελέσματα συχνά εξαρτώνται από την θεματική περιοχή και σαφώς από την γλώσσα. Συνεπώς, αν μια τέτοια γραμματική είναι δυνατόν να αποκτηθεί αυτόματα με την χρήση μηχανικής μάθησης, τότε η προσαρμογή συστημάτων που χρησιμοποιούν τέτοιες γραμματικές σε νέες θεματικές περιοχές ή γλώσσες, είναι δυνατόν να απλοποιηθεί σημαντικά.

Η συμβολή των συστημάτων που αναπτύχθηκαν είναι σημαντική, τόσο σε Ελληνικό όσο και σε διεθνές επίπεδο. Τα συστήματα αναγνώρισης ονομάτων οντοτήτων που αναπτύχθηκαν για την Ελληνική γλώσσα είναι τα πρώτα συστήματα στο είδος τους που αναφέρονται στη βιβλιογραφία. Ταυτόχρονα, η απόδοση των υλοποιηθέντων συστημάτων κρίνεται ιδιαίτερα ικανοποιητική, αφού είναι συγκρίσιμα με τις αποδόσεις συστημάτων που παρουσιάζονται στην διεθνή βιβλιογραφία την αντίστοιχη χρονική περίοδο.

Όσον αφορά το δεύτερο άξονα, και έχοντας σαν στόχο την αντιμετώπιση των προβλημάτων που εμφανίζει η εφαρμογή υπαρχουσών τεχνικών, αναπτύχθηκε μία νέα τεχνική μηχανικής μάθησης. Η νέα τεχνική εντάσσεται στην κατηγορία *της επαγωγικής εξαγωγής γραμματικών (inductive grammar learning)*. Τα κύρια πλεονεκτήματα της μεθόδου αυτής σε σχέση με άλλες μεθόδους μηχανικής μάθησης, είναι η δυνατότητα χειρισμού δεδομένων σε μορφή κειμένου, καθώς και η δυνατότητα ενσωμάτωσής της σε υπάρχοντα συστήματα αντικαθιστώντας χειρωνακτικά κατασκευασμένες γραμματικές. Κύριος στόχος της νέας αυτής τεχνικής είναι η αυτοματοποίηση της διαδικασίας δημιουργίας γραμματικών, οι οποίες να μπορούν να συνεργαστούν με την πληθώρα των συντακτικών αναλυτών που εμφανίζονται στην διεθνή βιβλιογραφία, αντικαθιστώντας υπάρχουσες (και πιθανώς χειρωνακτικά κατασκευασμένες) γραμματικές για διάφορες υπο-εργασίες συστημάτων εξαγωγής πληροφορίας.

Για την εφαρμογή της επαγωγικής εξαγωγής γραμματικών *αναπτύχθηκε ένας νέος αλγόριθμος επαγωγικής εξαγωγής γραμματικών* που λειτουργεί μόνο με θετικά παραδείγματα. Ο νέος αυτός αλγόριθμος μπορεί να επάγει *γραμματικές ανεξάρτητες από τα συμφραζόμενα (context free grammars)*, και βασίστηκε στον υπάρχοντα αλγόριθμο GRIDS, βελτιώνοντας τόσο το χρησιμοποιούμενο ευριστικό, όσο και την διαδικασία αναζήτησης στον χώρο των πιθανών γραμματικών, αυξάνοντας ταυτόχρονα την εφαρμοσιμότητα του νέου αλγορίθμου σε μεγαλύτερα σύνολα δεδομένων. Η απαίτηση ο αλγόριθμος να λειτουργεί μόνο με θετικά παραδείγματα προέρχεται από την συχνή ανυπαρξία αρνητικών παραδειγμάτων στην περιοχή της επεξεργασίας φυσικής γλώσσας. Σημειώνεται ότι η παρουσία αρνητικών παραδειγμάτων αποτελεί προϋπόθεση για την λειτουργία της μεγαλύτερης πλειοψηφίας των υπαρχόντων αλγορίθμων εξαγωγής γραμματικών. Η σχεδίαση του νέου αλγορίθμου έγινε με τέτοιο τρόπο ώστε να μπορεί να χρησιμοποιηθεί σε εργασίες κατηγοριοποίησης, όπως π.χ. στην αναγνώριση ονομάτων οντοτήτων. Η λειτουργία αυτή είναι διαφορετική από την συνήθη εφαρμογή αλγορίθμων εξαγωγής γραμματικών, καθώς δεν απαιτείται ούτε ο χαρακτηρισμός προτάσεων ως γραμματικά ορθές ή μη, ούτε η συντακτική ανάλυση προτάσεων, αλλά μόνο ο εντοπισμός *τμημάτων των προτάσεων* και η κατηγοριοποίησή τους σε κατάλληλες κατηγορίες. Η αποτίμηση του αλγορίθμου αυτού έγινε τόσο σε

συνθετικές γλώσσες, όσο και στην υπο-εργασία της εξαγωγής σχέσεων μεταξύ ονομάτων οντοτήτων.

## ΣΥΝΕΙΣΦΟΡΑ ΤΗΣ ΔΙΑΤΡΙΒΗΣ

Η παρούσα διατριβή προτείνει την αξιοποίηση της μηχανικής μάθησης σε κομβικά σημεία ενός τυπικού συστήματος εξαγωγής πληροφορίας, έχοντας ως σκοπό την υποβοήθηση της προσαρμογής του συστήματος σε νέες περιοχές και ίσως και σε γλώσσες. Το πρώτο πεδίο έρευνας αυτής της διατριβής αποτελεί η αναγνώριση μερών του λόγου για την Ελληνική γλώσσα. Η *μάθηση στηριζόμενη σε κανόνες μετασχηματισμού καθοδηγούμενη από σφάλματα (transformation-based error-driven learning – TBED)* εφαρμόστηκε για πρώτη φορά στην Ελληνική γλώσσα, πετυχαίνοντας υψηλές αποδόσεις, άμεσα συγκρίσιμες με αντίστοιχα συστήματα για την Ελληνική γλώσσα, απαιτώντας ταυτόχρονα σημαντικά μειωμένα δεδομένα εκπαίδευσης. Ταυτόχρονα, η προσέγγιση που περιγράφεται σε αυτή την διατριβή αποτέλεσε τον πρώτο αναγνωριστή μερών του λόγου που διατέθηκε ελεύθερα σαν εφαρμογή ανοικτού λογισμικού, με σημαντική αποδοχή από την επιστημονική κοινότητα, όπως καταδεικνύει ο αριθμός των ετεροαναφορών στις σχετικές δημοσιεύσεις.

Το δεύτερο πεδίο έρευνας αφορά την περιοχή της αναγνώρισης ονομάτων οντοτήτων. Τρεις αλγόριθμοι μηχανικής μάθησης δοκιμάστηκαν σε αυτή την εργασία, τόσο συμβολικοί όσο και στοχαστικοί, πετυχαίνοντας ικανοποιητικά αποτελέσματα. Οι αλγόριθμοι δοκιμάστηκαν σε διάφορες θεματικές περιοχές, τόσο σε Αγγλικά, όσο και Ελληνικά, επιβεβαιώνοντας όχι μόνο την ικανότητα της μηχανικής μάθησης να υποστηρίξει την εργασία της αναγνώρισης των ονομάτων οντοτήτων, αλλά και την προσαρμοστικότητα των εν λόγω συστημάτων όχι μόνο σε νέες θεματικές περιοχές, αλλά και σε γλώσσες. Η εργασία που υλοποιήθηκε στα πλαίσια αυτής της διατριβής συγκαταλέγεται ανάμεσα στα πρώτα συστήματα εξαγωγής πληροφορίας για την Ελληνική γλώσσα που αναφέρονται στην διεθνή βιβλιογραφία. Επιπρόσθετα, στο πλαίσιο της πειραματικής αποτίμησης των αλγορίθμων μηχανικής μάθησης είχε παρατηρηθεί ότι, τουλάχιστον για την εργασία της αναγνώρισης ονομάτων οντοτήτων, η σειρά των λέξεων σε μια πρόταση δεν διαδραματίζει ιδιαίτερο λόγο. Αν και αρχικά η συγκεκριμένη διαπίστωση προκάλεσε έκπληξη, εντούτοις η διαδεδομένη χρήση της αναπαράστασης *«συνόλου λέξεων» (bag-of-words representation)* – η οποία αγνοεί την σειρά των λέξεων – όχι μόνο για την αναγνώριση ονομάτων οντοτήτων, αλλά και για αρκετά ακόμα προβλήματα επεξεργασίας φυσικής γλώσσας, καταδεικνύει την ορθότητα της αρχικής εκείνης παρατήρησης.

Το τρίτο πεδίο έρευνας αφορά την ανάπτυξη ενός νέου αλγόριθμου μηχανικής μάθησης, και συγκεκριμένα ενός αλγορίθμου επαγωγικής εξαγωγής γραμματικών, ικανού να εξάγει γραμματικές ανεξάρτητες από συμφραζόμενα μόνο από θετικά παραδείγματα. Σημαντική ιδιότητα του νέου αυτού αλγορίθμου είναι η ικανότητα να επεξεργαστεί μεγάλους όγκους δεδομένων, απόρροια της παρατήρησης ότι είναι υπολογιστικά φθηνότερο να προβλέψεις το αποτέλεσμα της εφαρμογής των τελεστών εκμάθησης, παρά να τους εφαρμόσεις και να αποτιμήσεις το αποτέλεσμα.

**ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ**: Επεξεργασία Φυσικής Γλώσσας

**ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ**: εξαγωγή πληροφορίας, μηχανική μάθηση, επαγωγική εξαγωγή γραμματικών, εξαγωγή, μάθηση, γραμματική

# ABSTRACT

This thesis examines the use of machine learning techniques in various tasks of natural language processing, mainly for the task of information extraction from texts. The objectives are the improvement of adaptability of information extraction systems to new thematic domains (or even languages), and the improvement of their performance using as fewer resources (either linguistic or human) as possible. This thesis has examined two main axes: a) the research and assessment of existing algorithms of machine learning mainly in the stages of linguistic pre-processing (such as part of speech tagging) and named-entity recognition, and b) the creation of a new machine learning algorithm and its assessment on synthetic data, as well as in real world data from the task of relation extraction between named entities.

This doctoral thesis researches the possibility of exploiting machine learning techniques in the research area of natural language processing, aiming at the confrontation of the problems of upgrade as well as adaptation of natural language processing systems in new thematic domains or languages. The research is delimited in three important axes of information extraction systems:

- Part of speech recognition for the Greek language.
- Named entity recognition.
- Relation extraction between recognised named entities.

This thesis examines how machine learning methods and techniques can be exploited for the development of systems that support these tasks, which can be adapted more easily in new thematic domains and languages in contrast to the conventional systems that are rule based, manufactured often by experts. More specifically, this thesis researches techniques of machine learning along two main axes:

1. The application of existing techniques (both symbolic and statistical) in selected tasks of information extraction. These techniques are evaluated comparatively to each other in both the Greek and English languages. All existing machine learning algorithms that were examined require a vector of constant length as input. However the transformation of natural language into vectors of constant length is not always easy, without the use of arbitrary limits regarding the maximum number of words. This observation constituted the motivation for the creation of a new machine learning algorithm, which does not require vectors of constant length as input.
2. The development of a new machine learning algorithm, without the requirement for vectors of constant length as input. This new algorithm learns context free grammars from positive examples, with guidance via heuristics, such as minimum description length.

Regarding the first axis, named entity recognition systems were developed and evaluated, based on existing machine learning algorithms, such as decision trees and neural networks. The systems that were developed concern various thematic domains (management succession events, financial news, and juridical decisions) both in the Greek and English languages. These systems were evaluated in Greek texts, and they led to the recognition of the disadvantages and restrictions imposed by the examined algorithms, when applied on natural language data. From this analysis we concluded that one of the main problems when applying machine learning is the difficulty in managing data of variable length, as for example the information concerning all words of a sentence. On the contrary, a syntactic analyser can easily decide if a sentence (or part of a sentence) is described by a provided grammar. However, the manual

development of grammars suitable for a specific task is a complex process, while the results frequently depend on the thematic domain and of course from the language. Consequently, if such a grammar can be automatically acquired with the use of machine learning, then the adaptation of systems that use such grammars to new thematic domains or languages can be considerably simplified.

The contribution of the developed systems is significant. The named entity recognition systems that were developed for the Greek language are among the first systems of their kind that have been reported in the bibliography. Simultaneously, the performance of the developed systems is satisfactory, and directly comparable to the performance of similar systems reported in the bibliography for the corresponding time period.

Regarding the second axis, and aiming at the confrontation of problems associated with the application of existing techniques, a new technique of machine learning has been developed. This new technique belongs to the category of inductive grammar learning. The main advantages of this method with respect to other machine learning methods are the ability to handle textual data, as well as the possibility of using learned grammars in existing systems, replacing manually developed grammars. The main objective of this new technique is the automatic grammar creation, which can be used with the plethora of available syntactic parsers that have been presented in the bibliography, replacing existing (and probably manually constructed) grammars for various tasks in information extraction systems.

For applying inductive grammar learning, a new algorithm has been developed that learns grammars from positive examples only. This new algorithm can infer context free grammars, and it has been based on the existing algorithm GRIDS, improving both the used heuristic, as well as the search process in the space of possible grammars, increasing simultaneously the applicability of the new algorithm in bigger collections of data. The requirement for the algorithm to function only with positive examples emanates from the frequent absence of negative examples in the area of natural language processing. It should be noted that the presence of negative examples constitutes a necessary condition for the operation of most existing grammatical inference algorithms. The design of this new algorithm has been done in such a way that it can be used in classification tasks, such as named entity recognition. This kind of usage differs from the usual application of grammatical inference algorithms, as the verification or the syntactic analysis of sentences according to a grammar is not required. Instead, we are interested mainly in recognising sentence parts (phrases) and their classification in predefined semantic categories. The evaluation of this new algorithm has been performed on both synthetic languages, as well as on real world data for the task of relation extraction between named entities.

## CONTRIBUTION

This thesis proposes the exploitation of machine learning in nodal points of a typical information extraction system, having as aim the assistance adapting the system into new thematic domains and perhaps languages. This first research topic of this thesis involves part of speech tagging for the Greek language. The *transformation-based error-driven learning (TBED)* has been applied for the first time in the Greek language, achieving high performance, directly comparable with corresponding systems for the Greek language, requiring at the same time considerably less training data. Simultaneously, the approach that is described in this thesis constituted the first Greek part-of-speech tagger that has been distributed freely, as an open source application, with important acceptance from the scientific community, as denoted by the number of citations in the relative publications.

The second research topic concerns the area of named entity recognition. Three machine learning algorithms were examined for this task, both symbolic and stochastic ones, achieving satisfactory results. The algorithms were examined in various thematic domains, in both the English and Greek languages, confirming not only the ability of machine learning to support the task of named entity recognition, but also the adaptability of the machine learning based approaches not only in new thematic domains, but also in languages. The research that has been contacted in the context of this thesis is included among the first information extraction systems for the Greek language that have been reported in the bibliography. In addition, it has been observed that, at least for the task of named entity recognition, the order of words in a sentence is not important. Despite the fact that initially this observation seemed surprising, the widespread use of the “bag-of-words” representation - which also ignores the word order - not only for named entity recognition, but also for other natural language processing tasks, it shows the correctness of this initial observation.

The third research topic concerns the development of new machine learning algorithm, able to infer context free grammars from positive only examples. An important characteristic of this new algorithm is its ability to processes large volumes of data, a consequence of the observation that it is computationally cheaper to predict the result of applying a learning operator, than to apply the operator and evaluate the produced grammar. The results achieved by the new algorithm on the “Omphalos” competition were also significant, where it solved the first problem without human intervention within the competition time period, while it has been successfully combined with the winning algorithm, removing the need of the winning algorithm for human intervention, in cases where its heuristic could not drive further the learning process.

**SUBJECT AREA**: Natural Language Processing

**KEYWORDS**: information extraction, machine learning, grammatical inference, extraction, learning, grammar

*Στους γονείς μου*

# ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα διατριβή δεν θα είχε πραγματοποιηθεί ποτέ αν δεν υπήρχε το Ινστιτούτο Πληροφορικής και Τηλεπικοινωνιών του Εθνικού Κέντρου Έρευνας Φυσικών Επιστημών (Ε.Κ.Ε.Φ.Ε.) “Δημόκριτος”, το οποίο μέσω εξετάσεων εμπιστεύτηκε έναν Φυσικό για να εκπονήσει διδακτορική διατριβή σε ένα θέμα που άπτεται της Επιστήμης Υπολογιστών. Η παρούσα διατριβή πραγματοποιήθηκε με τη βοήθεια υποτροφίας που χορηγήθηκε από το Ινστιτούτο Πληροφορικής και Τηλεπικοινωνιών του Εθνικού Κέντρου Έρευνας Φυσικών Επιστημών (Ε.Κ.Ε.Φ.Ε.) “Δημόκριτος”. Ευχαριστώ όλους όσοι συντέλεσαν στην επιτυχή ολοκλήρωση της διατριβής αυτής.

Ευχαριστώ ιδιαίτερα τον κ. Κωνσταντίνο Σπυρόπουλο, Διευθυντή Έρευνας του Ε.Κ.Ε.Φ.Ε. “Δημόκριτος” για τα χρήσιμα σχόλιά του καθ’ όλη τη διάρκεια του διδακτορικού, στο τελικό κείμενο της διατριβής, καθώς και για το γεγονός ότι εξασφάλισε χρηματοδότηση για μια σειρά από θερινά σχολεία και συνέδρια τα οποία με βοήθησαν να προσεγγίσω περιοχές στις οποίες δεν είχα καμία εμπειρία (όπως αυτή της επαγωγικής εξαγωγής γραμματικών) και ανέδειξαν την ερευνητική μου εργασία σε διεθνή επίπεδο. Ευχαριστώ ιδιαίτερα τον κ. Ευάγγελο Καρκαλέτση, Διευθυντή Έρευνας του Ε.Κ.Ε.Φ.Ε. “Δημόκριτος”, ο οποίος με εισήγαγε στον χώρο της επεξεργασίας φυσικής γλώσσας και συνεισέφερε ουσιαστικά στην βελτίωση του τρόπου συγγραφής των επιστημονικών μου άρθρων. Επίσης, ευχαριστώ ιδιαίτερα των Γεώργιο Παλιούρα, Ερευνητή Β του Ε.Κ.Ε.Φ.Ε. “Δημόκριτος”, ο οποίος με εισήγαγε στην περιοχή της μηχανικής μάθησης. Τον ευχαριστώ για τις δημιουργικές συζητήσεις που είχαμε σχετικά με την ανάπτυξη του αλγορίθμου egGRIDS+, και για την συνεισφορά του στην βελτίωση του τρόπου συγγραφής των επιστημονικών μου άρθρων. Ευχαριστώ επίσης τα υπόλοιπα δύο μέλη της τριμελούς μου επιτροπής από το Πανεπιστήμιο Αθηνών, τον καθηγητή κ. Κωνσταντίνο Χαλάτση και τον επίκουρο καθηγητή κ. Παναγιώτη Σταματόπουλο για τη δημιουργική συνεργασία που είχαμε, αλλά και την υπομονή που επέδειξαν.

Όσον αφορά την εκπόνηση της παρούσας διατριβής, θα ήθελα να ευχαριστήσω τους Pat Langley, διευθυντή του Center for the Study of Language and Information, του Stanford University, αλλά και του Sean Stromsten, Lead Research Engineer at BAE Systems Advanced Information Technologies, για τον κώδικα που μου έστειλαν. Θα ήθελα επίσης να ευχαριστήσω την Claire Grover, Senior Research Fellow, School of Informatics, University of Edinburgh, για τα σώματα κειμένων και τα δεδομένα αξιολόγησης συστημάτων του Πανεπιστημίου του Εδιμβούργου.

Ευχαριστώ τον Δρ. Γεώργιο Συγλέτο, για την βοήθεια που μου έδωσε όταν την χρειάστηκα. Ευχαριστώ τον Δρ. Δημήτρη Σπηλιωτόπουλο για την φιλία και υποστήριξή του.

Τέλος, ένα πολύ μεγάλο ευχαριστώ στους γονείς μου και τα αδέρφια μου, που με στήριξαν οικονομικά και ψυχολογικά από το πρώτο έτος των βασικών σπουδών μου έως και σήμερα, συνεισφέροντας σημαντικά στην επιτυχή ολοκλήρωση της διατριβής αυτής.

# ΠΕΡΙΕΧΟΜΕΝΑ


**ΠΡΟΛΟΓΟΣ ..... 29**

**1. ΕΙΣΑΓΩΓΗ ..... 31**

**1.1 Σύντομη επισκόπηση του αντικειμένου της διατριβής ..... 31**

**1.2 Ορισμός προβλήματος ..... 33**

**1.3 Η προσέγγιση της διατριβής ..... 33**

**1.4 Συνεισφορά ..... 35**

**1.5 Δημοσιεύσεις ..... 36**

**1.6 Αξιοποίηση των αποτελεσμάτων της διατριβής ..... 41**

**1.7 Άλλες δημοσιεύσεις ..... 42**

**1.8 Διάρθρωση της διατριβής ..... 44**

**2. ΒΑΣΙΚΕΣ ΈΝΝΟΙΕΣ ΚΑΙ ΓΕΝΙΚΗ ΕΠΙΣΚΟΠΗΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑΣ ..... 47**

**2.1 Εξαγωγή πληροφορίας ..... 47**
2.1.1 Ορισμός και υπο-προβλήματα ..... 48

**2.2 Μηχανική μάθηση ..... 49**
2.2.1 Ορισμός ..... 49
2.2.2 Κατηγορίες μηχανικής μάθησης ..... 50

**2.3 Εκτίμηση επίδοσης αλγορίθμων μηχανικής μάθησης ..... 50**
2.3.1 Ακρίβεια (precision) ..... 50
2.3.2 Ανάκληση (recall) ..... 51
2.3.3 F-Measure ..... 51
2.3.4 Διασταυρωμένη επικύρωση (cross validation) ..... 51

**3. ΕΞΑΓΩΓΗ ΠΛΗΡΟΦΟΡΙΑΣ: ΑΝΑΓΝΩΡΙΣΗ ΜΕΡΩΝ ΤΟΥ ΛΟΓΟΥ ..... 53**

**3.1 Ορισμός προβλήματος ..... 53**

**3.2 Βιβλιογραφική επισκόπηση ..... 54**

**3.3 Μάθηση στηριζόμενη σε κανόνες μετασχηματισμού καθοδηγούμενη από σφάλματα ..... 56**

**3.4 Η προτεινόμενη προσέγγιση ..... 57**

**3.5 Το σύνολο των ετικετών ..... 57**

**3.6 Ο κανόνας αρχικοποίησης ..... 58**

**3.7 Τροποποίηση των πρωτοτύπων κανόνων ..... 59**

**3.8 Τα σώματα κειμένων (δεδομένα εκπαίδευσης) ..... 59**

**3.9 Πειραματική αξιολόγηση και αποτελέσματα ..... 60**
3.9.1 Αναγνώριση μερών του λόγου (γενική θεματική περιοχή) ..... 60
3.9.2 Αναγνώριση μερών του λόγου (συγκεκριμένη θεματική περιοχή) ..... 62

**3.10 Συνδυασμός μηχανικής μάθησης και μορφολογικού λεξικού ..... 64**
3.10.1 Πειραματική αξιολόγηση και αποτελέσματα ..... 65

**3.11 Συνεισφορά ..... 65**

**4. ΕΞΑΓΩΓΗ ΠΛΗΡΟΦΟΡΙΑΣ: ΑΝΑΓΝΩΡΙΣΗ ΟΝΟΜΑΤΩΝ ΟΝΤΟΤΗΤΩΝ ..... 67**

**4.1 Ορισμός προβλήματος ..... 67**

**4.2 Βιβλιογραφική επισκόπηση ..... 68**
4.2.1 Επιδόσεις συστημάτων στα συνέδρια MUC ..... 71
4.2.2 Προσεγγίσεις των συστημάτων του MUC-7 ..... 72

**4.3 Τυπική Αρχιτεκτονική ενός συστήματος αναγνώρισης ονομάτων οντοτήτων ..... 76**

**4.4 Η προτεινόμενη προσέγγιση ..... 77**

**4.5 Αντικατάσταση του υποσυστήματος της γραμματικής ..... 77**

**4.6 Προσέγγιση Α: μηχανική μάθηση στην αναγνώριση ονομάτων οντοτήτων ..... 78**
4.6.1 Μηχανική μάθηση και αναπαράσταση γνώσης ..... 79
4.6.2 Σύστημα NERC βασισμένο σε δέντρα αποφάσεων ..... 80
4.6.3 Σύστημα NERC βασισμένο σε νευρωνικά δίκτυα ..... 81
4.6.4 Συμβολικό διάνυσμα χαρακτηριστικών ..... 82
4.6.5 Αριθμητικό διάνυσμα χαρακτηριστικών ..... 82
4.6.6 Πειραματική αξιολόγηση και αποτελέσματα ..... 85

**4.7 Προσέγγιση Β: συνδυασμός συστημάτων κατηγοριοποίησης σε επίπεδο λέξεων και φράσεων ..... 88**
4.7.1 Γλωσσική προ-επεξεργασία ..... 89
4.7.2 Αναγνώριση ονομάτων οντοτήτων σε επίπεδο λέξης ..... 90
4.7.3 Αναγνώριση ονομάτων οντοτήτων σε επίπεδο φράσης ..... 93
4.7.4 Φιλτράρισμα (filtering) ..... 95
4.7.5 Πειραματική αξιολόγηση και αποτελέσματα ..... 96

**4.8 Προσαρμογή/εμπλουτισμός του υποσυστήματος του λεξικού ..... 105**
4.8.1 Η προσέγγιση ..... 106
4.8.2 Πειραματική αξιολόγηση και αποτελέσματα ..... 107

**4.9 Ενημέρωση συστήματος αναγνώρισης ονομάτων οντοτήτων ..... 111**
4.9.1 Η προσέγγιση ..... 111
4.9.2 Πειραματική αξιολόγηση και αποτελέσματα ..... 113

**4.10 Συνεισφορά ..... 115**

**5. ΕΠΑΓΩΓΙΚΗ ΕΞΑΓΩΓΗ ΓΡΑΜΜΑΤΙΚΩΝ: Ο ΑΛΓΟΡΙΘΜΟΣ EGGRIDS+ ..... 119**

**5.1 Ορισμός προβλήματος ..... 119**

**5.2 Βιβλιογραφική επισκόπηση ..... 119**

**5.3 Ο αλγόριθμος egGRIDS+: επαγωγική εξαγωγή γραμματικών από θετικά παραδείγματα ..... 123**
5.3.1 Αναπαράσταση γνώσης στον egGRIDS+ ..... 124
5.3.2 Αναζήτηση στον egGRIDS+: προτίμηση προς «απλές» γραμματικές ..... 125
5.3.3 Υπολογιστική πολυπλοκότητα της μέτρησης του μήκους του μοντέλου ..... 133

5.3.4 Η Αρχιτεκτονική του egGRIDS+ και οι τελεστές εκμάθησης ..... 134

**5.4 Οι τελεστές αναζήτησης του egGRIDS+ ..... 136**
5.4.1 Ο τελεστής “Create NT” ..... 136
5.4.2 Η πολυπλοκότητα της κατάστασης “Create NT” ..... 137
5.4.3 Η επίδραση του τελεστή “Create NT” στο μήκος περιγραφής γραμματικής ..... 138
5.4.4 Επιταχύνοντας τον τελεστή CreateNT ..... 139
5.4.5 Ο τελεστής “Merge NT” ..... 141
5.4.6 Η πολυπλοκότητα της κατάστασης “ Merge NT” ..... 142
5.4.7 Η επίδραση του τελεστή “Merge NT” στο μήκος περιγραφής γραμματικής ..... 144
5.4.8 Η επίδραση του τελεστή “Merge NT” στο μήκος περιγραφής παραγωγών ..... 145
5.4.9 Συνολική επίδραση του τελεστή “Merge NT” στο μήκος περιγραφής μοντέλου γραμματικής ..... 149
5.4.10 Επιταχύνοντας τον τελεστή MergeNT ..... 149
*5.4.11* Ο τελεστής “Create Optional NT” ..... 150
5.4.12 Η πολυπλοκότητα της κατάστασης “Create Optional NT” ..... 152
5.4.13 Η επίδραση του τελεστή “Create Optional NT” στο μήκος περιγραφής γραμματικής ..... 154
5.4.14 Η επίδραση του τελεστή “Create Optional NT” στο μήκος περιγραφής παραγωγών ..... 155
5.4.15 Συνολική επίδραση του “Create Optional NT” στο μήκος περιγραφής μοντέλου γραμματικής ..... 156

**5.5 Πειραματική αξιολόγηση και αποτελέσματα ..... 156**
5.5.1 Πειράματα σε Τεχνητές Γραμματικές ..... 156
5.5.2 Πείραμα 1: Αξιολόγηση σε μικρές γραμματικές ..... 159
5.5.3 Πείραμα 2: Διαφοροποιώντας το μέγεθος της δέσμης ..... 161
5.5.4 Πείραμα 3: Η γλώσσα ισορροπημένων παρενθέσεων ..... 162
5.5.5 Πειράματα με σώματα κειμένων μεγάλου μεγέθους ..... 163
5.5.6 Ο διεθνής διαγωνισμός “Omphalos” ..... 166

**5.6 Συνεισφορά ..... 167**

## 6. ΕΞΑΓΩΓΗ ΠΛΗΡΟΦΟΡΙΑΣ: ΕΞΑΓΩΓΗ ΣΧΕΣΕΩΝ ΜΕΤΑΞΥ ΟΝΟΜΑΤΩΝ ΟΝΤΟΤΗΤΩΝ ..... 169

**6.1 Ορισμός προβλήματος ..... 169**

**6.2 Βιβλιογραφική επισκόπηση ..... 169**

**6.3 Η προσέγγιση ..... 170**
6.3.1 Εξάγοντας συσχετίσεις ..... 172
6.3.2 Ο αλγόριθμος egGRIDS+: η τάση προς απλές γραμματικές ..... 172
6.3.3 Αρχιτεκτονική και τελεστές μάθησης ..... 173

**6.4 Τα σώματα κειμένων (δεδομένα εκπαίδευσης) ..... 175**

**6.5 Πειραματική αξιολόγηση και αποτελέσματα ..... 175**
6.5.1 Δημιουργία παραδειγμάτων εκπαίδευσης/αποτίμησης ..... 175
6.5.2 Συγκριτική αξιολόγηση με έτερο αλγόριθμο μηχανικής μάθησης ..... 179

**6.6 Συνεισφορά ..... 182**

## 7. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ ..... 185

**7.1 Συμπεράσματα ..... 185**
7.1.1 Αξιοποίηση και αποτίμηση υπαρχόντων αλγορίθμων μηχανικής μάθησης ..... 185
7.1.2 egGRIDS+: ένας νέος αλγόριθμος επαγωγικής εξαγωγής γραμματικών ..... 189

**7.2 Προοπτικές μελλοντικές έρευνας ..... 190**

**ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ .....193**

**ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ.....197**

**ΠΑΡΑΡΤΗΜΑ Ι.....199**

**ΠΑΡΑΡΤΗΜΑ ΙΙ.....203**

**ΑΝΑΦΟΡΕΣ.....207**

# ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

Εικόνα 1: Η ορθότητα (accuracy) του αναγνωριστή μερών του λόγου TBED, σε συνάρτηση με το μέγεθος των δεδομένων εκπαίδευσης. .....61

Εικόνα 2: Ο αριθμός των λεκτικών και συμφραζόμενων κανόνων, σε συνάρτηση με το μέγεθος των δεδομένων εκπαίδευσης. .....62

Εικόνα 3: Η ορθότητα (accuracy) του αναγνωριστή μερών του λόγου TBED, σε συνάρτηση με το μέγεθος των δεδομένων εκπαίδευσης. (Θεματική περιοχή «γεγονότων διαδοχής διαχείρισης») .....63

Εικόνα 4: Ο αριθμός των λεκτικών και συμφραζόμενων κανόνων, σε συνάρτηση με το μέγεθος των δεδομένων εκπαίδευσης. (Θεματική περιοχή «γεγονότων διαδοχής διαχείρισης») .....64

Εικόνα 5: Υποσυστήματα ενός τυπικού συστήματος αναγνώρισης ονομάτων οντοτήτων. .....76

Εικόνα 6: Κωδικοποίηση μιας λέξης σε αριθμητικό διάνυσμα. .....83

Εικόνα 7: Κωδικοποίηση μιας φράσης σε αριθμητικό διάνυσμα. .....84

Εικόνα 8: Η αρχιτεκτονική του συστήματος ML-HNERC (δεύτερη προτεινόμενη προσέγγιση). .....89

Εικόνα 9: Η αρχιτεκτονική του αναγνωριστή ονομάτων οντοτήτων σε επίπεδο λέξης *(token-based NERC)*. .....91

Εικόνα 10: Η αναπαράσταση του διανύσματος εισόδου του τέταρτου ταξινομητή (Decision Tree Classifier 1). .....93

Εικόνα 11: Η αναπαράσταση του διανύσματος εισόδου του πέμπτου ταξινομητή (Decision Tree Classifier 2). .....93

Εικόνα 12: Η αρχιτεκτονική του αναγνωριστή ονομάτων οντοτήτων σε επίπεδο φράσης *(phrase-based NERC)* .....94

Εικόνα 13: Η αναπαράσταση ενός διανύσματος του υποσυστήματος κατηγοριοποίησης φράσεων (Decision Tree Phrase Classifier). .....94

Εικόνα 14: Παράδειγμα της εξόδου του συστήματος αναγνώρισης ονομάτων οντοτήτων σε επίπεδο λέξης. .....96

Εικόνα 15: Αποτελέσματα του συστήματος ML-HNERC για την Ελληνική γλώσσα.....97

Εικόνα 16: Αποτελέσματα του συστήματος ML-HNERC για την Αγγλική γλώσσα.....99

Εικόνα 17: Αποτελέσματα του συστήματος ML-HNERC για την Ελληνική γλώσσα (θεματική περιοχή: περιγραφές φορητών υπολογιστών από ιστοσελίδες ηλεκτρονικών καταστημάτων). ..... 102

Εικόνα 18: Αποτελέσματα του συστήματος ML-HNERC για την Ελληνική γλώσσα, χρησιμοποιώντας εμπλουτισμένες λίστες γνωστών ονομάτων οντοτήτων. ..... 104

Εικόνα 19: Η αρχιτεκτονική του συστήματος προσαρμογής/εμπλουτισμού του υποσυστήματος του λεξικού. ..... 106

Εικόνα 20: Η αναπαράσταση ενός διανύσματος του συστήματος προσαρμογής/εμπλουτισμού του υποσυστήματος του λεξικού..... 107

Εικόνα 21: Η αρχιτεκτονική του συστήματος σηματοδότησης ανάγκης ενημέρωσης, για το στάδιο της εκπαίδευσης..... 112

Εικόνα 22: Η αρχιτεκτονική του συστήματος σηματοδότησης ανάγκης ενημέρωσης, για το στάδιο του ελέγχου. ..... 112

Εικόνα 23: Η Αρχιτεκτονική ενός τυπικού αλγορίθμου επαγωγικής εξαγωγής γραμματικών. ..... 121

Εικόνα 24: Ψευδοκώδικας για τον υπολογισμό του μήκους μοντέλου $ML$ μιας γραμματικής. ..... 133

Εικόνα 25: Η αρχιτεκτονική του egGRIDS+. ..... 135

Εικόνα 26: Ψευδοκώδικας ενός βήματος (κατάστασης λειτουργίας του egGRIDS+) του τελεστή CreateNT. ..... 138

Εικόνα 27: : Ψευδοκώδικας ενός βήματος (κατάστασης λειτουργίας του egGRIDS+) του τελεστή MergeNT. ..... 143

Εικόνα 28: Ψευδοκώδικας ενός βήματος (κατάστασης λειτουργίας του egGRIDS+) του τελεστή CreateOptionalNT. ..... 153

Εικόνα 29: Πιθανότητα ανάλυσης μιας έγκυρης πρότασης μήκους μέχρι 15 λέξεις. (1-errors of omission) ..... 160

Εικόνα 30: Πιθανότητα ανάλυσης μιας έγκυρης πρότασης μήκους μεταξύ 16 και 20 λέξεων. (1-errors of omission)..... 160

Εικόνα 31: Πιθανότητα παραγωγής μιας έγκαιρης πρότασης. (1-errors of commission) ..... 160

Εικόνα 32: Πιθανότητα ανάλυσης μιας έγκυρης πρότασης μεγέθους μέχρι 15 λέξεις. (B = 10) ..... 162

Εικόνα 33: Πιθανότητα ανάλυσης μιας έγκυρης πρότασης μεγέθους μεταξύ 16 και 20 λέξεων. (B = 10) ..... 162

Εικόνα 34: Πιθανότητα παραγωγής μιας έγκυρης πρότασης. (B = 10) ..... 162

Εικόνα 35: Πιθανότητα ανάλυσης μιας έγκυρης πρότασης μεγέθους μέχρι 20 λέξεις. (1-errors of omission) ..... 163

Εικόνα 36: Πιθανότητα ανάλυσης μιας έγκυρης πρότασης μεγέθους από 21 μέχρι 25 λέξεων. (1-errors of omission) ..... 163

Εικόνα 37: Πιθανότητα παραγωγής μιας έγκυρης πρότασης. (1-errors of commission) ..... 163

Εικόνα 38: Ο χρόνος που απαιτείται από τον egGRIDS+, με και χωρίς τη χρήση αποτελεσμάτων θεωρητικής ανάλυσης. ..... 166

Εικόνα 39: Η αρχιτεκτονική του egGRIDS+. ..... 174

Εικόνα 40: Παράδειγμα πρότασης επισημειωμένης με ονόματα οντοτήτων. ..... 176

Εικόνα 41: Παραδείγματα εκπαίδευσης που εξήχθησαν από την πρόταση του παραδείγματος (Εικόνα 40). ..... 177

Εικόνα 42: Παράδειγμα πρότασης επισημειωμένης με ονόματα οντοτήτων. ..... 179

Εικόνα 43: Παράδειγμα διανύσματος εκπαίδευσης. ..... 180

Εικόνα 44: Ο αναγνωριστής ονομάτων οντοτήτων για την Ελληνική γλώσσα, όπως εμφανίζεται στην πλατφόρμα επεξεργασίας φυσικής γλώσσας GATE. [144] ..... 203

Εικόνα 45: Ενδεικτικό αποτέλεσμα της εφαρμογής της γραμματικής για τα Ελληνικά σε μια πρόταση. ..... 205

# ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 1: Τμήμα εγγράφου στο οποίο οι λεκτικές μονάδες με τονισμένη γραφή αποτελούν ονόματα οντοτήτων. .....48

Πίνακας 2: Το συμπληρωμένο σχεδιότυπο που πρέπει να παραχθεί από ένα σύστημα εξαγωγής πληροφορίας όταν εφαρμοστεί στο τμήμα εγγράφου του πίνακα: Πίνακας 1. .....49

Πίνακας 3: Αποτελέσματα του υβριδικού αναγνωριστή μερών του λόγου, για την θεματική περιοχή των περιγραφών εκθεμάτων μουσείων. .....65

Πίνακας 4: Τμήμα εγγράφου στο οποίο οι λεκτικές μονάδες με τονισμένη γραφή αποτελούν ονόματα οντοτήτων. .....67

Πίνακας 5: Τμήμα εγγράφου με αμφίσημα ονόματα οντοτήτων. .....68

Πίνακας 6: Τμήμα εγγράφου επισημειωμένο κατά τα πρότυπα του MUC7 [65]. .....69

Πίνακας 7: Παράδειγμα ισοδυναμίας κλάσεων κατά τα πρότυπα του MUC7 [65]. .....70

Πίνακας 8: Παράδειγμα συμπληρωμένου σχεδιότυπου οντότητας κατά τα πρότυπα του MUC7 [65]. .....70

Πίνακας 9: παράδειγμα συσχέτισης οντοτήτων κατά τα πρότυπα του MUC7 [65]. .....71

Πίνακας 10: Οι εργασίες αξιολόγησης των συνεδρίων MUC-3 έως MUC-7. .....71

Πίνακας 11: Μέγιστα αποτελέσματα για κάθε εργασία αξιολόγησης (MUC-3 έως MUC-7). .....72

Πίνακας 12: Παραδείγματα “σίγουρων κανόνων”. Σαν Xxxx+ σημειώνεται μια σειρά από λέξεις που αρχίζουν με κεφαλαίο χαρακτήρα, σαν DD ένας αριθμός, σαν PROF ένα επάγγελμα (director, manager, analyst κ.α.), σαν REL μια σχέση (sister, nephew κ.α.), σαν JJ* μια σειρά από κανένα ή περισσότερα επίθετα, σαν LOC μια γνωστή τοποθεσία, σαν PERSON-NAME ένα έγκυρο όνομα προσώπου αναγνωρισμένο από γραμματική ονομάτων προσώπων. [73] .....74

Πίνακας 13: Η απόδοση του συστήματος LTG έχοντας ολοκληρώσει διάφορα στάδια του συστήματος. R = ανάκτηση (recall) P = ακρίβεια (precision). .....76

Πίνακας 14: Η απόδοση των χειρωνακτικά κατασκευασμένων συστημάτων αναγνώρισης ονομάτων οντοτήτων. .....85

Πίνακας 15: Αποτελέσματα για το πείραμα με τις τρεις κατηγορίες (πρόσωπο – οργανισμός – όχι-NE). (DT: δέντρο απόφασης με τη συμβολική αντιπροσώπευση

(ενότητα 4.6.4), NN: νευρικό δίκτυο με την αριθμητική αντιπροσώπευση (ενότητα 4.6.5), DT-N: δέντρο απόφασης με την αριθμητική αντιπροσώπευση). ..... 86

Πίνακας 16: Αποτελέσματα για το πείραμα με τις δύο κατηγορίες (NE – όχι-NE). (DT: δέντρο απόφασης με τη συμβολική αντιπροσώπευση (ενότητα 4.6.4), NN: νευρικό δίκτυο με την αριθμητική αντιπροσώπευση (ενότητα 4.6.5), DT-N: δέντρο απόφασης με την αριθμητική αντιπροσώπευση). ..... 87

Πίνακας 17: Αποτελέσματα για το πείραμα με τις δύο κατηγορίες (πρόσωπο – οργανισμός). (DT: δέντρο απόφασης με τη συμβολική αντιπροσώπευση (ενότητα 4.6.4), NN: νευρικό δίκτυο με την αριθμητική αντιπροσώπευση (ενότητα 4.6.5), DT-N: δέντρο απόφασης με την αριθμητική αντιπροσώπευση). ..... 88

Πίνακας 18: Το πλήρες σύνολο ετικετών του αναγνωριστή ονομάτων οντοτήτων σε επίπεδο λέξης *(token-based NERC)*. ..... 90

Πίνακας 19: Τα χαρακτηριστικά του σωμάτων κειμένων αποτίμησης (θεματική περιοχή: αγγελίες προσφοράς εργασίας). ..... 97

Πίνακας 20: Αναλυτικά αποτελέσματα (ανά τύπο οντότητας) του συστήματος ML-HNERC για την Ελληνική γλώσσα. ..... 98

Πίνακας 21: Αναλυτικά αποτελέσματα (ανά τύπο οντότητας) του συστήματος ML-HNERC για την Αγγλική γλώσσα. ..... 100

Πίνακας 22: Αναλυτικά αποτελέσματα (ανά τύπο οντότητας) του συστήματος ML-ENERC για την Αγγλική γλώσσα. ..... 101

Πίνακας 23: Τα χαρακτηριστικά του σωμάτων κειμένων αποτίμησης (θεματική περιοχή: περιγραφές φορητών υπολογιστών από ιστοσελίδες ηλεκτρονικών καταστημάτων).. 102

Πίνακας 24: Αναλυτικά αποτελέσματα (ανά τύπο οντότητας) του συστήματος ML-HNERC για την Ελληνική γλώσσα (θεματική περιοχή: περιγραφές φορητών υπολογιστών από ιστοσελίδες ηλεκτρονικών καταστημάτων). ..... 103

Πίνακας 25: Αναλυτικά αποτελέσματα (ανά τύπο οντότητας) του συστήματος ML-HNERC για την Ελληνική γλώσσα, χρησιμοποιώντας εμπλουτισμένες λίστες γνωστών ονομάτων οντοτήτων. ..... 105

Πίνακας 26: Οι ετικέτες επιπρόσθετης μορφολογικής πληροφορίας για κάθε μέρος του λόγου. ..... 108

Πίνακας 27: Κατανομή των ονομάτων οντοτήτων στις τρεις σημασιολογικές κατηγορίες, για τα δεδομένα εκπαίδευσης και αξιολόγησης. ..... 109

Πίνακας 28: Παράδειγμα παραγόμενων διανυσμάτων από ένα στιγμιότυπο ονόματος οντότητας σε ένα κείμενο. ..... 110

Πίνακας 29: Αποτελέσματα (ανά τύπο οντότητας) του συστήματος προσαρμογής/εμπλουτισμού του υποσυστήματος του λεξικού για την Αγγλική γλώσσα ..... 111

Πίνακας 30: Αποτελέσματα αξιολόγησης του συστήματος ενημέρωσης αναγνωριστή ονομάτων οντοτήτων. ..... 114

Πίνακας 31: Παράδειγμα μιας απλής CFG, μαζί με το πλήρες σύνολο προτάσεων που μπορούν να παραχθούν ή να αναλυθούν από αυτήν την γραμματική, όπως παρουσιάζεται στην εργασία [13]. ..... 125

Πίνακας 32: Υπολογίζοντας το μήκος περιγραφής (GDL) μιας γραμματικής $G$. ..... 130

Πίνακας 33: Υπολογίζοντας την συνεισφορά στο μήκος *DDL* δύο προτάσεων. ..... 131

Πίνακας 34: Η γραμματική $G$ και οι αντίστοιχες συχνότητες κανόνων. ..... 131

Πίνακας 35: Μετατροπή προτάσεων εκπαίδευσης σε μια αρχική γραμματική. ..... 135

Πίνακας 36: Η επίδραση του τελεστή CreateNT, όπως παρουσιάζεται στην αναφορά [13]. ..... 136

Πίνακας 37: Η επίδραση του τελεστή MergeNT, όπως παρουσιάζεται στην αναφορά [13]. ..... 141

Πίνακας 38: Συγχωνεύοντας σύνολα από κανόνες. ..... 147

Πίνακας 39: Η επίδραση του τελεστή CreateOptionalNT. ..... 151

Πίνακας 40: Αναλύοντας το αποτέλεσμα της επίδρασης του τελεστή CreateOptionalNT. ..... 151

Πίνακας 41: Αντικαθιστώντας όλες τις εμφανίσεις του $WXY$, όπου το $WX$ μπορεί να επεκταθεί με το προαιρετικό σύμβολο $Y$. ..... 152

Πίνακας 42: Δύο τεχνητές γραμματικές: η γραμματική (a) περιλαμβάνει τυχαίες σειρές επιθέτων, και η (b) υποστηρίζει τυχαίες δευτερεύουσες προτάσεις (relative clauses) [13]. ..... 158

Πίνακας 43: Συνοπτική περιγραφή των προβλημάτων του διαγωνισμού “Omphalos”. 167

Πίνακας 44: Τα αποτελέσματα της αξιολόγησης του egGRIDS+ στην εργασία της αναγνώρισης συσχετίσεων μεταξύ ονομάτων οντοτήτων. ..... 177

Πίνακας 45: Τα αποτελέσματα της αξιολόγησης του egGRIDS+ στην εργασία της αναγνώρισης συσχετίσεων μεταξύ ονομάτων οντοτήτων (μοναδικά παραδείγματα εκπαίδευσης). ..... 179

Πίνακας 46: Τα αποτελέσματα της αξιολόγησης της προσέγγισης που βασίζεται στον CRF++, στην εργασία της αναγνώρισης συσχετίσεων μεταξύ ονομάτων οντοτήτων.. 181

Πίνακας 47: Τα αποτελέσματα της αξιολόγησης της εναλλακτικής (δεύτερης) προσέγγισης που βασίζεται στον CRF++, στην εργασία της αναγνώρισης συσχετίσεων μεταξύ ονομάτων οντοτήτων..... 182

Πίνακας 48: Το πλήρες σύνολο των ετικετών του αναγνωριστή μερών του λόγου για την Ελληνική γλώσσα..... 202

Πίνακας 49: Ενδεικτικά παραδείγματα από το λεξικό του χειρονακτικού αναγνωριστή ονομάτων οντοτήτων. ..... 204

Πίνακας 50: Αποτελέσματα αξιολόγησης του συστήματος VIE [84] και του χειρωνακτικού αναγνωριστή ονομάτων για τα Ελληνικά. ..... 206

# ΠΡΟΛΟΓΟΣ

Η παρούσα διατριβή δεν θα είχε πραγματοποιηθεί ποτέ αν δεν υπήρχε το Ινστιτούτο Πληροφορικής και Τηλεπικοινωνιών του Εθνικού Κέντρου Έρευνας Φυσικών Επιστημών (Ε.Κ.Ε.Φ.Ε.) “Δημόκριτος”, το οποίο μέσω εξετάσεων εμπιστεύτηκε έναν Φυσικό για να εκπονήσει διδακτορική διατριβή σε ένα θέμα που άπτεται της Επιστήμης Υπολογιστών. Η παρούσα διατριβή πραγματοποιήθηκε με τη βοήθεια υποτροφίας που χορηγήθηκε από το Ινστιτούτο Πληροφορικής και Τηλεπικοινωνιών του Εθνικού Κέντρου Έρευνας Φυσικών Επιστημών (Ε.Κ.Ε.Φ.Ε.) “Δημόκριτος”. Ευχαριστώ όλους όσοι συντέλεσαν στην επιτυχή ολοκλήρωση της διατριβής αυτής.

Ευχαριστώ ιδιαίτερα τον κ. Κωνσταντίνο Σπυρόπουλο, Διευθυντή Έρευνας του Ε.Κ.Ε.Φ.Ε. “Δημόκριτος” για τα χρήσιμα σχόλιά του καθ’ όλη τη διάρκεια του διδακτορικού, στο τελικό κείμενο της διατριβής, καθώς και για το γεγονός ότι εξασφάλισε χρηματοδότηση για μια σειρά από θερινά σχολεία και συνέδρια τα οποία με βοήθησαν να προσεγγίσω περιοχές στις οποίες δεν είχα καμία εμπειρία (όπως αυτή της επαγωγικής εξαγωγής γραμματικών) και ανέδειξαν την ερευνητική μου εργασία σε διεθνή επίπεδο. Ευχαριστώ ιδιαίτερα τον κ. Ευάγγελο Καρκαλέτση, Διευθυντή Έρευνας του Ε.Κ.Ε.Φ.Ε. “Δημόκριτος”, ο οποίος με εισήγαγε στον χώρο της επεξεργασίας φυσικής γλώσσας και συνεισέφερε ουσιαστικά στην βελτίωση του τρόπου συγγραφής των επιστημονικών μου άρθρων. Επίσης, ευχαριστώ ιδιαίτερα των Γεώργιο Παλιούρα, Ερευνητή Β του Ε.Κ.Ε.Φ.Ε. “Δημόκριτος”, ο οποίος με εισήγαγε στην περιοχή της μηχανικής μάθησης. Τον ευχαριστώ για τις δημιουργικές συζητήσεις που είχαμε σχετικά με την ανάπτυξη του αλγορίθμου egGRIDS+, και για την συνεισφορά του στην βελτίωση του τρόπου συγγραφής των επιστημονικών μου άρθρων. Ευχαριστώ επίσης τα υπόλοιπα δύο μέλη της τριμελούς μου επιτροπής από το Πανεπιστήμιο Αθηνών, τον καθηγητή κ. Κωνσταντίνο Χαλάτση και τον επίκουρο καθηγητή κ. Παναγιώτη Σταματόπουλο για τη δημιουργική συνεργασία που είχαμε, αλλά και την υπομονή που επέδειξαν.

Όσον αφορά την εκπόνηση της παρούσας διατριβής, θα ήθελα να ευχαριστήσω τους Pat Langley, διευθυντή του Center for the Study of Language and Information, του Stanford University, αλλά και του Sean Stromsten, Lead Research Engineer at BAE Systems Advanced Information Technologies, για τον κώδικα που μου έστειλαν. Θα ήθελα επίσης να ευχαριστήσω την Claire Grover, Senior Research Fellow, School of Informatics, University of Edinburgh, για τα σώματα κειμένων και τα δεδομένα αξιολόγησης συστημάτων του Πανεπιστημίου του Εδιμβούργου.

Ευχαριστώ τον Δρ. Γεώργιο Συγλέτο, για την βοήθεια που μου έδωσε όταν την χρειάστηκα. Ευχαριστώ τον Δρ. Δημήτρη Σπηλιωτόπουλο για την φιλία και υποστήριξή του.

Τέλος, ένα πολύ μεγάλο ευχαριστώ στους γονείς μου και τα αδέρφια μου, που με στήριξαν οικονομικά και ψυχολογικά από το πρώτο έτος των βασικών σπουδών μου έως και σήμερα, συνεισφέροντας σημαντικά στην επιτυχή ολοκλήρωση της διατριβής αυτής.

# 1. Εισαγωγή

Η παρούσα διατριβή εξετάζει την χρήση τεχνικών μηχανικής μάθησης σε διάφορα στάδια της επεξεργασίας φυσικής γλώσσας, κυρίως για σκοπούς εξαγωγής πληροφορίας από κείμενα. Στόχος είναι τόσο η βελτίωση της προσαρμοστικότητας των συστημάτων εξαγωγής πληροφορίας σε νέες θεματικές περιοχές (ή ακόμα και γλώσσες), όσο και η επίτευξη καλύτερης απόδοσης χρησιμοποιώντας όσο το δυνατό λιγότερους πόρους (τόσο γλωσσικούς όσο και ανθρώπινους).

Η διατριβή κινείται σε δύο κύριους άξονες: α) την έρευνα και αποτίμηση υπαρχόντων αλγορίθμων μηχανικής μάθησης κυρίως στα στάδια της προ-επεξεργασίας (όπως η αναγνώριση μερών του λόγου) και της αναγνώρισης ονομάτων οντοτήτων, και β) τη δημιουργία ενός νέου αλγορίθμου μηχανικής μάθησης και αποτίμησής του, τόσο σε συνθετικά δεδομένα, όσο και σε πραγματικά δεδομένα από το στάδιο της εξαγωγής σχέσεων μεταξύ ονομάτων οντοτήτων.

Στις επόμενες ενότητες θα παρουσιαστεί μια γενική εισαγωγή σε τεχνολογίες εντοπισμού (ανάκτησης, φιλτραρίσματος και εξαγωγής) πληροφορίας, και θα οριστεί το αντικείμενο της διατριβής αυτής (ενότητα 1.2). Η προσέγγιση που ακολουθήθηκε στα πλαίσια αυτής της διατριβής παρουσιάζεται στην ενότητα 1.3, ενώ η ενότητα 1.4 παρουσιάζει με συνοπτικό τρόπο τη συνεισφορά στην ερευνητική περιοχή της χρήσης μηχανικής μάθησης για εξαγωγή πληροφορίας από κείμενα. Οι σχετικές με τη διατριβή δημοσιεύσεις και ο αριθμός των ετεροαναφορών σε αυτές παρουσιάζονται στην ενότητα 1.5, ενώ η ενότητα 1.6 ολοκληρώνει αυτό το κεφάλαιο με την παρουσίαση της διάρθρωσης της διατριβής.

## 1.1 Σύντομη επισκόπηση του αντικειμένου της διατριβής

Η σημαντική τεχνολογική ανάπτυξη της πληροφορικής και κυρίως η ραγδαία εξάπλωση του παγκόσμιου ιστού έχουν επιτρέψει την πρόσβαση σε τεράστιο όγκο πληροφοριών και γνώσης. Ωστόσο, η ευκολία προσπέλασης των πληροφοριών δεν συνοδεύεται από τον εύκολο εντοπισμό τους, καθώς η άναρχη δομή του διαδικτύου και ο τεράστιος όγκος της περιεχόμενης πληροφορίας καθιστούν τον εντοπισμό συγκεκριμένων πληροφοριών αρκετά ενδιαφέρον αλλά δύσκολο πρόβλημα. Ο εντοπισμός της επιθυμητής πληροφορίας γίνεται ακόμα πιο περίπλοκος, αν αναλογιστούμε το γεγονός ότι η πληροφορία μπορεί να είναι εκφρασμένη σε περισσότερες από μία γλώσσες. Επομένως, είναι αναγκαία η έρευνα και ανάπτυξη κατάλληλων τεχνολογιών οι οποίες να μπορούν να εξάγουν από τον τεράστιο όγκο της διαθέσιμης πληροφορίας την πληροφορία που ενδιαφέρει τον χρήστη.

Το πρόβλημα αυτό είναι γνωστό σαν *υπερπληροφόρηση (information overloading)*: ο χρήστης κυριολεκτικά “βομβαρδίζεται” με πληροφορία, αδυνατώντας να την φιλτράρει με αποδοτικό τρόπο, ώστε να εντοπίσει την πληροφορία που τον αφορά και ενδιαφέρει. Για την αντιμετώπιση του προβλήματος της υπερπληροφόρησης διάφορες τεχνολογίες έχουν προταθεί, με κύριους εκπροσώπους την *ανάκτηση πληροφορίας (information retrieval)*, το *φιλτράρισμα πληροφορίας (information filtering)* και την *εξαγωγή πληροφορίας (information extraction)*. Κάθε τεχνολογία προσπαθεί να καλύψει διαφορετικές ανάγκες αναζήτησης, και σαν αποτέλεσμα παρουσιάζει διαφορετικές απαιτήσεις. Οι δύο πρώτες τεχνολογίες (δηλ. η ανάκτηση και το φιλτράρισμα πληροφορίας) επικεντρώνονται κυρίως σε επίπεδο εγγράφου, αφού προσπαθούν να εντοπίσουν συγκεκριμένα έγγραφα που περιέχουν τις επιθυμητές πληροφορίες - συνήθως βασιζόμενες σε λέξεις κλειδιά (ανάκτηση πληροφορίας) - ή να κατηγοριοποιήσουν έγγραφα σε θεματικές κατηγορίες (φιλτράρισμα πληροφορίας). Αντίθετα, η εξαγωγή πληροφορίας έχει ως στόχο τον εντοπισμό της επιθυμητής

πληροφορίας εντός των εγγράφων, την εξαγωγή της πληροφορίας και την παρουσίασή της στον χρήστη σε συνοπτική μορφή. Ακριβώς επειδή η εξαγωγή πληροφορίας παρουσιάζει συνοπτικά την επιθυμητή πληροφορία, αντί να εντοπίζει μόνο τα έγγραφα που την περιέχουν, μπορεί να βοηθήσει στην προσπέλαση πληροφορίας εκφρασμένη σε διαφορετικές γλώσσες.

Στην διεθνή βιβλιογραφία έχουν παρουσιαστεί μια πληθώρα συστημάτων εξαγωγής πληροφορίας, ενώ υπάρχουν και αρκετά εμπορικά συστήματα. Τα περισσότερα συστήματα εξαγωγής πληροφορίας ωστόσο, εξάγουν πληροφορία από συγκεκριμένες θεματικές περιοχές, εκφρασμένη σε συγκεκριμένη γλώσσα. Για να μπορέσει η τεχνολογία εξαγωγής πληροφορίας να εφαρμοστεί στην πράξη, πρέπει τα συστήματα εξαγωγής πληροφορίας να μπορούν εύκολα να προσαρμόζονται σε νέες θεματικές περιοχές (δηλ. νέες κατηγορίες πληροφορίας που πρέπει να εξαχθεί, πιθανώς από κείμενα διαφορετικού τύπου) και γλώσσες. Την τελευταία δεκαετία έχει επιτευχθεί σημαντική πρόοδος όσον αφορά την ανάπτυξη αξιόπιστων συστημάτων εξαγωγής πληροφορίας, καθώς έχουν αναπτυχθεί εφαρμογές εξαγωγής πληροφορίας για διάφορες θεματικές περιοχές, όπως η εξαγορά εταιριών [1], [2] [3], η μεταβίβαση μετοχών [4], κέρδη και ζημίες εταιριών [5], μετακινήσεις στελεχών επιχειρήσεων [6], [7], [8] καθώς και κατανόησης στρατιωτικών μηνυμάτων [9] και αστυνομικών αναφορών [10], [11], [12]. Ταυτόχρονα, η πλειοψηφία των συστημάτων αυτών αφορούν συγκεκριμένες γλώσσες (ευρέως ομιλούμενες) και κυρίως την Αγγλική και Ιαπωνική γλώσσα.

Λόγω του γεγονότος ότι η εξαγωγή πληροφορίας προϋποθέτει μια μορφή κατανόησης των κειμένων, η τεχνολογία αυτή εμφανίζεται περισσότερο απαιτητική, όσον αφορά την επεξεργασία της φυσικής γλώσσας, σε σχέση με τις υπόλοιπες δύο τεχνολογίες. Σαν αποτέλεσμα, τα συστήματα εξαγωγής πληροφορίας εμφανίζουν ισχυρότερη εξάρτηση από την θεματική περιοχή και γλώσσα σε σχέση με τα συστήματα των άλλων δύο κατηγοριών. Ταυτόχρονα, επειδή η επεξεργασία φυσικής γλώσσας είναι μια αρκετά απαιτητική εργασία, κυρίως λόγω της αμφισημίας της και των γλωσσικών πόρων που απαιτούνται, τα σύγχρονα συστήματα εξαγωγής πληροφορίας επιλέγουν να ασχοληθούν με την επεξεργασία υπογλωσσών, δηλαδή με το περιορισμένο εκείνο υποσύνολο μιας γλώσσας που απαντάται σε μια περιορισμένη θεματική περιοχή. Ένας από τους στόχους αυτής της διδακτορικής διατριβής είναι η υποβοήθηση των διαφόρων υποσυστημάτων επεξεργασίας φυσικής γλώσσας με τεχνικές μηχανικής μάθησης, είτε βοηθώντας την απόκτηση κάποιων απαραίτητων γλωσσικών πόρων ή αντικαθιστώντας ολόκληρα υποσυστήματα, με απώτερο σκοπό την διευκόλυνση προσαρμογής αυτών σε νέες θεματικές περιοχές ή ακόμα και γλώσσες.

Η εξαγωγή πληροφορίας μπορεί να χωριστεί σε δύο υπο-εργασίες: α) την αναγνώριση ονομάτων οντοτήτων (όπως ονόματα ανθρώπων, οργανισμών ή τοποθεσιών) που συμμετέχουν σε ένα γεγονός και β) την αναγνώριση των συσχετίσεων μεταξύ των οντοτήτων που μετέχουν στο γεγονός. Η πρώτη υπο-εργασία αποτελεί βασικό συστατικό κάθε συστήματος εξαγωγής πληροφορίας. Στην πλειονότητα των υπαρχόντων συστημάτων οι απαιτήσεις από το σύστημα αναγνώρισης ονομάτων οντοτήτων είναι σαφώς καθορισμένες, και αφορούν τον εντοπισμό ονομάτων οντοτήτων και την κατηγοριοποίησή τους σε κατάλληλες θεματικές κατηγορίες. Η δεύτερη υπο-εργασία βασίζεται και αξιοποιεί τα αποτελέσματα του αναγνωριστή ονομάτων οντοτήτων. Σε αντίθεση όμως με την αναγνώριση ονομάτων οντοτήτων, δεν υπάρχουν κοινές απαιτήσεις για την υπο-εργασία αυτή μεταξύ των διαφόρων συστημάτων, καθώς οι απαιτήσεις συχνά εξαρτώνται από τον βαθμό πολυπλοκότητας και ειδίκευσης του κάθε συστήματος.

## 1.2 Ορισμός προβλήματος

Η παρούσα διδακτορική διατριβή ερευνά τη δυνατότητα αξιοποίησης τεχνικών μηχανικής μάθησης στην περιοχή της επεξεργασίας φυσικής γλώσσας, με σκοπό την αντιμετώπιση του προβλήματος της αναβάθμισης καθώς και της προσαρμογής συστημάτων επεξεργασίας φυσικής γλώσσας σε νέες θεματικές περιοχές ή γλώσσες. Η έρευνα οριοθετείται σε τρεις σημαντικούς άξονες ενός συστήματος εξαγωγής πληροφορίας:

- Αναγνώριση μερών του λόγου για την Ελληνική γλώσσα.
- Αναγνώριση ονομάτων οντοτήτων.
- Αναγνώριση σχέσεων ανάμεσα σε αναγνωρισμένα ονόματα οντοτήτων.

Η διατριβή εξετάζει το πώς μπορούν να αξιοποιηθούν μέθοδοι και τεχνικές μηχανικής μάθησης για την κατασκευή συστημάτων που υποστηρίζουν τις εργασίες αυτές, τα οποία θα προσαρμόζονται ευκολότερα σε νέες θεματικές περιοχές και γλώσσες σε σχέση με τα χειρωνακτικά κατασκευασμένα συστήματα που βασίζονται σε κανόνες, κατασκευασμένους συχνά από ειδικούς.

## 1.3 Η προσέγγιση της διατριβής

Ειδικότερα, η διατριβή ερευνά τεχνικές μηχανικής μάθησης προς δύο κύρους άξονες:

1. Την εφαρμογή υπαρχουσών τεχνικών (τόσο συμβολικών όσο και στατιστικών) σε επιλεγμένα στάδια της εξαγωγής πληροφορίας. Οι τεχνικές αυτές αποτιμούνται συγκριτικά μεταξύ τους σε κείμενα τόσο στην Ελληνική όσο και την Αγγλική γλώσσα. Όλοι οι υπάρχοντες αλγόριθμοι μηχανικής μάθησης που εξετάστηκαν χρειάζονται σαν είσοδο ένα διάνυσμα σταθερού μήκους. Ωστόσο η μετατροπή της φυσικής γλώσσας σε διάνυσμα σταθερού μήκους δεν είναι πάντα εύκολη χωρίς την χρήση αυθαιρέτων ορίων όσον αφορά τον μέγιστο αριθμό λέξεων. Η παρατήρηση αυτή αποτέλεσε και το βασικό κίνητρο για την δημιουργία ενός νέου αλγορίθμου μηχανικής μάθησης χωρίς τον περιορισμό των διανυσμάτων σταθερού μήκους σαν είσοδο.
2. Την ανάπτυξη ενός νέου αλγορίθμου μηχανικής μάθησης, χωρίς την απαίτηση για είσοδο διανυσμάτων σταθερού μήκους. Ο νέος αυτός αλγόριθμος μαθαίνει *γραμματικές ανεξάρτητες από τα συμφραζόμενα (context free grammars)* από θετικά παραδείγματα, με καθοδήγηση μέσω ευριστικών, όπως το *ελάχιστο μήκος περιγραφής (minimum description length)*.

Όσον αφορά τον πρώτο άξονα, αναπτύχθηκαν και αξιολογήθηκαν συστήματα αναγνώρισης ονομάτων οντοτήτων, βασισμένα σε υπάρχοντες αλγορίθμους μηχανικής μάθησης, όπως δέντρα αποφάσεων και νευρωνικά δίκτυα. Τα συστήματα που αναπτύχθηκαν αφορούν διάφορες θεματικές περιοχές (μετακινήσεις στελεχών επιχειρήσεων, χρηματο-οικονομικές ειδήσεις, δικαστικές αποφάσεις) τόσο στην Ελληνική όσο και στην Αγγλική γλώσσα. Τα συστήματα αυτά αξιολογήθηκαν σε κείμενα της Ελληνικής γλώσσας, και οδήγησαν στον προσδιορισμό των μειονεκτημάτων και των περιορισμών που επιβάλλουν οι εξετασθέντες αλγόριθμοι όταν εφαρμόζονται σε δεδομένα φυσικής γλώσσας. Από την παραπάνω ανάλυση, προέκυψε ότι ένα από τα σημαντικότερα προβλήματα της εφαρμογής μηχανικής μάθησης είναι η δυσκολία διαχείρισης δεδομένων μεταβλητού μήκους, όπως π.χ. χαρακτηριστικά που αφορούν όλες τις λέξεις μιας πρότασης. Αντίθετα, ένας συντακτικός αναλυτής μπορεί να εξετάσει εύκολα αν μια πρόταση ή μέρος αυτής περιγράφεται από μια δεδομένη γραμματική. Ωστόσο, η χειροκίνητη ανάπτυξη γραμματικών κατάλληλων για μια εργασία είναι μια σύνθετη διαδικασία, ενώ τα αποτελέσματα συχνά εξαρτώνται από την θεματική περιοχή και σαφώς από την γλώσσα. Συνεπώς, αν μια τέτοια γραμματική είναι δυνατόν να

αποκτηθεί αυτόματα με την χρήση μηχανικής μάθησης, τότε η προσαρμογή συστημάτων που χρησιμοποιούν τέτοιες γραμματικές σε νέες θεματικές περιοχές ή γλώσσες, είναι δυνατόν να απλοποιηθεί σημαντικά.

Η συμβολή των συστημάτων που αναπτύχθηκαν είναι σημαντική, τόσο σε Ελληνικό όσο και σε διεθνές επίπεδο. Τα συστήματα αναγνώρισης ονομάτων οντοτήτων που αναπτύχθηκαν για την Ελληνική γλώσσα είναι τα πρώτα συστήματα στο είδος τους που αναφέρονται στη βιβλιογραφία. Ταυτόχρονα, η απόδοση των υλοποιηθέντων συστημάτων κρίνεται ιδιαίτερα ικανοποιητική, αφού είναι συγκρίσιμα με τις αποδόσεις συστημάτων που παρουσιάζονται στην διεθνή βιβλιογραφία την αντίστοιχη χρονική περίοδο.

Όσον αφορά το δεύτερο άξονα, και έχοντας σαν στόχο την αντιμετώπιση των προβλημάτων που εμφανίζει η εφαρμογή υπαρχουσών τεχνικών, αναπτύχθηκε μία νέα τεχνική μηχανικής μάθησης. Η νέα τεχνική εντάσσεται στην κατηγορία της *επαγωγικής εξαγωγής γραμματικών (inductive grammar learning)*. Τα κύρια πλεονεκτήματα της μεθόδου αυτής σε σχέση με άλλες μεθόδους μηχανικής μάθησης, είναι η δυνατότητα χειρισμού δεδομένων σε μορφή κειμένου, καθώς και η δυνατότητα ενσωμάτωσής της σε υπάρχοντα συστήματα αντικαθιστώντας χειρωνακτικά κατασκευασμένες γραμματικές. Κύριος στόχος της νέας αυτής τεχνικής είναι η αυτοματοποίηση της διαδικασίας δημιουργίας γραμματικών, οι οποίες να μπορούν να συνεργαστούν με την πληθώρα των συντακτικών αναλυτών που εμφανίζονται στην διεθνή βιβλιογραφία, αντικαθιστώντας υπάρχουσες (και πιθανώς χειρωνακτικά κατασκευασμένες) γραμματικές για διάφορες υπο-εργασίες συστημάτων εξαγωγής πληροφορίας.

Για την εφαρμογή της επαγωγικής εξαγωγής γραμματικών *αναπτύχθηκε ένας νέος αλγόριθμος επαγωγικής εξαγωγής γραμματικών* που λειτουργεί μόνο με θετικά παραδείγματα. Ο νέος αυτός αλγόριθμος μπορεί να επάγει *γραμματικές ανεξάρτητες από τα συμφραζόμενα (context free grammars)*, και βασίστηκε σε έναν υπάρχοντα αλγόριθμο [13], βελτιώνοντας τόσο το χρησιμοποιούμενο ευριστικό, όσο και την διαδικασία αναζήτησης στον χώρο των πιθανών γραμματικών, αυξάνοντας ταυτόχρονα την εφαρμοσιμότητα του νέου αλγορίθμου σε μεγαλύτερα σύνολα δεδομένων. Η απαίτηση ο αλγόριθμος να λειτουργεί μόνο με θετικά παραδείγματα προέρχεται από την συχνή ανυπαρξία αρνητικών παραδειγμάτων στην περιοχή της επεξεργασίας φυσικής γλώσσας. Σημειώνεται ότι η παρουσία αρνητικών παραδειγμάτων αποτελεί προϋπόθεση για την λειτουργία της μεγαλύτερης πλειοψηφίας των υπαρχόντων αλγορίθμων εξαγωγής γραμματικών. Η σχεδίαση του νέου αλγορίθμου έγινε με τέτοιο τρόπο ώστε να μπορεί να χρησιμοποιηθεί σε εργασίες κατηγοριοποίησης, όπως π.χ. στην αναγνώριση ονομάτων οντοτήτων. Η λειτουργία αυτή είναι διαφορετική από την συνήθη εφαρμογή αλγορίθμων εξαγωγής γραμματικών, καθώς δεν απαιτείται ούτε ο χαρακτηρισμός προτάσεων ως γραμματικά ορθές ή μη, ούτε η συντακτική ανάλυση προτάσεων, αλλά μόνο ο εντοπισμός *τμημάτων των προτάσεων* και η κατηγοριοποίησή τους σε κατάλληλες κατηγορίες. Η αποτίμηση του αλγορίθμου αυτού έγινε τόσο σε συνθετικές γλώσσες, όσο και στην υπο-εργασία της εξαγωγής σχέσεων μεταξύ ονομάτων οντοτήτων, ενώ συμμετείχε και στον διεθνή διαγωνισμό “Omphalos” [14]. Το αποτέλεσμα του νέου αλγορίθμου στον διαγωνισμό “Omphalos” ήταν σημαντικό, πετυχαίνοντας να επιλύσει το πρώτο πρόβλημα χωρίς ανθρώπινη επέμβαση εντός των προθεσμιών του διαγωνισμού, ενώ συνδυάζεται επιτυχώς και με τον αλγόριθμο που έλυσε το δυσκολότερο πρόβλημα, αφαιρώντας την ανάγκη που εμφάνιζε ο αλγόριθμος που έλυσε το δυσκολότερο πρόβλημα εντός των προθεσμιών του διαγωνισμού για ανθρώπινη παρέμβαση, σε περιπτώσεις όπου το ευριστικό του δεν μπορούσε να οδηγήσει περεταίρω την διαδικασία αναζήτησης λόγω ισοβαθμιών.

## 1.4 Συνεισφορά

Η εργασία για αυτή την διδακτορική διατριβή ξεκίνησε σε μια περίοδο όπου η μηχανική μάθηση άρχισε να βρίσκει ένα ενδιαφέρον πεδίο εφαρμογής, την επεξεργασία φυσικής γλώσσας. Η διδακτορική διατριβή συνέβαλε στην πορεία αυτή, εξετάζοντας την αξιοποίηση αρκετών μεθόδων μηχανικής μάθησης σε διάφορες φάσεις της διαδικασίας εξαγωγής πληροφορίας.

Η παρούσα διατριβή προτείνει την αξιοποίηση της μηχανικής μάθησης σε κομβικά σημεία ενός τυπικού συστήματος εξαγωγής πληροφορίας, έχοντας ως σκοπό την υποβοήθηση της προσαρμογής του συστήματος σε νέες περιοχές και ίσως και σε γλώσσες. Το πρώτο πεδίο έρευνας αυτής της διατριβής, αποτελεί η αναγνώριση μερών του λόγου για την Ελληνική γλώσσα. Η μάθηση στηριζόμενη σε *κανόνες μετασχηματισμού καθοδηγούμενη από σφάλματα (transformation-based error-driven learning – TBED)* [15] εφαρμόστηκε για πρώτη φορά στην Ελληνική γλώσσα, πετυχαίνοντας υψηλές αποδόσεις, άμεσα συγκρίσιμες με αντίστοιχα συστήματα για την Ελληνική γλώσσα, απαιτώντας ταυτόχρονα σημαντικά μειωμένα δεδομένα εκπαίδευσης. Ταυτόχρονα, η προσέγγιση που περιγράφεται σε αυτή την διατριβή αποτέλεσε τον πρώτο αναγνωριστή μερών του λόγου που διατέθηκε ελεύθερα σαν εφαρμογή ανοικτού λογισμικού, με σημαντική αποδοχή από την επιστημονική κοινότητα, όπως καταδεικνύει ο αριθμός των ετεροαναφορών στις σχετικές δημοσιεύσεις.

Το δεύτερο πεδίο έρευνας αφορά την περιοχή της αναγνώρισης ονομάτων οντοτήτων. Τέσσερις προσεγγίσεις βασισμένες σε μηχανική μάθηση (χρησιμοποιώντας τόσο συμβολικούς όσο και στοχαστικούς αλγορίθμους μηχανικής μάθησης) δοκιμάστηκαν σε αυτή την εργασία, πετυχαίνοντας ικανοποιητικά αποτελέσματα. Οι προσεγγίσεις δοκιμάστηκαν σε διάφορες θεματικές περιοχές, τόσο σε Αγγλικά, όσο και Ελληνικά, επιβεβαιώνοντας όχι μόνο την ικανότητα της μηχανικής μάθησης να υποστηρίξει την εργασία της αναγνώρισης των ονομάτων οντοτήτων, αλλά και την προσαρμοστικότητα των εν λόγω συστημάτων όχι μόνο σε νέες θεματικές περιοχές, αλλά και σε γλώσσες. Η εργασία που υλοποιήθηκε στα πλαίσια αυτής της διατριβής συγκαταλέγεται ανάμεσα στα πρώτα συστήματα εξαγωγής πληροφορίας για την Ελληνική γλώσσα που αναφέρονται στην διεθνή βιβλιογραφία. Επιπρόσθετα, στο πλαίσιο της πειραματικής αποτίμησης των αλγορίθμων μηχανικής μάθησης είχε παρατηρηθεί ότι, τουλάχιστον για την εργασία της αναγνώρισης ονομάτων οντοτήτων, η σειρά των λέξεων σε μια πρόταση δεν διαδραματίζει ιδιαίτερο λόγο [16]. Αν και αρχικά η συγκεκριμένη διαπίστωση προκάλεσε έκπληξη, εντούτοις η διαδεδομένη χρήση της αναπαράστασης *«συνόλου λέξεων» (bag-of-words representation)* – η οποία αγνοεί την σειρά των λέξεων – όχι μόνο για την αναγνώριση ονομάτων οντοτήτων, αλλά και για αρκετά ακόμα προβλήματα επεξεργασίας φυσικής γλώσσας, καταδεικνύει την ορθότητα της αρχικής εκείνης παρατήρησης.

Το τρίτο πεδίο έρευνας αφορά την ανάπτυξη ενός νέου αλγόριθμου μηχανικής μάθησης, και συγκεκριμένα ενός αλγορίθμου επαγωγικής εξαγωγής γραμματικών, ικανό να εξάγει γραμματικές ανεξάρτητες από συμφραζόμενα μόνο από θετικά παραδείγματα. Σημαντική ιδιότητα του νέου αυτού αλγορίθμου είναι η ικανότητα να επεξεργαστεί μεγάλους όγκους δεδομένων, απόρροια της παρατήρησης ότι είναι υπολογιστικά φθηνότερο να προβλέψεις το αποτέλεσμα της εφαρμογής των τελεστών εκμάθησης, παρά να τους εφαρμόσεις και να αποτιμήσεις το αποτέλεσμα. Σημαντικό ήταν το αποτέλεσμα του αλγορίθμου και στον διαγωνισμό “Omphalos” [14], όπου κατόρθωσε να επιλύσει το πρώτο πρόβλημα χωρίς ανθρώπινη επέμβαση εντός των προθεσμιών του διαγωνισμού, ενώ συνδυάζεται επιτυχώς και με τον αλγόριθμο που έλυσε το δυσκολότερο πρόβλημα, αφαιρώντας την ανάγκη που εμφάνιζε ο αλγόριθμος που

έλυσε το δυσκολότερο πρόβλημα εντός των προθεσμιών του διαγωνισμού για ανθρώπινη παρέμβαση, σε περιπτώσεις όπου το ευριστικό του δεν μπορούσε να οδηγήσει περεταίρω την διαδικασία αναζήτησης λόγω ισοβαθμιών.

## 1.5 Δημοσιεύσεις

Στα πλαίσια της διατριβής προέκυψαν οι παρακάτω δημοσιεύσεις. Επίσης, αναφέρεται ο αριθμός των ετεροαναφορών που εντοπίστηκαν μέχρι τη μέρα ολοκλήρωσης της διατριβής.

1. G. Petasis, V. Karkaletsis, G. Paliouras, C. D. Spyropoulos, “Learning context-free grammars to extract relations from text”. In Proceedings of the 18th European Conference on Artificial Intelligence (ECAI 2008), pp. 303 – 307, Patras, Greece, July 2008. (cited by 1)
    1. Ronald Fagin , Benny Kimelfeld , Yunyao Li , Sriram Raghavan , Shivakumar Vaithyanathan, Understanding queries in a search database system, Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems of data, June 06-11, 2010, Indianapolis, Indiana, USA.
2. G. Petasis, G. Paliouras, C. D. Spyropoulos and C. Halatsis, “eg-GRIDS: Context-Free Grammatical Inference from Positive Examples using Genetic Search”. In Proceedings of the 7th International Colloquium on Grammatical Inference (ICGI 2004), Lecture Notes in Artificial Intelligence 3264, pp. 223 – 234, Springer, 2004. (cited by 8)
    2. C de la Higuera, ‘Grammatical inference: learning automata and grammars’, Cambridge University Press, 2010.
    3. Mernik, M., Hrncic, D., Bryant, B.R., Sprague, A.P., Gray, J., Qichao Liu, Javed, F., “Grammar inference algorithms and applications in software engineering”. In Proceedings of Information, Communication and Automation Technologies, XXII International Symposium on Information, Communication and Automation Technologies (ICAT 2009), 29-31 Oct. 2009, Bosnia and Herzegovina.
    4. Salman Saghafi, “State-Machine Generation Using a Genetic Algorithm”. Masters project Dissertation, Department of Computer Science, University of Sheffield, 2009. [http://www.dcs.shef.ac.uk/intranet/teaching/projects/archive/msc2009/pdf/acr08ss.pdf]
    5. Massimiliano Di Penta, Pierpaolo Lombardi, Kunal Taneja , and Luigi Troiano, “Focus - Search-based inference of dialect grammars”, Soft Computing Journal - A Fusion of Foundations, Methodologies and Applications, Special Issue on Software Engineering and Soft Computing, Volume 12, Number 1, pp. 51-66, 2008. DOI: 10.1007/s00500-007-0216-5 [http://www.rcost.unisannio.it/mdipenta/papers/softcomp08.pdf]
    6. Nurfadhlina Mohd Sharef, Trevor Martin, Yun Shen, “Incremental Evolutionary Grammar Fragments”. In Proceedings of the 8th Annual UK Workshop on Computational Intelligence (UKCI'08), 2008. [http://www.cci.dmu.ac.uk/conferences/ukci2008/papers/Incremental-Evolutionary-Grammar-Fragments.pdf]
    7. Ernesto Luis Malta Rodrigues, “INFERÊNCIA DE GRAMÁTICAS FORMAIS LIVRES DE CONTEXTO UTILIZANDO COMPUTAÇÃO EVOLUCIONÁRIA COM APLICAÇÃO EM BIOINFORMÁTICA”. PhD Thesis, UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ, Brazil, 2007. [http://arquivos.cpgei.ct.utfpr.edu.br/Ano_2007/teses/Tese_30_2007.pdf]
    8. Faizan Javed, Marjan Mernik, Alan Sprague, and Barrett Bryant, “Incrementally Inferring Context-Free Grammars for Domain-Specific Languages”, Proceedings of The Eighteenth International Conference on Software Engineering and Knowledge Engineering (SEKE'06), July 5th -July 7th, pgs 363 - 368, San Francisco, CA, 2006.
    9. DeYoung, Mark E., “Dynamic Protocol Reverse Engineering A Grammatical Inference Approach”. Master Thesis. DEPARTMENT OF THE AIR FORCE, AIR UNIVERSITY, AIR FORCE INSTITUTE OF TECHNOLOGY, Air Force Base, Ohio, 2008. [http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA484312&Location=U2&doc=GetTRDoc.pdf]

3. G. Petasis, V. Karkaletsis, C. Grover, B. Hachey, M. T. Pazienza, M. Vindigni, J. Coch: "Adaptive, Multilingual Named Entity Recognition in Web Pages". In Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004), pp. 1073 – 1074, Valencia, Spain, August 22 – 27, 2004. [Extended Version]
4. G. Petasis, G. Paliouras, V. Karkaletsis, C. Halatsis, and C.D. Spyropoulos, "egGRIDS+: Computationally Efficient Grammatical Inference from Positive Examples". GRAMMARS, (7), pp. 69 – 110, 2004. (http://grammars.grlmc.com/special.asp) (cited by 11)
    10. C de la Higuera, 'Grammatical inference: learning automata and grammars', Cambridge University Press, 2010.
    11. R Reichart, A Rappoport "Unsupervised induction of labeled parse trees by clustering with syntactic features", in Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1, pp 721-8, 2008
    12. W Zuidema "An annotated bibliography of grammar induction models for natural language learning", paper at Institute for Logic, Language and Computation, University of Amsterdam, May 23, 2008
    13. R. Eyraud, C. de la Higuera and J-C Janodet, "LARS: A Learning Algorithm for Rewriting Systems", Machine Learning Journal, 2007, vol.66, No1, pp.7-31.
    14. G. Borensztajn, and W. Zuidema. 2007. Bayesian Model Merging for Unsupervised Constituent Labeling and Grammar Induction. Technical Report. ILLC.
    15. MATSUNO, I.P. Um Estudo do Processo de Inferência de Gramáticas Regulares e Livres de Contexto Baseados em Modelos Adaptativos. Dissertação de Mestrado, USP, São Paulo, 2006
    16. Alpana Dubey, "Inferring Grammar Rules from Programs", PhD Thesis, Indian Institute Of Technology Kanpur, 2006
    17. Rémi Eyraud, "Inférence grammaticale de langages hors-contextes", PhD thesis, University of Saint-Etienne, 2006
    18. C. de la Higuera, & J. Oncina (2006). Learning context-free languages. Artificial Intelligence Reviews.
    19. Ivone Penque Matsuno, "Um Estudo dos Processos de Inferencia de Gramaticas Regulares e Livres de Contexto Baseados em Modelos Adaptivos" Master Dissertation, 2006.
    20. R. Eyraud, C. de la Higuera and J-C Janodet, "Representing Languages by Learnable Rewriting Systems", Proceedings of the 7th International Colloquium on Grammatical Inference (ICGI), Lecture Notes in Computer Science, 3264, pp. 139- 150, 2004.
5. G. Petasis, V. Karkaletsis, and C. D. Spyropoulos, "Cross-lingual Information Extraction from Web pages: the use of a general-purpose Text Engineering Platform". In Proceedings of the 4th International Conference on Recent Advances in Natural Language Processing (RANLP 2003), Borovets, Bulgaria, pp 381 – 388, September 10 – 12, 2003.
6. G. Petasis, F. Vichot, F. Wolinski, G. Paliouras, V. Karkaletsis, C. D. Spyropoulos, "Using Machine Learning to Maintain Rule-based Named – Entity Recognition and Classification Systems". In Proceedings of the 39th Conference of Association for Computational Linguistics (ACL-EACL 2001), pp. 426 – 433, July 9 – 11 2001, Toulouse, France. (cited by 9)
    21. Gu, Baohua "Recognizing named entities in biomedical texts", Thesis (Ph.D.) - School of Computing Science - Simon Fraser University, 2008
    22. P Srikanth, KN Murthy "Named Entity Recognition for Telugu", in Proceedings of the Workshop NER for South and South East Asian Languages, in IJCNLP-08, 2008
    23. G. Lucarelli, X.Vasilakos and I. Androutsopoulos "Named Entity Recognition In Greek Texts With An Ensemble Of SMVS And Active Learning", International Journal on Artificial Intelligence Tools, vol. 16, No 6, pp-1015-1045, 2007.

24. Nadeau, David, "Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision", thesis, 2007

25. Georgios Lucarelli and Ion Androutsopoulos "A Greek Named-Entity Recognizer That Uses Support Vector Machines and Active Learning", Springer Berlin / Heidelberg, In "Lecture Notes in Computer Science", Advances in Artificial Intelligence, Volume 3955, pp 203-213, 2006.

26. I. Michailidis, K. Diamantaras, S. Vasileiadis, Y. Frère "Greek Named Entity Recognition using Support Vector Machines, Maximum Entropy and Onetime", LREC, 2006

27. D. Nadeau, P. D. Turney and S. Matwin, "Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity", Proceedings of the 19th Canadian Conference on Artificial Intelligence (CAI), pp. 266-277, 2006.

28. C. C. Yang and K. W. Li, "Automatic construction of English/Chinese parallel corpora", Journal of the American Society for Information Science and Technology, v. 54, n. 8, pp. 730 - 742, 2003.

29. F. Pachet, D. Laigre, "A Naturalist Approach to Music File Name Analysis," 2nd Annual International Symposium on Music Information Retrieval 2001.

7. G. Petasis, S. Petridis, G. Paliouras, V. Karkaletsis, S. Perantonis, and C.D. Spyropoulos, "Symbolic and Neural Learning of Named-Entity Recognition and Classification Systems in Two Languages", In Advances in Computational Intelligence and Learning: Methods and Applications, H–J. Zimmermann, G. Tselentis, M. van Someren and G. Dounias (eds), Kluwer Academic Publishers, 2001.

8. G. Petasis, S. Petridis, G. Paliouras, V. Karkaletsis, S. J. Perantonis, C. D. Spyropoulos, "Symbolic and Neural Learning for Named-Entity Recognition". In Proceedings of European Best Practice Workshops and Symposium on Computational Intelligence and Learning (COIL 2000), pp. 58 – 66, June 19 – 23 2000, Chios, Greece.

9. G. Petasis, A. Cucchiarelli, P. Velardi, G. Paliouras, V. Karkaletsis, C. D. Spyropoulos, "Automatic adaptation of Proper Noun Dictionaries through cooperation of machine learning and probabilistic methods". In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 128 – 135, July 24 – 28 2000, Athens, Greece. (cited by 16)

30. Seco Naveiras, Diego, "Técnicas de indexación y recuperación de documentos utilizando referencias geográficas y textuales", 2009

31. J. Wang and N. Ge, "Automatic feature thesaurus enrichment: extracting generic terms from digital gazetteer", Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital libraries (ICDL), pp. 326-333, 2006.

32. Feng Chen "Minimize Test Collection for Geographic Retrieval Evaluation", 2006.

33. D. Koning, I.N. Sarkar and T. Moritz, "TaxonGrab: Extracting Taxonomic Names From Text". Biodiversity Informatics, n.2, pp. 79-82, 2005.

34. B. Martins, M. J. Silva and M. Chaves, "Challenges and Resources for Evaluating Geographical IR", Proceedings of the Workshop on Geographic Information Retrieval, ACM International Conference on Information and Knowledge Management (CIKM), Bremen, Germany, October 2005

35. T. Solorio and A. López López, "Learning named entity recognition in Portuguese from Spanish", Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), Lecture Notes in Computer Science, 3406, pp. 762-768, 2005.

36. P. Li, Y. Guan, X.-L. Wang and J. Sun, "Automatic and efficient recognition of proper nouns based on maximum entropy model", Proceedings of the International Conference on Machine Learning and Cybernetics, Vol. 6, pp. 3775- 3780, August 18-21, 2005.

37. Solorio T. "Improvement of Named Entity Tagging by Machine Learning." Ph.D. thesis, National Institute of Astrophysics, Optics and Electronics, Puebla, Mexico, September 2005.

38. T. Solorio and A. López López. “Learning named entity classifiers using support vector machines”, Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), Lecture Notes in Computer Science, 2945, pp. 158-167, 2004.

39. T. Solorio, Improvement of Named Entity Tagging by Machine Learning, Technical Report CCC-04-004, National Institute of Astrophysics, Optics and Electronics, Puebla, Mexico, March 2004.

40. T. Solorio and A. López López. “Adapting a named entity recognition system for Spanish to Portuguese”, Proceedings of the IX Iberoamerican Workshops on Artificial Intelligence: Workshop on Herramientas y Recursos Lingüísticos para el Español y el Portugués, pages 292–297, Puebla, Mexico, November 2004.

41. C. C. Yang and K. W. Li, “Automatic construction of English/Chinese parallel corpora”, Journal of the American Society for Information Science and Technology, v. 54, n. 8, pp. 730 - 742, 2003.

42. J-H Kim, I-H Kang and K-S Choi. “Unsupervised Named Entity Classification Models and their Ensembles”, Proceedings of the 19th International Conference on Computational Linguistics (COLING), 2002.

43. F. Sebastiani “Machine learning in automated text categorization”, ACM Computer Surveys 34(1), pp. 1-47, 2002.

44. AB Jonsdottir, “Named Entity Recognition for Norwegian”, Proceedings of the Seventh ESSLLI Student Session, Malvina Nissim (editor), Chapter 8,, 2002.

45. Jae Ho Kim “Named Entity Classification using Unsupervised Learning Models and Their Ensemble”, Master Thesis, 2001

10. G. Petasis, G. Paliouras, V. Karkaletsis, C.D. Spyropoulos and I. Androutsopoulos, “Resolving Part-of-Speech Ambiguity in the Greek Language Using Learning Techniques”. In Fakotakis, N. et al. (Eds.), Machine Learning in Human Language Technology (Proceedings of the ECCAI Advanced Course on Artificial Intelligence – ACAI ’99), pp. 29 – 34, July 5 – 16, 1999, Chania, Greece. (cited by 7)

46. Fotis Lazarinis, “Automatic Extraction of Knowledge from Greek Web Documents”. In Proceedings of the 6th Dutch-Belgian Information Retrieval Workshop (DIR 2006), TNO ICT, Delft, The Netherlands, March 13-14, 2006.

47. M. Maragoudakis, K. Kermanidis and N. Fakotakis: Towards a Bayesian Stochastic Part-Of-Speech and Case Tagger of Natural Language Corpora , CORPUS LINGUISTICS 2003, Lancaster University (UK), 28 - 31 March, pp 486-495, 2003.

48. Harry Kornilakis, Maria Grigoriadou, Eleni Galiotou, Evangelos Papakitsos, Aligning, Annotating And Lemmatizing A Corpus For The Validation Of Balkan Wordnets, Workshop on Balkan Language Resources and Tools, 2003.

49. Nikos Fakotakis, Kyriakos N. Sgarbas, “Machine Learning in Human Language Technology”. In AC - Advanced Courses, pp. 267-273, 2001.

50. H. Cunningham, D. Maynard, K. Bontcheva, B. Tablan, and Y. Wilks. “Experience of using GATE for NLP R&D”. Workshop: Using Toolsets and Architectures To Build NLP Systems, 18th International Conference on Computational Linguistics (COLING), 2000.

51. D. Maynard, H. Cunningham, et al. A Survey of Uses of GATE. Technical Report CS-00-06, University of Sheffield, UK, 2000.

52. H. Papageorgiou, P. Prokopidis, V. Giouli, S. Piperidis, “A Unified POS Tagging Architecture and its Application to Greek”, Proc. of the 2nd International Conference on Language Resources and Evaluation (LREC), vol. III, pp. 1455–1462, Athens, Greece, June 2000.

11. G. Petasis, G. Paliouras, V. Karkaletsis, C. D. Spyropoulos, I. Androutsopoulos, “Using Machine Learning Techniques for Part-Of-Speech Tagging in the Greek Language”. In “Advances in Informatics: Proceedings of the Pan-Hellenic Conference on Informatics (EPY)”, D.I. Fotiadis and S.D. Nikolopoulos (eds.), pp. 273 – 281, World Scientific, 2000 (Post-Proceedings volume of the 7th Hellenic Conference on Informatics, August 26 – 29 1999, Ioannina, Greece). (cited by 9)

53. F. Lazarinis, "Automatic Extraction of Knowledge from Greek Web Documents", Proceedings of the 6th Dutch-Belgian Information Retrieval Workshop (DIR), TNO ICT, Delft, The Netherlands, March 13-14, 2006.

54. G. Xydas, D. Spiliotopoulos and G. Kouroupetroglou, "Modeling Improved Prosody Generation from High-Level Linguistically Annotated Corpora", IEICE Trans. of Inf. and Syst., Special Section on "Corpus-Based Speech Technologies", vol. E88-D, no 3, pp. 510-518, March 2005.

55. S. D. Baldzis, S. A. Kolalas and E. Eumeridou, "The Computational Modern Greek Morphological Lexicon—An Efficient and Comprehensive System for Morphological Analysis and Synthesis". Literary and Linguistic Computing, vol. 20, Issue 2, pp. 153-187, Oxford University Press, Nov. 2005.

56. H. Kornilakis, M. Grigoriadou, E. Galiotou, E. Papakitsos, "Aligning, Annotating and Lemmatizing a Corpus for the Validation of Balkan WordNets", Proceedings of the Workshop on Balkan Language Resources and Tools, at the Balkan Conference for Informatics (BCI), 2003.

57. M. Maragoudakis, K. Kermanidis and N. Fakotakis, "Towards a Bayesian Stochastic Part-Of-Speech and Case Tagger of Natural Language Corpora", Proceedings of Corpus Linguistics, pp 486-495, 28 - 31 March, Lancaster University, UK, 2003.

58. S. Baldzis, E. Eumeridou and S. Kolalas, "A Complete and Comprehensive System for Modern Greek Language Processing Proposed as a Modern Greek Language Call Method Developer", Literary and Linguistic Computing, vol. 17, Issue 4, pp. 373-400, Oxford University Press, Nov. 2002.

59. H. Cunningham, D. Maynard, K. Bontcheva, B. Tablan, and Y. Wilks, "Experience of using GATE for NLP R&D," Proceedings of the Workshop on Using Toolsets and Architectures to Build NLP Systems, International Conference on Computational Linguistics (COLING), Luxembourg, 2000.

60. H. Papageorgiou, P. Prokopidis, V. Giouli and S. Piperidis, "A Unified POS Tagging Architecture and its Application to Greek," Proceedings of the International Conference on Language Resources and Evaluation (LREC), vol. III, pp. 1455–1462, 2000.

61. D. Maynard, H. Cunningham, et al. A Survey of Uses of GATE, Technical Report CS-00-06, University of Sheffield, UK, 2000.

12. V. Karkaletsis, G. Paliouras, G. Petasis, N. Manousopoulou and C. D. Spyropoulos, "Named-Entity Recognition from Greek and English Texts". Journal of Intelligent and Robotic Systems v. 26, n.2, pp. 123 – 135, 1999. (cited by 9)

62. M Mcshane, "Developing Proper Name Recognition, Translation and Matching Capabilities for Low- and Middle-Density Languages", in book Language Engineering for Lesser-studied Languages, editors S. Nierenburg, IOS Press 2009, pp 81-115

63. G. Lucarelli, X.Vasilakos and I. Androutsopoulos "Named Entity Recognition In Greek Texts With An Ensemble Of SMVS And Active Learning", International Journal on Artificial Intelligence Tools, vol. 16, No 6, pp-1015-1045, 2007

64. B. Bekavac, M. Tadić, (2007), Implementation of Croatian NERC System, Balto-Slavonic Natural Language Processing 2007, ACL 2007, Prague, str. 11-18, [http://acl.ldc.upenn.edu/W/W07/W07-1702.pdf]

65. Georgios Lucarelli and Ion Androutsopoulos, "A Greek Named-Entity Recognizer That Uses Support Vector Machines and Active Learning", in "Lecture Notes in Computer Science", pp 203-213, Volume 3955, Springer, 2006.

66. I. Michailidis, K. Diamantaras, S. Vasileiadis, and Y. Frμere. Greek named entity recognition using Support Vector Machines, Maximum Entropy and Onetime. In Proceedings of the 5th International Conference on Language Resources and Evaluation, pages 45{72, Genova, Italy, 2006.

67. Manolis Maragoudakis§, Katia Kermanidis§, Aristogiannis Garbis‡ and Nikos Fakotakis, "Dealing with Imbalanced Data using Bayesian Techniques", in LREC 2006, pp 1045-1050, 2006

68. Marjorie McShane, Ron Zacharski, Sergei Nirenburg, Stephen Beale, "The Boas II Named Entity Elicitation System", 2005.

69. H. Dalianis and E. Astrom, SweNam - A Swedish Named Entity Recogniser - Its construction, training and evaluation, Technical Report TRITA-NA-P0113, IPLab-189, KTH NADA, University of Stockholm, July 2001.

70. F. Vichot, F. Wolinski, H.-C. Ferri, and D. Urbani, "Feeding a Financial Decision Support System with Textual Information," Journal of Intelligent and Robotic Systems, v. 26, n. 2, pp. 157-166, 1999.

13. V. Karkaletsis, C. D. Spyropoulos, and G. Petasis, "Named Entity Recognition from Greek texts: the GIE Project". In "Advances in Intelligent Systems: Concepts, Tools and Applications", ed. S. Tzafestas, Kluwer Academic Publishers, Part II – Chapter 12, pp. 131 – 142. (Presented at the 3rd European Robotics Intelligent Systems & Control Conference (EURISCON'98), June 22 – 25 1998, Athens, Greece.) (cited by 5)

71. Diana Maynard, "D1.2.2.1.3 Benchmarking of annotation tools", 2007

72. Diana Maynard, Valentin Tablan, Kalina Bontcheva, Hamish Cunningham, Yorick Wilks, "MUSE: a MUlti-Source Entity recognition system", 2003

73. D. Maynard, V. Tablan, C. Ursu, H. Cunningham and Y. Wilks, "Named Entity Recognition from Diverse Text Types". Proc. of the Recent Advances in Natural Language Processing 2001 Conference, Tzigov Chark, Bulgaria.

74. Demiros, S. Boutsis, V. Giouli, M. Liakata, H. Papageorgiou, S. Piperidis, "Named Entity Recognition in Greek Texts", Proc. of the 2nd International Conference on Language Resources and Evaluation (LREC 2000), Athens, Greece, June 2000, vol. III, pp. 1223–1228.

75. S. Boutsis, I. Demiros, V. Giouli, M. Liakata, H. Papageorgiou, S. Piperidis, "A System for Recognition of Named Entities in Greek", In Christodoulakis, D.N. (Ed.), Proceedings of the 2nd International Conference on Natural Language Processing (NLP 2000), Patra, Greece. Lecture Notes in Artificial Intelligence, 1835, Springer, 2000, pp. 424-436.

## 1.6 Αξιοποίηση των αποτελεσμάτων της διατριβής

1. D. Spiliotopoulos, G. Petasis, and G. Kouroupetroglou, "A Framework for Language-Independent Analysis and Prosodic Feature Annotation of Text Corpora". In Proceedings of the 11th International Conference on Text, Speech and Dialogue (TSD 2008), Lecture Notes in Artificial Intelligence 5246, Text Speech and Dialogue, Springer-Verlag Berlin Heidelberg, pp. 517 – 524, Brno, Czech Republic, September 8 – 12, 2008.

2. D. Spiliotopoulos, G. Petasis, and G. Kouroupetroglou, "Prosodically Enriched Text Annotation for High Quality Speech Synthesis". In Proceedings of the 10th International Conference on Speech and Computer (SPECOM-2005), pp. 313 – 316, Patras, Greece, 17 – 19 October, 2005.

3. C. Grover, S. McDonald, D. N. Gearailt, V. Karkaletsis, D. Farmakiotou, G. Samaritakis, G. Petasis, M. T. Pazienza, M. Vindigni, F. Vichot and F. Wolinski, "Multilingual XML-Based Named Entity Recognition for E-Retail Domains". In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands, Spain, May 2002. (cited by 5)

76. Jianhan Zhu, Victoria Uren, and Enrico Motta, "ESpotter: Adaptive Named Entity Recognition for Web Browsing". In Proceedings of the Professional Knowledge Management Conference (WM2005), Kaiserslautern, Germany, 2005.

77. Osenova P. and Kolkovska S., "Combining the named-entity recognition task and NP chunking strategy for robust pre-processing". In Proceedings of the Workshop on Linguistic Theories and Treebanks, 20-21 Sept., Sozopol, Bulgaria, 2002.

78. Labsky M., Vacura M., Praks P., "Web Image Classification for Information Extraction". In Proceedings of the First International Workshop on Representation and Analysis of Web Space (RAWS-05), Prague-Tocna, Sep. 15-16, 2005.

79. G. Sigletos, G. Paliouras, C. D. Spyropoulos, M. Hatzopoulos. "Mining Web sites using wrapper induction, named entities and post-processing". In Proceedings of the 1st European Web Mining Forum Workshop, Joint European Conference on Machine Learning and on Principles and Practices of Knowledge Discovery in Databases (ECML/PKDD), Cavtat-Dubrovnik, Croatia, 2003.

80. Philipp Masche, “Multilingual Information Extraction”. Master's Thesis, University of Helsinki, Faculty of Science, Department of Computer Science, April 2004.

4. D. Farmakiotou, V. Karkaletsis, G. Samaritakis, G. Petasis, and C.D. Spyropoulos “Named Entity Recognition from Greek Web Pages”, Proceedings Companion Volume of 2nd Hellenic Conference on Artificial Intelligence (SETN-02), I.P. Vlahavas and C.D. Spyropoulos (eds), 2002, pp. 91 – 102. (cited by 2)

81. G. Lucarelli, X. Vasilakos and I. Androutsopoulos “Named Entity Recognition In Greek Texts With An Ensemble Of SMVS And Active Learning”, International Journal on Artificial Intelligence Tools, vol. 16, No 6, pp-1015-1045, 2007.

82. Georgios Lucarelli and Ion Androutsopoulos “A Greek Named-Entity Recognizer That Uses Support Vector Machines and Active Learning”, Springer Berlin / Heidelberg, In “Lecture Notes in Computer Science”, Advances in Artificial Intelligence, Volume 3955, pp 203-213, 2006.

5. G. Paliouras, V. Karkaletsis, G. Petasis and C. D. Spyropoulos, “Learning Decision Trees for Named-Entity Recognition and Classification”. In Proceedings of the 14th European Conference on Artificial Intelligence (ECAI 2000), August 20 – 25 2000, Berlin, Germany. (cited by 5)

83. Y. Xia, K-F Wong, and W. Gao, “NIL Is Not Nothing: Recognition of Chinese Network Informal Language Expressions”, Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, Second International Joint Conference on Natural Language Processing (IJCNLP), Jeju Island, Republic of Korea, October 11-13, 2005.

84. Y. Xia and K-F Wong, “Methods and Practice in Chinese Network Informal Language Processing”, Proceedings of the 8th Joint Seminar on Computational Linguistics (JSCL), Nanjing, China, 2005.

85. T. Bogers. Dutch Named Entity Recognition: Optimizing Features, Algorithms, and Output. Master's thesis, Tilburg University, September 2004.

86. H. Isozaki. “Japanese Named Entity Recognition based on a Simple Rule Generator and Decision Tree Learning,” Proceedings of ACL, pp.306--313, 2001.

87. A. Cucchiarelli and P. Velardi, “Unsupervised named entity recognition using syntactic and semantic contextual evidence,” Computational Linguistics, 27 (1), 123-131, 2001.

## 1.7 Άλλες δημοσιεύσεις

1. G. Petasis, V. Karkaletsis, G. Paliouras, A. Krithara, E. Zavitsanos, “Ontology Population and Enrichment: State of the Art”. In “Knowledge-Driven Multimedia Information Extraction and Ontology Evolution”, G. Paliouras, C.D. Spyropoulos, G. Tsatsaronis (Eds.), Lecture Notes in Artificial Intelligence, Vol. 6050, pp. 134 – 166, Springer, ISBN 978-3-642-20794-5, 2011.
2. V. Karkaletsis, P. Fragkou, G. Petasis, E. Iosif, “Ontology Based Information Extraction from Text”. In “Knowledge-Driven Multimedia Information Extraction and Ontology Evolution”, G. Paliouras, C.D. Spyropoulos, G. Tsatsaronis (Eds.), Lecture Notes in Artificial Intelligence, Vol. 6050, pp. 89 – 109, Springer, ISBN 978-3-642-20794-5, 2011.
3. G. Petasis. “TkRibbon: Windows Ribbons for Tk”. In Proceedings of the 17th Annual Tcl/Tk Conference (Tcl 2010), Hilton Suites Chicago/Oakbrook Terrace, 10 Drury Lane, Oakbrook Terrace, Illinois, United States 60181, October 11 – 15, 2010.
4. G. Petasis. “TkGecko: Another Attempt for an HTML Renderer for Tk”. In Proceedings of the 17th Annual Tcl/Tk Conference (Tcl 2010), Hilton Suites Chicago/Oakbrook Terrace, 10 Drury Lane, Oakbrook Terrace, Illinois, United States 60181, October 11 – 15, 2010.

5. G. Petasis. “TileQt and TileGtk: current status”. In Proceedings of the 17th Annual Tcl/Tk Conference (Tcl 2010), Hilton Suites Chicago/Oakbrook Terrace, 10 Drury Lane, Oakbrook Terrace, Illinois, United States 60181, October 11 – 15, 2010.
6. G. Petasis. “Ellogon and the challenge of threads”. In Proceedings of the 17th Annual Tcl/Tk Conference (Tcl 2010), Hilton Suites Chicago/Oakbrook Terrace, 10 Drury Lane, Oakbrook Terrace, Illinois, United States 60181, October 11 – 15, 2010.
7. G. Petasis. “TkDND: a cross-platform drag’n’drop package”. In Proceedings of the 17th Annual Tcl/Tk Conference (Tcl 2010), Hilton Suites Chicago/Oakbrook Terrace, 10 Drury Lane, Oakbrook Terrace, Illinois, United States 60181, October 11 – 15, 2010.
8. G. Petasis and D. Petasis. “BlogBuster: A tool for extracting corpora from the blogosphere”. In Proceedings of the 7th Language Resources and Evaluation Conference (LREC 2010), Valletta, Malta, May 19 – 21, 2010.
9. G. Petasis, V. Karkaletsis, A. Krithara, G. Paliouras and C.D. Spyropoulos. “Semi-automated ontology learning: the BOEMIE approach”. In Proceedings of the International Workshop on Inductive Reasoning and Machine Learning on the Semantic Web (IRMLeS 2009), 6th European Semantic Web Conference (ESWC 2009), Hersonissos, Crete, Greece, May 31, 2009.
10. S. Castano, I.S.E. Peraldi, A. Ferrara, V. Karkaletsis, A. Kaya, R. Möller, S. Montanelli, G. Petasis, and M. Wessel, “Multimedia Interpretation for Dynamic Ontology Evolution”. In Journal of Logic and Computation, September 2008. http://logcom.oxfordjournals.org/cgi/content/abstract/exn049
11. P. Fragkou, G. Petasis, A. Theodorakos, V. Karkaletsis, and C. D. Spyropoulos, “BOEMIE ontology-based text annotation tool”. In Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008), Marrakech (Morocco), May 28 – 30, 2008.
12. G. Petasis, P. Fragkou, A. Theodorakos, V. Karkaletsis, and C. D. Spyropoulos, “Segmenting HTML pages using visual and semantic information”. In Proceedings of the 4th Web as a Corpus Workshop (WAC2008), 6th Language Resources and Evaluation Conference (LREC 2008), pp. 18 – 25, Marrakech (Morocco), June 1, 2008.
    Procedings: The 4th Web as Corpus: Can we do better than Google?
13. Silvana Castano, Sofia Espinosa, Alfio Ferrara, Vangelis Karkaletsis, Atila Kaya, Sylvia Melzer, Ralf Moller , Stefano Montanelli, and Georgios Petasis, “Ontology Dynamics with Multimedia Information: The BOEMIE Evolution Methodology”. In Proceedings of International Workshop on Ontology Dynamics (IWOD) ESWC 2007 Workshop, pp. 41 – 54, Innsbruck, Austria, 7 June 2007. (download – http://kmi.open.ac.uk/events/iwod/)
14. B. Gatos, S. J. Perantonis, V. Maragos, V. Karkaletsis and G. Petasis, “Text Area Identification in Web Images”. In Proceedings of the Panhellenic Conference in Artificial Intelligence (SETN 04), Lecture Notes in Artificial Intelligence, n. 3025, pp. 82 – 92, Springer Verlag, Samos, Greece, May 2004.
15. G. Petasis, V. Karkaletsis, D. Farmakiotou, I. Androutsopoulos, and C.D. Spyropoulos. “A Greek Morphological Lexicon and its Exploitation by Natural Language Processing Applications”. In Lecture Notes on Computer Science (LNCS), vol. 2563, "Advances in Informatics - Post-proceedings of the 8th Panhellenic

Conference in Informatics", Springer Verlag, pp. 401 – 419, 2003. (http://www.springerlink.com/content/hcdjrlvj5nlybf5c/ )

16. G. Petasis, V. Karkaletsis, G. Paliouras and C. D. Spyropoulos, “Using the Ellogon Natural Language Engineering Infrastructure”. In Proceedings of the Workshop on Balkan Language Resources and Tools, 1st Balkan Conference in Informatics (BCI’2003), Thessaloniki, Greece, November 21, 2003. (http://www.iit.demokritos.gr/skel/bci03_workshop/papers/SESSION5_2-23_Petasis.pdf)
17. G. Petasis, V. Karkaletsis, G. Paliouras, I. Androutsopoulos and C. D. Spyropoulos, “Ellogon: A New Text Engineering Platform”. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands, Spain, vol. I, pp. 72 – 78, May 2002.
18. D. Farmakiotou, V. Karkaletsis, I. Koutsias, G. Petasis, C.D. Spyropoulos, “PatEdit: An Information Extraction Pattern Editor for Fast System Customization”. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Spain, May 2002.
19. G. Petasis, V. Karkaletsis, D. Farmakiotou, G. Samaritakis, I. Androutsopoulos, C. D. Spyropoulos, “A Greek Morphological Lexicon and its Exploitation by a Greek Controlled Language Checker”. In Proceedings of the 8th Panhellenic Conference on Informatics, Nicosia, Cyprus, vol. 1, pp. 80 – 89, 8 – 10 November 2001.
20. V. Karkaletsis, G. Samaritakis, G. Petasis, D. Farmakiotou, I. Androutsopoulos, S. Markantonatou, and C.D. Spyropoulos, “A Controlled Language Checker Based on the Ellogon Text Engineering Platform”. In Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001), demonstration notes (companion volume to the proceedings), pp. 74 – 75, Carnegie Mellon University, Pittsburgh, USA, 2001.
21. G. Petasis, “Machine Learning and Named-Entity Recognition”. Presentation in the 8th ELSNET European Summer School on Language and Speech Communication on the subject of Text and Speech Triggered Information Access (TeSTIA 2000), Chios, Greece, July 15 – 30, 2000.
22. G. Petasis, “Exploiting Learning in Bilingual Named Entity Recognition”. In Proceedings of the ECCAI Advanced Course on Artificial Intelligence (ACAI ‘99), July 5 – 16, 1999, Chania, Greece.

## 1.8 Διάρθρωση της διατριβής

Η δομή αυτής της διατριβής είναι η ακόλουθη: Στο κεφάλαιο 2 παρουσιάζονται βασικές έννοιες καθώς και μια σύντομη επισκόπηση της διεθνούς βιβλιογραφίας, που αφορά τις περιοχές της εξαγωγής πληροφορίας και μηχανικής μάθησης. Ταυτόχρονα γίνεται μια αναλυτική παρουσίαση της εργασίας που έχει επιλεχθεί ως πεδίο εφαρμογής της διατριβής αυτής, της εξαγωγής πληροφορίας από κείμενα, εστιάζοντας κυρίως στα στάδια της αναγνώρισης ονομάτων οντοτήτων και της εξαγωγής συσχετίσεων μεταξύ αναγνωρισμένων ονομάτων οντοτήτων. Το κεφάλαιο 3 εξετάζει την ερευνητική περιοχή της αναγνώρισης μερών του λόγου για την Ελληνική γλώσσα, με χρήση τεχνικών μηχανικής μάθησης. Στη βιβλιογραφική επισκόπηση παρουσιάζοντα οι σημαντικότερες μέθοδοι μηχανικής μάθησης που έχουν εφαρμοστεί για την Αγγλική γλώσσα, και δίνει το κίνητρο της επιλογής της μάθησης στηριζόμενης σε *κανόνες μετασχηματισμού καθοδηγούμενη από σφάλματα (transformation-based error-driven learning – TBED)* [15] για να εφαρμοστεί στην Ελληνική γλώσσα.

Το κεφάλαιο 4 εξετάζει την χρήση μηχανικής μάθησης στην εργασία της αναγνώρισης ονομάτων οντοτήτων. Στόχος της ερευνητικής εργασίας που περιγράφεται στο κεφάλαιο αυτό είναι η εφαρμογή υπαρχουσών τεχνικών μηχανικής μάθησης για την δημιουργία συστημάτων αναγνώρισης ονομάτων οντοτήτων. Παρουσιάζονται συνοπτικά οι δύο τεχνικές που εξετάστηκαν, τα δέντρα αποφάσεως και τα νευρωνικά δίκτυα. Ακολούθως, παρουσιάζεται ο τρόπος με τον οποίο κάθε τεχνική χρησιμοποιήθηκε για την κατασκευή ενός αναγνωριστή ονομάτων οντοτήτων. Για κάθε τεχνική γίνεται εκτενής αξιολόγηση καθώς και σύγκριση με υπάρχοντα συστήματα της διεθνούς βιβλιογραφίας. Τέλος, το ίδιο κεφάλαιο εξετάζει και δύο προβλήματα που απασχολούν τους δημιουργούς συστημάτων αναγνώρισης ονομάτων οντοτήτων. Το πρώτο αφορά την απόκτηση των καταλόγων γνωστών ονομάτων οντοτήτων ενώ το δεύτερο, τον προσδιορισμό του πότε ένα σύστημα πρέπει να αναβαθμιστεί ώστε να μπορεί να αντιμετωπίσει τις νέες συνθήκες, όπως έχουν τροποποιηθεί με την πάροδο του χρόνου. Οι *κατάλογοι γνωστών ονομάτων οντοτήτων (gazetteer lists)* χρησιμοποιούνται για τον εντοπισμό *πιθανών* ονομάτων οντοτήτων (π.χ. ονόματα εταιριών) ή μερών ονομάτων οντοτήτων (π.χ. μικρά ονόματα προσώπων οι προσδιοριστές εταιριών, όπως Α.Ε., Ε.Π.Ε, κλπ), κατά τα αρχικά στάδια ενός αναγνωριστή ονομάτων οντοτήτων. Ταυτόχρονα, ο προσδιορισμός της χρονικής στιγμής όπου ένα σύστημα αναγνώρισης ονομάτων οντοτήτων πρέπει να αναβαθμιστεί είναι σημαντικός, αφού το ύφος γραφής των κειμένων ενδέχεται να μεταβάλλεται με την πάροδο του χρόνου, την οποία μεταβολή πρέπει να ακολουθεί ένας αναγνωριστής ονομάτων οντοτήτων για να διατηρήσει την απόδοσή του.

Το κεφάλαιο 5 εξετάζει την ερευνητική περιοχή της επαγωγικής εξαγωγής γραμματικών, και παρουσιάζει έναν νέο αλγόριθμο εξαγωγής γραμματικών ανεξάρτητων από συμφραζόμενα, γνωστό με την ονομασία egGRIDS+. Παρουσιάζεται αναλυτικά ο νέος αλγόριθμος, η αρχιτεκτονική του καθώς και οι τελεστές μάθησης που αυτός χρησιμοποιεί. Ιδιαίτερη βαρύτητα δίνεται στην δυναμική συμπεριφορά των τελεστών αυτών, καθώς και στη βελτιστοποίηση της διαδικασίας αναζήτησης του αλγορίθμου. Τέλος, ο νέος αλγόριθμος αξιολογείται σε τεχνητά παραδείγματα, δημιουργηθέντα από γνωστές γλώσσες που χρησιμοποιούνται για την αξιολόγηση παρόμοιων συστημάτων. Στο κεφάλαιο 6 ο αλγόριθμος egGRIDS+ εφαρμόζεται και αποτιμάται στην εργασία της εξαγωγής σχέσεων μεταξύ ονομάτων οντοτήτων, χρησιμοποιώντας δεδομένα τα οποία έχουν εξαχθεί από πραγματικά σώματα κειμένων. Τέλος, με το κεφάλαιο 7 ολοκληρώνεται η παρούσα διατριβή, παρουσιάζοντας τα συμπεράσματα και προτάσεις για μελλοντική εργασία.

# 2. Βασικές Έννοιες και Γενική Επισκόπηση Βιβλιογραφίας

Στο κεφάλαιο αυτό παρουσιάζονται οι βασικές έννοιες που απαιτούνται για την κατανόηση των μεθόδων και τεχνικών που αναπτύχθηκαν στα πλαίσια της διδακτορικής διατριβής. Συγκεκριμένα, γίνεται αναφορά στις έννοιες που αφορούν την εξαγωγή πληροφορίας, την μηχανική μάθηση και την εκτίμηση της επίδοσης αλγορίθμων μηχανικής μάθησης.

## 2.1 Εξαγωγή πληροφορίας

Ο όρος «εξαγωγή πληροφορίας» αναφέρεται στην αυτόματη εξαγωγή *δομημένης πληροφορίας (structured information)* από αδόμητο κείμενο, κυρίως σε φυσική γλώσσα. Λόγω της γενικότητας του όρου «δομημένη πληροφορία», η εξαγωγή πληροφορίας καλύπτει μια ευρεία ερευνητική περιοχή, από τον απλό εντοπισμό στοιχείων από ιστοσελίδες με χρήση *προτύπων (patterns)* και *κανονικών γραμματικών (regular grammars)*, μέχρι την σημασιολογική ανάλυση της γλώσσας για εξαγωγή νοήματος και εννοιών, όπως οι ερευνητικές περιοχές της *αποσαφήνισης εννοιών λέξεων (word sense disambiguation)* ή της *ανάλυσης συναισθήματος (sentiment analysis)*. Η βασική ιδέα της εξαγωγής πληροφορίας (η συγκέντρωση της βασικής πληροφορίας ενός εγγράφου σε δομημένη μορφή, κυρίως με την μορφή πίνακα) είναι αρκετά παλαιά, με τις πρώτες προσεγγίσεις να εμφανίζονται την δεκαετία του 1950, όπου η εφαρμοσιμότητα της εξαγωγής πληροφορίας προτάθηκε από τον Zellig Harris για υπογλώσσες, με τα πρώτα πρακτικά συστήματα να εμφανίζονται στα τέλη της δεκαετίας του 1970, όπως τα συστήματα του Roger Schank [17], [18], που εξήγαγαν *“scripts”* από ειδησεογραφικά άρθρα εφημερίδων.

Το γεγονός ότι η έξοδος ενός συστήματος εξαγωγής πληροφορίας είναι δομημένη πληροφορία, συχνά σε μορφή πίνακα, ήταν ένα σημαντικό πλεονέκτημα που βοήθησε στην εξέλιξη της ερευνητικής περιοχής. Η ευκολία *αποτίμησης (evaluation)* των συστημάτων εξαγωγής πληροφορίας, έναντι άλλων τεχνολογιών επεξεργασίας φυσικής γλώσσας όπως η μηχανική μετάφραση ή η εξαγωγή περιλήψεων, όπου η αποτίμηση είναι ακόμα ένα ανοικτό ερευνητικό ζήτημα, τα έκανε αρκετά δημοφιλή. Η ευκολία αποτίμησης σε συνδυασμό με τις άμεσες εφαρμογές, οδήγησε σε χρηματοδότηση της ερευνητικής περιοχής καθώς και στα συνέδρια αξιολόγησης Message Understanding Conferences (MUC) [7], [8], [19], τα οποία επαναπροσδιόρισαν την περιοχή.

Η εξαγωγή πληροφορίας δεν πρέπει να συγχέεται με την *ανάκτηση πληροφορίας (information retrieval)* όπου το πρόβλημα είναι ο εντοπισμός και η ανάκτηση σχετικών εγγράφων από ένα σύνολο εγγράφων. Ταυτόχρονα, δεν πρέπει να συγχέεται και με την *κατανόηση κειμένου (text understanding)*, το οποίο είναι ένα περιπλοκότερο πρόβλημα. Η εξαγωγή πληροφορίας τοποθετείται μεταξύ της ανάκτησης πληροφορίας και της κατανόησης κειμένων. Αντίθετα από την ανάκτηση πληροφορίας, όπου στόχος είναι να βρεθούν κείμενα (ή περιοχές τους) σχετικά με ένα θέμα ή ένα ερώτημα, η εξαγωγή πληροφορίας στοχεύει στον εντοπισμό σε κείμενα προ-διευκρινισμένων γεγονότων, εξαρτώμενα από την θεματική περιοχή, (π.χ., στοιχεία ενός αθλητή που συμμετέχει σε μια δοκιμασία άλματος επί κοντώ, όπως το όνομα, η υπηκοότητα, η απόδοση, αλλά επίσης και στοιχεία της δοκιμασίας, όπως το όνομα της διοργάνωσης, ή η τοποθεσία). Αντίθετα με την κατανόηση κειμένου, συχνά ένα μικρό μέρος ενός κειμένου είναι σχετικό με την εξαχθείσα πληροφορία, αντί για ολόκληρο το έγγραφο.

Η εξαγωγή πληροφορίας ορίζεται σαν τον αυτόματο προσδιορισμό καθορισμένων τύπων οντοτήτων, σχέσεων ή γεγονότων σε ελεύθερο κείμενο (μια εισαγωγή στις πιο διαδεδομένες προσεγγίσεις παρουσιάζεται στην εργασία [20]). Πιο συγκεκριμένα, η

εξαγωγή πληροφορίας ασχολείται με την εξαγωγή των ακόλουθων τύπων πληροφορίας:

- *Οντότητες (entities)*: κειμενικές περιοχές ιδιαίτερου ενδιαφέροντος, όπως ονόματα προσώπων, τοποθεσιών, οργανισμών, κλπ., καθώς και αριθμητικές εκφράσεις (π.χ. ημερομηνίες ή άλλες χρονικές εκφράσεις).
- *Αναφορές (mentions)*: ο προσδιορισμός όλων των λεκτικών μορφών/αναφορών *(lexicalisations)* μιας οντότητας στο κείμενο. Παραδείγματος χάριν, το όνομα ενός προσώπου μπορεί να αναφερθεί με διαφορετικούς τρόπους μέσα σε ένα έγγραφο, όπως «Γεώργιος Πετάσης», «Γεώργιος Π. Πετάσης», «Πετάσης», «Γιώργος», αλλά ακόμα και «αυτός».
- *Συσχετίσεις/σχέσεις μεταξύ οντοτήτων (relations among entities)*: ο προσδιορισμός των σχέσεων που υπάρχουν μεταξύ των εξαχθεισών οντοτήτων, σύμφωνα με προϋπάρχουσα προδιαγραφή (γνώση θεματικής περιοχής – domain knowledge). Συνήθως αυτές οι σχέσεις εξάγονται βάσει γλωσσικής πληροφορίας, η οποία υπάρχει μεταξύ των αναφορών των οντοτήτων, και μπορεί να περιγραφεί από *λεξικο-συντακτικά πρότυπα (lexico-syntactic patterns)*.
- *Γεγονότα που περιλαμβάνουν τις οντότητες (events involving the entities)*: ο προσδιορισμός όλων των οντοτήτων που συμμετέχουν σε ένα γεγονός που περιγράφεται σε ένα έγγραφο, καθώς επίσης και ο προσδιορισμός άλλων πιθανά σχετικών γεγονότων με αυτό.

### 2.1.1 Ορισμός και υπο-προβλήματα

Για τον ορισμό της εξαγωγής πληροφορίας προ-απαιτούνται οι ακόλουθες έννοιες και ορισμοί: Έστω ένα έγγραφο $D$, το οποίο τμηματοποιείται σε *λεκτικές μονάδες (tokens)* $\{t_1, \dots, t_n\}$, οι οποίες ορίζονται ως τα τμήματα κειμένου (συνήθως λέξεις) μεταξύ δύο διαδοχικών συμβόλων που έχουν ρόλο διαχωριστικού (π.χ. κενά ή σημεία στίξης). Ως *όριο (boundary)* μιας λεκτικής μονάδας ορίζεται το εικονικό διάστημα μεταξύ δύο γειτονικών λεκτικών μονάδων, δηλαδή τα κενά διαστήματα ή οποιαδήποτε άλλη ακολουθία χαρακτήρων. Έστω ένα *σχεδιότυπο (template)* $T_j$, το οποίο αποτελείται από ένα σύνολο προ-επιλεγμένων *πεδίων (fields)* $\{f_1^j, \dots, f_i^j\}$. Η τιμή ενός πεδίου συμπληρώνεται με τμήματα κειμένου $t_{(s,e)}$ από το έγγραφο, όπου $s$ και $e$ τα όρια αρχής και τέλους του κειμενικού τμήματος. Ένα πεδίο του σχεδιοτύπου ονομάζεται και προς εξαγωγή/συμπλήρωση *θέση-στόχος (target-slot)*, ενώ το τμήμα κειμένου $t_{(s,e)}$ με το οποίο θα συμπληρωθεί, ονομάζεται *δεδομένα πλήρωσης θέσης (slot-filler)*. Για παράδειγμα, ο Πίνακας 1 δείχνει ένα τμήμα μιας ιστοσελίδας η οποία περιγράφει ένα αθλητικό γεγονός, όπου τα προς εξαγωγή τμήματα κειμένου είναι τονισμένα με έντονη γραφή. Ο Πίνακας 2 δείχνει ένα συμπληρωμένο σχεδιότυπο που αντιστοιχεί στην πρώτη οντότητα που περιγράφεται από το τμήμα κειμένου του πίνακα: Πίνακας 1.

| |
|---|
| **Kenya**'s **Richard Limo** the World **5000m** champion (eventual **third** **26:50.20**) came the nearest during the first 300m of the lap, until in the finishing straight, Ethiopia's Olympic bronze Assefa Mezegebu started a drive to the line which took second place (26:49.90). |

**Πίνακας 1: Τμήμα εγγράφου στο οποίο οι λεκτικές μονάδες με τονισμένη γραφή αποτελούν ονόματα οντοτήτων.**

| Πεδίο $f_i$ | $(s, e)$ | $t_{(s,e)}$ |
|---|---|---|
| **Όνομα αθλητή** | (8, 20) | Richard Limo |
| **Εθνικότητα** | (0, 5) | Kenya |
| **Άθλημα** | (31, 36) | 5000m |
| **Κατάταξη** | (56, 61) | third |
| **Επίδοση** | (62, 70) | 26:50.20 |

**Πίνακας 2: Το συμπληρωμένο σχεδιότυπο που πρέπει να παραχθεί από ένα σύστημα εξαγωγής πληροφορίας όταν εφαρμοστεί στο τμήμα εγγράφου του πίνακα: Πίνακας 1.**

Στην γενικότερη μορφή του, το πρόβλημα της εξαγωγής πληροφορίας μπορεί να οριστεί ως εξής: δοθέντος ενός κειμένου $D$, να βρεθούν και να συμπληρωθούν όλα τα δεδομένα πλήρωσης θέσης για κάθε θέση-στόχο $f_i^j$ από το $D$ για κάθε σχεδιότυπο $T_j$. Κατά την διάρκεια των συνεδρίων MUC, το πρόβλημα της εξαγωγής πληροφορίας διαιρέθηκε και τυποποιήθηκε στα ακόλουθα υπο-προβλήματα, τα οποία έχουν ειδικευμένους και διακριτούς στόχους [21]:

1. *Αναγνώριση ονομάτων οντοτήτων (named entity recognition)*: Αναγνώριση και απόδοση κατηγορίας σε οντότητες (ονόματα οντοτήτων και αναφορές). Οι οντότητες εξαρτώνται από την θεματική περιοχή, ενώ συχνά αφορούν πρόσωπα, οργανισμούς, αντικείμενα καθώς και χρηματικές, ημερολογιακές και χρονικές εκφράσεις.
2. *Συν-αναφορά (co-reference)*: Ομαδοποίηση/συσχέτιση αναφορών (*mentions*) που αναφέρονται στην ίδια οντότητα.
3. *Σχεδιότυπο οντότητας (template element)*. Ομαδοποίηση όλων των χαρακτηριστικών μιας οντότητας υπό την μορφή ενός σχεδιοτύπου, το οποίο αναπαριστά ένα πραγματικό αντικείμενο ή γεγονός.
4. *Σχεδιότυπο σχέσης (template relation)*. Αναγνώριση συσχετίσεων μεταξύ σχεδιοτύπων οντοτήτων υπό τη μορφή ενός σχεδιοτύπου σχέσης.
5. *Σχεδιότυπο σεναρίου (scenario template)*. Προσαρμογή των σχεδιοτύπων οντοτήτων και σχέσεων σε ένα συγκεκριμένο *σενάριο γεγονότος (event)*, το οποίο παριστάνεται επίσης με τη μορφή σχεδιοτύπου.

## 2.2 Μηχανική μάθηση

Η *μηχανική μάθηση (machine learning)* αποτελεί έναν από τους σημαντικότερους τομείς έρευνας της τεχνητής νοημοσύνης. Στόχος της είναι η δημιουργία συστημάτων που να είναι σε θέση να εκπαιδεύονται από εμπειρικά δεδομένα που έχουν παρατηρήσει στο παρελθόν, ώστε να εκτελούν την εργασία για την οποία προορίζονται αποτελεσματικότερα. Η διαδικασία εκμάθησης μπορεί να αναλυθεί στα παρακάτω στάδια:

- Απόκτηση εμπειρικών δεδομένων (παραδειγμάτων εκπαίδευσης) από την αλληλεπίδραση με το περιβάλλον.
- Επεξεργασία των δεδομένων, ούτως ώστε να βρεθούν πιθανές γενικεύσεις ή εξειδικεύσεις (διαδικασία μάθησης).
- Χρησιμοποίηση των αποτελεσμάτων της επεξεργασίας για την εκτέλεση της εργασίας στόχου.

### 2.2.1 Ορισμός

*Μάθηση* είναι η διαδικασία εκτίμησης μιας άγνωστης συνάρτησης ή δομής που εμφανίζεται στα δεδομένα εισόδου και εξόδου ενός συστήματος χρησιμοποιώντας έναν περιορισμένο αριθμό παρατηρήσεων (διαθεσίμων δεδομένων που συσχετίζουν είσοδο

με έξοδο ενός συστήματος). Μια *μέθοδος μάθησης* είναι ένας αλγόριθμος (λογισμικό) ο οποίος *εκτιμά (estimates)* την άγνωστη απεικόνιση εξάρτησης *(dependency)* μεταξύ δεδομένων εισόδου και εξόδου ενός συστήματος από τα διαθέσιμα δεδομένα. Μετά την εκτίμηση μιας τέτοιας εξάρτησης *(dependency)*, αυτή μπορεί να χρησιμοποιηθεί για την πρόβλεψη μελλοντικών εξόδων από γνωστές τιμές εισόδου [22]. Σύμφωνα με τον Mitchell [23], «ένα πρόγραμμα υπολογιστή θεωρείται ότι *μαθαίνει* από εμπειρία $E$, σε σχέση με κάποια κατηγορία εργασιών $T$ και μετρική αποτίμησης $P$, εάν η απόδοση στις εργασίες του $T$, όπως μετριέται από το $P$, βελτιώνεται με την εμπειρία $E$».

### 2.2.2 Κατηγορίες μηχανικής μάθησης

Η μηχανική μάθηση μπορεί να διακριθεί στην *επιβλεπόμενη μάθηση (supervised learning)* και στη *μάθηση χωρίς επίβλεψη (unsupervised learning)*. Ένα σύστημα επιβλεπόμενης μάθησης, εκπαιδεύεται αρχικά σε ένα σύνολο παραδειγμάτων εκπαίδευσης όπου κάθε παράδειγμα χαρακτηρίζεται από μια κατηγορία. Τυπικό παράδειγμα επιβλεπόμενης μάθησης αποτελούν τα προβλήματα *ταξινόμησης (classification)*. Σε ένα πρόβλημα ταξινόμησης, κάθε παράδειγμα εκπαίδευσης αντιστοιχεί σε ένα διάνυσμα. Ένα τέτοιο διάνυσμα είναι ένα σύνολο τιμών χαρακτηριστικών, ή αλλιώς γνωρισμάτων, το οποίο περιέχει και μια τιμή κατηγορίας (ή κλάσης – class) η οποία περιγράφει το επιθυμητό αποτέλεσμα, ή αλλιώς, την *έννοια στόχο*. Πληθώρα αλγορίθμων μηχανικής μάθησης είναι σχεδιασμένοι για προβλήματα ταξινόμησης, όπως είναι οι αλγόριθμοι *ID3* [24] και *C4.5* [25] για την εκμάθηση *δέντρων αποφάσεων (decision trees)*, η *μάθηση στηριζόμενη σε κανόνες μετασχηματισμού καθοδηγούμενη από σφάλματα (transformation-based error-driven learning – TBED)* [26], [27], [15], [28] για την εκμάθηση *λιστών αποφάσεων*, ο αλγόριθμος *Naive Bayes* [29], ο αλγόριθμος των *k-κοντινότερων γειτόνων (k-nearest-neighbours)* [30], τα κρυφά μοντέλα Markov [31], κ.α. Το εκπαιδευμένο μοντέλο που προκύπτει από την εφαρμογή ενός αλγορίθμου ταξινόμησης σε ένα σύνολο διανυσμάτων χαρακτηριστικών συχνά καλείται και *ταξινομητής (classifier)*.

Στη μάθηση χωρίς επίβλεψη, δεν υπάρχει προκαθορισμένο σύνολο τιμών. Τα παραδείγματα εκπαίδευσης χωρίζονται σε, άγνωστες εκ των προτέρων, ομάδες με βάση τα χαρακτηριστικά τους, μια διαδικασία που συχνά αναφέρεται σαν *κατηγοριοποίηση (clustering)*. Παραδείγματα αλγορίθμων μη επιβλεπόμενης μάθησης αποτελούν οι αλγόριθμοι *COBWEB* [32], *Apriori* [33], *AutoClass* [34] κ.α. Στόχος της ενότητας αυτής ήταν να παράσχει κάποιες πολύ βασικές έννοιες που αφορούν τη μηχανική μάθηση. Μια περισσότερο λεπτομερής περιγραφή του πεδίου της μηχανικής μάθησης υπάρχει στο βιβλίο [23].

## 2.3 Εκτίμηση επίδοσης αλγορίθμων μηχανικής μάθησης

Στην ενότητα αυτή θα παρουσιαστούν δημοφιλείς μετρικές αποτίμησης/αξιολόγησης που χρησιμοποιούνται για την μέτρηση της επίδοσης πολλών εργασιών στην περιοχή της επεξεργασίας φυσικής γλώσσας.

### 2.3.1 Ακρίβεια (precision)

Η μετρική της *ακρίβειας (precision)* εκτιμά την ορθότητα των αποτελεσμάτων μιας εργασίας. Αν υποθέσουμε ένα σύστημα, στο οποίο όλες οι σωστές απαντήσεις που υπάρχουν είναι $X$, και λαμβάνουμε $Y$ απαντήσεις, εκ των οποίων οι $N$ είναι σωστές, τότε η ακρίβεια ορίζεται σαν:

$$ακρίβεια = \frac{N}{Y}$$

### 2.3.2 Ανάκληση (recall)

Η μετρική της *ανάκλησης (recall)* εκτιμά την *πληρότητα (completeness)* των αποτελεσμάτων μιας εργασίας. Αν υποθέσουμε ένα σύστημα, στο οποίο όλες οι σωστές απαντήσεις που υπάρχουν είναι $X$, και λαμβάνουμε $Y$ απαντήσεις, εκ των οποίων οι $N$ είναι σωστές, τότε η ακρίβεια ορίζεται σαν:

$$ανάκληση = \frac{N}{X}$$

### 2.3.3 F-Measure

Η μετρική F-Measure δίνει μια εκτίμηση της επίδοσης μιας εργασίας, συνδυάζοντας την ακρίβεια και την ανάκληση. Για την ακρίβεια η μετρική F-Measure αποτελεί τον *αρμονικό μέσο όρο (harmonic mean)* της ακρίβειας και της ανάκλησης, ενώ ορίζεται από την ακόλουθη εξίσωση:

$$F - measure = 2 \cdot \frac{ακρίβεια \cdot ανάκληση}{ακρίβεια + ανάκληση}$$

Η μετρική αυτή είναι γνωστή και σαν $F_1$, επειδή προσδίδει την ίδια βαρύτητα στην ακρίβεια και την ανάκληση.

### 2.3.4 Διασταυρωμένη επικύρωση (cross validation)

Η *διασταυρωμένη επικύρωση (cross validation)* είναι ένας τρόπος να λάβουμε μια αξιόπιστη εκτίμηση για την επίδοση ενός συστήματος βασισμένου σε μηχανική μάθηση. Υποθέτοντας ένα σύστημα βασισμένο σε μηχανική μάθηση και ένα σύνολο δεδομένων εκπαίδευσης, το σύστημα καλείται να προσαρμόσει το μοντέλο του όσον το δυνατόν καλύτερα στα δεδομένα εκπαίδευσης. Μια ιδανική διαδικασία αποτίμησης θα περιλάμβανε ένα δεύτερο σύνολο δεδομένων αποτίμησης, το οποίο θα είναι ανεξάρτητο από τα δεδομένα εκπαίδευσης, αλλά να ακολουθεί την ίδια κατανομή με τα δεδομένα εκπαίδευσης (π.χ. κατανομή επιλογής από κάποιο πληθυσμό δεδομένων). Η διασταυρωμένη επικύρωση αποτελεί έναν τρόπο αποτίμησης που προβλέπει την επίδοση ενός συστήματος σε ένα τέτοιο σύνολο δεδομένων αποτίμησης, στην περίπτωση που αυτό δεν είναι διαθέσιμο. Συχνά η διασταυρωμένη επικύρωση θεωρείται ότι δίνει μια καλύτερη εκτίμηση της επίδοσης ενός συστήματος στην πράξη από την απλή αποτίμηση ενός συστήματος σε ένα μικρό μέρος των διαθέσιμων δεδομένων εκπαίδευσης, καθιστώντας την διασταυρωμένη επικύρωση έναν δημοφιλή τρόπο αποτίμησης συστημάτων μηχανικής μάθησης.

Η διαδικασία εφαρμογής της διασταυρωμένης επικύρωσης είναι σχετικά απλή: το σύνολο των δεδομένων εκπαίδευσης χωρίζεται σε $k$ ισομεγέθη (κατά το δυνατόν) τμήματα. Έχοντας $k$ σύνολα δεδομένων εκπαίδευσης, διενεργούνται $k$ πειράματα. Σε κάθε πείραμα, το σύστημα εκπαιδεύεται στην ένωση $k-1$ συνόλων δεδομένων εκπαίδευσης, και αποτιμάται στο εναπομένων σύνολο δεδομένων, το οποίο είναι διαφορετικό σε κάθε πείραμα. Η επίδοση του συστήματος εξάγεται υπολογίζοντας τον μέσο όρο των αποδόσεων που μετρήθηκαν στα $k$ πειράματα. Δημοφιλείς τιμές για το $k$

αποτελούν οι τιμές 5 και 10, οδηγώντας αντίστοιχα στην *πενταπλή και δεκαπλή διασταυρωμένη επικύρωση (5-fold and 10-fold cross validation)*.

# 3. Εξαγωγή Πληροφορίας: Αναγνώριση Μερών του Λόγου

Η διατριβή αυτή εξετάζει την αξιοποίηση μεθόδων μηχανικής μάθησης για τις ανάγκες της εξαγωγής πληροφορίας από μια συνεργατική σκοπιά: τη συνεργασία μηχανικής μάθησης και επεξεργασίας φυσικής γλώσσας για την δημιουργία αξιοποιήσιμων συστημάτων, άμεσα εφαρμόσιμων για την επεξεργασία κειμένων. Η αυξανόμενη διαθεσιμότητα *σωμάτων κειμένων (corpora)*, οφειλόμενη εν πολλοίς στην ανάπτυξη του παγκόσμιου ιστού, και ο συνδυασμός τους με μεθόδους μηχανικής μάθησης, οδηγεί στην ολοένα αυξανόμενη χρήση των συστημάτων επεξεργασίας φυσικής γλώσσας. Το μεγαλύτερο όφελος από την χρήση *εμπειρικών μεθόδων (empiricism)* είναι η πιθανή αντιμετώπιση ενός διαρκούς προβλήματος στο χώρο της επεξεργασίας φυσικής γλώσσας, της *δυσκολίας απόκτησης γλωσσολογικής γνώσης (linguistic knowledge acquisition bottleneck)*. Η δυσκολία συνήθως έγκειται στο ότι οι απαραίτητοι γλωσσικοί πόροι (λεξικά, κανόνες, γραμματικές, κλπ.) πρέπει να δημιουργηθούν από την αρχή ή να προσαρμοστούν για κάθε νέα ή παρεμφερή εφαρμογή.

Η γλωσσική προ-επεξεργασία αποτελεί ένα από τα πρώτα στάδια ενός συστήματος επεξεργασίας φυσική γλώσσας, και συνεπώς και της εξαγωγής πληροφορίας. Η προ-επεξεργασία συνήθως περιλαμβάνει κάποιες βασικές εργασίες, όπως η αναγνώριση λέξεων, προτάσεων, μερών του λόγου των λέξεων, ενώ συχνά περιλαμβάνει και πιο σύνθετη μορφολογική ανάλυση, όπως η εύρεση θεμάτων ή λημμάτων λέξεων. Το κεφάλαιο αυτό εξετάζει το πρόβλημα της αναγνώρισης μερών του λόγου, τόσο από ερευνητική σκοπιά, όσο και από πρακτική, αφού ένας από τους στόχους είναι και η ανάπτυξη ενός αναγνωριστή ονομάτων οντοτήτων για την Ελληνική γλώσσα ο οποίος να μπορεί να εφαρμοστεί στην πράξη. Στις επόμενες ενότητες ορίζεται το πρόβλημα της αναγνώρισης ονομάτων οντοτήτων, παρουσιάζεται η σχετική διεθνή βιβλιογραφία στην οποία συμβάλλει η παρούσα διατριβή, δίδεται η μεθοδολογία που ακολουθήθηκε, και τέλος η προτεινόμενη μεθοδολογία αποτιμάται σε κείμενα εκφρασμένα στην Ελληνική γλώσσα.

Συγκεκριμένα, στην ενότητα 3.1 ορίζεται το πρόβλημα της αναγνώρισης μερών του λόγου, ενώ στις ενότητες 3.2 και 3.3 παρουσιάζεται η βιβλιογραφική επισκόπηση της ερευνητικής περιοχής, μαζί με λεπτομέρειες της συγκεκριμένης μεθόδου που τελικά επιλέχθηκε να εφαρμοστεί στα Ελληνικά. Η προσέγγιση που ακολουθείται παρουσιάζεται στην ενότητα 3.4, ενώ λεπτομέρειες σχετικά με το σύνολο ετικετών, τον κανόνα αρχικοποίησης, τους πρότυπους κανόνες και τα σώματα κειμένων δίνονται στις ενότητες 3.5, 3.6, 3.7, και 3.8 αντίστοιχα. Η πειραματική αξιολόγηση της προτεινόμενης προσέγγισης καθώς και τα συμπεράσματα, παρουσιάζονται στην ενότητα 3.9. Η ενότητα 3.10 ασχολείται με την συνεργασία της προτεινόμενης προσέγγισης, η οποία βασίζεται σε μηχανική μάθηση, με λεκτικούς πόρους όπως ένα ηλεκτρονικό μορφολογικό λεξικό, ενώ η ενότητα 3.11 ολοκληρώνει αυτό το κεφάλαιο, παρουσιάζοντας την συνεισφορά της παρουσιαζόμενης προσέγγισης για την αναγνώριση μερών του λόγου στην ερευνητική περιοχή.

## 3.1 Ορισμός προβλήματος

Ο όρος *«αναγνώριση μερών του λόγου» (part of speech tagging)* αναφέρεται στη διαδικασία απόδοσης μιας μοναδικής *ετικέτας (tag)* σε κάθε λέξη ενός κειμένου, με τέτοιο τρόπο ώστε να μπορεί να προσδιοριστεί το μέρος του λόγου στο οποίο ανήκει η λέξη από την συσχετισμένη με την λέξη ετικέτα. Πιο συγκεκριμένα, με δεδομένα ένα σύνολο ετικετών (κατηγορίες) που αντιστοιχούν στα μέρη του λόγου (άρθρο, επίθετο, ουσιαστικό, ρήμα, κλπ.) και ένα κείμενο, ένα σύστημα αναγνώρισης μερών του λόγου πρέπει να αντιστοιχίσει κάθε λέξη του κειμένου σε μία και μόνο κατηγορία από το σύνολο των διαθέσιμων ετικετών. Η αναγνώριση μερών του λόγου συχνά αποτελεί

μέρος μιας ευρύτερης ανάλυσης, γνωστή με την ονομασία μορφολογική ανάλυση, η οποία απαντάται σε αρκετά συστήματα επεξεργασίας φυσικής γλώσσας. Η ευρύτητα χρήσης των αναγνωριστών μερών του λόγου αποτελεί μια ισχυρή ένδειξη της σημαντικότητας αυτής της ερευνητικής περιοχής, η οποία είναι ταυτόχρονα εξαιρετικά ενδιαφέρουσα ερευνητικά, ιδιαίτερα όταν αφορά γλώσσες με πλούσια μορφολογία όπως η Ελληνική.

Η ευρεία χρήση της αναγνώρισης μερών του λόγου μπορεί εν μέρει να αποδοθεί στο ότι τα μέρη του λόγου προσφέρουν ένα *επίπεδο απόκρυψης (abstraction layer)* για τις ίδιες τις λέξεις. Για παράδειγμα, είναι ευκολότερο να αναπτυχθεί μια γραμματική ευρείας κάλυψης βασισμένη σε μέρη του λόγου, παρά στις ίδιες τις λέξεις. Όμως όλες οι εφαρμογές δεν έχουν τις ίδιες απαιτήσεις όσον αφορά την λεπτομέρεια που προσφέρει αυτό το επίπεδο απόκρυψης. Υπάρχουν εφαρμογές που μπορούν να αρκεστούν στα μέρη του λόγου που ορίζονται στην χρησιμοποιούμενη γλώσσα, αλλά συχνά τα συστήματα επεξεργασίας φυσικής γλώσσας απαιτούν μεγαλύτερη λεπτομέρεια. Η λεπτομέρεια αυτή μπορεί να περιλαμβάνει συντακτικά χαρακτηριστικά της λέξης που μπορούν να εξαχθούν από την μορφολογία της λέξης και το περιβάλλον της, όπως το γένος, τον αριθμό ή ακόμα και την πτώση μιας λέξης. Αυτό έχει άμεσο αντίκτυπο στο σύνολο των ετικετών που καλείται ένας αναγνωριστής μερών του λόγου να χρησιμοποιήσει κατά την διαδικασία της επισημείωσης. Αν και σε μια τέτοια περίπτωση το σύστημα δεν εκτελεί μια απλή αναγνώριση των μερών του λόγου αλλά μια γενικότερη μορφολογική/συντακτική ανάλυση, έχει επικρατήσει να χρησιμοποιείται ο όρος «αναγνώριση μερών του λόγου» και για αυτά τα συστήματα.

## 3.2 Βιβλιογραφική επισκόπηση

Η ιστορία της επεξεργασίας φυσικής γλώσσας είναι συνυφασμένη με την εξέλιξη των ηλεκτρονικών υπολογιστών (Η/Υ) και την εξέλιξη της τεχνητής νοημοσύνης. Από πολύ νωρίς η επεξεργασία κειμένων ήταν ένα επιθυμητό πεδίο εφαρμογής των Η/Υ, έχοντας σαν αποτέλεσμα μια πλούσια βιβλιογραφία η οποία περιλαμβάνει αρκετές δεκαετίες έρευνας. Η αναγνώριση μερών του λόγου είναι μια περιοχή με σημαντικό ερευνητικό ενδιαφέρον, αφού η μορφολογική πληροφορία που αποδίδεται σε κάθε λέξη ενός κειμένου αποτελεί την βάση για την περεταίρω επεξεργασία του. Φυσικά οι πρώτες προσπάθειες αφορούσαν την κλασική ανάπτυξη αναγνωριστών μερών του λόγου, με την μορφή *έμπειρων συστημάτων (expert systems)*. Οι ειδικοί (γλωσσολόγοι) κωδικοποιούσαν τη γλωσσολογική πληροφορία με την μορφή κανόνων και περιορισμών, μέσω των οποίων γινόταν ο χαρακτηρισμός των λέξεων. Παραδείγματα τέτοιων συστημάτων μπορούν να βρεθούν από αρκετές δεκαετίες πριν [35] μέχρι σχετικά πρόσφατα [36]. Ωστόσο, η ανάπτυξη κάθε έμπειρου συστήματος είναι μια χρονοβόρα και ακριβή διαδικασία, ενώ η εφαρμοσιμότητά του είναι περιορισμένη, εξαρτώμενη από την ειδικευμένη γνώση των κανόνων και των περιορισμών, καθώς και την προσέγγιση του ειδικού στο πρόβλημα.

Η αναγνώριση μερών του λόγου ήταν μια από τις πρώτες μορφές επεξεργασίας φυσικής γλώσσας όπου εφαρμόστηκαν τεχνικές μηχανικής μάθησης, η εκπαίδευση των οποίων βασιζόταν στην διαθεσιμότητα κατάλληλα επισημειωμένων σωμάτων κειμένων. Οι πρώτες προσπάθειες αφορούσαν την χρήση *κρυφών μοντέλων Markov (hidden Markov models - HMMs)* [37], [38]. Τα κρυφά μοντέλα Markov υποθέτουν ότι οι πιθανότητες κατηγοριοποίησης μιας λέξης $\nu$ εξαρτώνται μόνο από τις πιθανότητες των προηγούμενων $\nu - 1$ λέξεων. Αν και αυτή η υπόθεση για την συγκεκριμένη εργασία είναι προφανώς λανθασμένη, η εφαρμογή των κρυφών μοντέλων Markov οδήγησε σε αρκετά ακριβή αποτελέσματα στην αναγνώριση μερών του λόγου, η επίδοση των οποίων κυμαίνεται από 95% έως 98% για την Αγγλική γλώσσα. Λόγω της υπόθεσης ότι η κατηγοριοποίηση μιας λέξης εξαρτάται μόνο από τις $\nu - 1$ προηγούμενες λέξεις, οι

αναγνωριστές αυτοί είναι γνωστοί και με την ονομασία *«ν-γραμματικοί αναγνωριστές» (n-gram taggers)*, ενώ μια συνήθης τιμή του $\nu$ είναι 3 *(«τρι-γραμματικοί αναγνωριστές – trigram taggers)*. Η καλή επίδοση των κρυφών μοντέλων Markov οδήγησε την χρήση τους σε μια πληθώρα γλωσσών, συμπεριλαμβανομένης και της Ελληνική γλώσσας [39]. Μια παρεμφερή προσέγγιση αποτελεί η χρήση νευρωνικών δικτύων [40], η οποία αποδίδει συγκρίσιμα ή και λίγο καλύτερα με τα κρυφά μοντέλα Markov. Πιο πρόσφατες μέθοδοι αφορούν την χρήση στατιστικών μεθόδων, κυρίως μεθόδων που βασίζονται στην *μεγιστοποίηση της εντροπίας (maximum entropy)* [41], [42], [43], [44], [45], αλλά και τεχνικές που βασίζονται σε *αποθήκευση στην μνήμη (memory-based learning)* [46].

Ωστόσο όλες οι προαναφερόμενες προσεγγίσεις αντιμετωπίζουν τη γλώσσα σαν ένα μαύρο κουτί γεμάτο πιθανότητες και βάρη μετάβασης, το εκπαιδευμένο μοντέλο των οποίων δεν μπορεί να ερμηνευθεί εύκολα, πόσο μάλλον να εκφραστεί με γλωσσολογικούς όρους, όπως κανόνες ή περιορισμούς πάνω σε γενικεύσεις που έχουν εξαχθεί από την μελέτη της γλώσσας. Αλγόριθμοι μηχανικής μάθησης όπως τα δέντρα ή οι λίστες αποφάσεων μαθαίνουν μοντέλα τα οποία μπορούν να κατανοηθούν ευκολότερα: κατασκευάζοντας έναν αναγνωριστή μερών του λόγου με έναν αλγόριθμο αυτής της κατηγορίας, παρέχεται η δυνατότητα να εξεταστεί κατά πόσο το μοντέλο κωδικοποιεί πραγματικά γλωσσικά φαινόμενα. Οι εκμάθηση κανόνων αποσαφήνισης εφαρμόστηκε από αρκετά νωρίς για την αναγνώριση μερών του λόγου [47], ενώ αρκετά επιτυχημένη όσον αφορά την επίδοση ήταν και η μάθηση στηριζόμενη σε *κανόνες μετασχηματισμού καθοδηγούμενη από σφάλματα (transformation-based error-driven learning – TBED)* [15]. Αρκετά συστήματα βασίστηκαν σε λίστες [48] και δέντρα αποφάσεων [49], [46], [50], [51]. Όλες αυτές οι προσεγγίσεις εμφάνισαν επίδοση συγκρίσιμη με τις στοχαστικές μεθόδους, στην περιοχή από 94% μέχρι 98% για την Αγγλική γλώσσα.

Όσον αφορά την Ελληνική γλώσσα, αρκετές από τις παραπάνω τεχνικές έχουν εφαρμοστεί με αρκετή επιτυχία. Οι Δερματάς και Κοκκινάκης [39] εφάρμοσαν κρυφά μοντέλα Markov, με το σύστημα να πετυχαίνει επίδοση 95% έχοντας εκπαιδευτεί με βάση ένα σώμα κειμένων αποτελούμενο από 110.000 λέξεις. Στην περίπτωση του αναγνωριστή μερών του λόγου που παρουσιάζεται στις εργασίες [50] και [51], εφαρμόζονται δέντρα αποφάσεων σε συνεργασία με ένα μορφολογικό λεξικό, όπου τα δέντρα αποφάσεων καλούνται να άρουν την αμφισημία λέξεων που ανήκουν σε περισσότερα από ένα μέρη του λόγου σύμφωνα με το λεξικό, καθώς και να αναγνωρίσουν το μέρος του λόγου σε λέξεις που δεν περιέχονται στο λεξικό. Η επίδοση του συστήματος κυμαίνεται από 93% έως 95% για την αποσαφήνιση, και από 82% έως 88% για την κατηγοριοποίηση άγνωστων στο λεξικό λέξεων. Τέλος, ο Μαλακασιώτης [52] εφάρμοσε *ενεργητική μάθηση (active learning)* πετυχαίνοντας μια επίδοση της τάξης του 80%.

Η μάθηση στηριζόμενη σε κανόνες μετασχηματισμού καθοδηγούμενη από σφάλματα (TBED) εξετάστηκε στο πλαίσιο της διατριβής, εμφανίζοντας μια επίδοση της τάξης του 95% έχοντας εκπαιδευτεί σε ένα σώμα κειμένων αποτελούμενο από 25.000 λέξεις. Το σύστημα αυτό θα παρουσιαστεί αναλυτικά σε αυτό το κεφάλαιο, καθώς είναι αποτέλεσμα αυτής της διατριβής, ενώ ταυτόχρονα αποτέλεσε το πρώτο σύστημα αναγνώρισης μερών του λόγου για την Ελληνική γλώσσα, το λογισμικό του οποίου διανεμήθηκε ελεύθερα για κάθε χρήση, κάτω από άδεια ανοικτού λογισμικού (μέσω της ανοικτού λογισμικού πλατφόρμας επεξεργασίας φυσικής γλώσσας «Έλλογον» [53]). Επίσης, ένας συνδυασμός αυτής της προσέγγισης με ένα μορφολογικό λεξικό παρουσιάζεται στο πλαίσιο της εργασίας [54]. Μια ανάλογη προσπάθεια περιγράφεται στην εργασία [55] χρησιμοποιώντας επίσης την τεχνική TBED, επιτυγχάνοντας επίδοση 96,28% έχοντας εκπαιδευτεί σε ένα σώμα αποτελούμενο από περίπου 356.000 λέξεις.

## 3.3 Μάθηση στηριζόμενη σε κανόνες μετασχηματισμού καθοδηγούμενη από σφάλματα

Η μάθηση στηριζόμενη σε *κανόνες μετασχηματισμού καθοδηγούμενη από σφάλματα (transformation-based error-driven learning – TBED)* είναι μια γενική μέθοδος μηχανικής μάθησης, η οποία προέκυψε από την διδακτορική διατριβή του Eric Brill [27], όπου και εφαρμόστηκε για πρώτη φορά στην αναγνώριση μερών του λόγου με σημαντική επιτυχία [26], [15], [28]. Η μέθοδος βασίζεται στην εκμάθηση κανόνων μετασχηματισμού οδηγούμενης από μια εξαντλητική αναζήτηση η οποία προσπαθεί να μειώσει το σφάλμα μεταξύ του μοντέλου και των δεδομένων εκπαίδευσης. Ένας αναγνωριστής μερών του λόγου καλείται να αποδώσει μια μοναδική κατηγορία (από ένα προκαθορισμένο σύνολο πιθανών κατηγοριών) σε κάθε λέξη ενός κειμένου. Κατά την διάρκεια της εκπαίδευσης, η τεχνική TBED αποδίδει μια κατηγορία σε κάθε λέξη του κειμένου βάση ενός *κανόνα αρχικοποίησης (start-state rule)*, και στην συνέχεια εξάγεται μια σειρά κανόνων μετασχηματισμού. Κάθε κανόνας μετασχηματισμού τροποποιεί την κατηγορία μιας λέξης, αν ικανοποιηθούν οι περιορισμοί του, ενώ ο χώρος αναζήτησης ορίζεται από ένα μικρό σύνολο από προκαθορισμένους, πρότυπους κανόνες. Τα πρότυπα των κανόνων μπορούν να κατηγοριοποιηθούν σε δύο σημαντικές κατηγορίες:

- *Λεκτικοί κανόνες (lexical rules)*: οι λεκτικοί κανόνες χρησιμοποιούν μορφολογική πληροφορία (όπως το πρόθεμα ή την κατάληξη της λέξης) για να μετασχηματίσουν μια κατηγορία $X$ σε μια διαφορετική κατηγορία $\Psi$ (όπου η κατηγορία $X$ συχνά ανήκει στους περιορισμούς του κανόνα). Οι λεκτικοί κανόνες είναι οι πρώτοι κανόνες που εφαρμόζονται κατά την εφαρμογή του παραγόμενου μοντέλου μετά την ολοκλήρωση της εκπαίδευσης.
- *Κανόνες συμφραζομένων (contextual rules)*: οι κανόνες συμφραζομένων εξετάζουν τις κατηγορίες γειτονικών λέξεων (εντός κάποιου παραθύρου λέξεων $N$), για να μετασχηματίσουν μια κατηγορία $X$ σε μια διαφορετική κατηγορία $\Psi$ (όπου η κατηγορία $X$ συχνά ανήκει στους περιορισμούς του κανόνα). Οι κανόνες συμφραζόμενων εφαρμόζονται μετά τους λεκτικούς κανόνες (ώστε να έχει εξαντληθεί η χρησιμοποίηση της μορφολογικής πληροφορίας για την απόδοση κατηγοριών), κατά την εφαρμογή του παραγόμενου μοντέλου σε κείμενα μετά την ολοκλήρωση της εκπαίδευσης.

Σε γενικές γραμμές η διαδικασία της εφαρμογής ενός μοντέλου σε ένα κείμενο, διενεργείται ως εξής:

1. Μια αρχική κατηγορία αποδίδεται σε κάθε λέξη του κειμένου (με τα σημεία στίξης να θεωρούνται επίσης «λέξεις», στα οποία θα αποδοθεί επίσης μια κατηγορία). Αν η λέξη είναι γνωστή (περιέχεται στο λεξικό που εξάγεται αυτόματα από τα κείμενα εκπαίδευσης), αποδίδεται η πιο συχνή κατηγορία για την λέξη αυτή, βάση συχνοτήτων που έχουν εξαχθεί από το σώμα εκπαίδευσης. Αν η λέξη δεν είναι γνωστή, εφαρμόζεται ένας κανόνας αρχικοποίησης, ο οποίος για την Αγγλική γλώσσα είναι:
   EAN (η λέξη ξεκινά με κεφαλαίο χαρακτήρα) ΤΟΤΕ
   Κατηγοριοποίησε την λέξη σαν κύριο όνομα (ενικού αριθμού)
   ΔΙΑΦΟΡΕΤΙΚΑ
   Κατηγοριοποίησε την λέξη σαν ουσιαστικό (ενικού αριθμού)
2. Μετά την απόδοση μιας αρχικής κατηγορίας σε κάθε λέξη, εφαρμόζεται ένα διατεταγμένο σύνολο (ordered list) από λεκτικούς κανόνες σε κάθε λέξη. Κάθε κανόνας που οι περιορισμοί του ταιριάζουν με την μορφολογία της λέξης (και πιθανώς και με την τρέχουσα κατηγορία της λέξης αν ο κανόνας περιλαμβάνει τέτοιο περιορισμό), εφαρμόζεται, αλλάζοντας την κατηγορία της λέξης με μια νέα. Ένα παράδειγμα λεκτικού κανόνα είναι το ακόλουθο:

EAN (η λέξη τελειώνει σε “ed”) TOTE
Κατηγοριοποίησε την λέξη σαν ρήμα παρελθοντικού χρόνου

3. Μετά την εφαρμογή των λεκτικών κανόνων, ακολουθεί η εφαρμογή των κανόνων συμφραζομένων, οι οποίοι είναι επίσης διατεταγμένοι. Κάθε ένας από αυτούς του κανόνες μπορεί να αλλάξει την κατηγορία μιας λέξης ανάλογα με τις κατηγορίες των γειτονικών λέξεων, του περιβάλλοντος δηλαδή μέσα στο οποίο εμφανίζεται η λέξη. Το παράθυρο που μπορεί να καλυφτεί από τους περιορισμούς ενός κανόνα είναι τέσσερις λέξεις, συμπεριλαμβανομένου και της εξεταζόμενης λέξης, της οποίας η κατηγορία τελικά θα αλλαχθεί από την εφαρμογή του κανόνα.

Όλοι οι απαιτούμενοι πόροι (το λεξικό, οι λεκτικοί κανόνες, οι κανόνες συμφραζομένων, κλπ) δημιουργούνται κατά την φάση της εκπαίδευσης.

## 3.4 Η προτεινόμενη προσέγγιση

Η μάθηση στηριζόμενη σε κανόνες μετασχηματισμού καθοδηγούμενη από σφάλματα (TBED) αποτελεί μια από τις πιο ενδιαφέρουσες τεχνικές για το πρόβλημα της αναγνώρισης μερών του λόγου, αφού συνδυάζει την υψηλή απόδοση για την Αγγλική γλώσσα με ένα μη-στοχαστικό μοντέλο, το οποίο μπορεί εύκολα να αναλυθεί και να κατανοηθεί από έναν ειδικό. Ταυτόχρονα, αλγοριθμικά υπάρχει σαφής διαχωρισμός της αναγνώρισης μερών του λόγου σε δύο στάδια, την αναγνώριση με βάση την μορφολογική ανάλυση και την αναγνώριση με βάση τα συμφραζόμενα, επιτρέποντας την εύκολη και απρόσκοπτη συνεργασία με γλωσσικούς πόρους, όπως μορφολογικά λεξικά. Οι παραπάνω λόγοι αποτέλεσαν τα βασικά κριτήρια επιλογής της μεθόδου αυτής ώστε να εξεταστεί η εφαρμοσιμότητά της στην Ελληνική γλώσσα.

Η εφαρμογή της τεχνικής TBED σε μια νέα γλώσσα (επομένως και στην Ελληνική γλώσσα) απαιτεί μια σειρά από προαπαιτούμενα, στα οποία περιλαμβάνονται τα ακόλουθα:

- Τον καθορισμό ενός *συνόλου ετικετών (tag set)*, για τα μέρη του λόγου που είναι επιθυμητά να αναγνωρίζονται. Το σύνολο ετικετών είναι πιθανό να είναι μεγαλύτερο από το σύνολο των μερών του λόγου, καθώς μπορεί να εμπεριέχεται επιπρόσθετη πληροφορία, όπως αριθμός, γένος, πτώση, κλπ.
- Την δημιουργία ενός κατάλληλου *κανόνα αρχικοποίησης (start-state rule)*, για την απόδοση κατάλληλων ετικετών σε άγνωστες λέξεις.
- Την τροποποίηση των προτύπων των κανόνων, αν αυτό κριθεί απαραίτητο.
- Την δημιουργία του σώματος εκπαίδευσης ικανού μεγέθους, επισημειωμένου με το επιθυμητό σύνολο ετικετών.

Η δημιουργία και τα χαρακτηριστικά των παραπάνω προαπαιτούμενων περιγράφονται στις ακόλουθες ενότητες. Στόχος της διατριβής αυτής είναι, πέρα από την εξέταση της εφαρμοσιμότητας της μεθόδου για την Ελληνική γλώσσα, η δημιουργία ενός πρακτικού αναγνωριστή μερών του λόγου, που να μπορεί να χρησιμοποιηθεί σε ένα ευρύ φάσμα εφαρμογών. Η μέθοδος εξετάστηκε σε δύο διαφορετικές θεματικές περιοχές ώστε να ελεγχθεί τυχόν εξάρτηση της απόδοσης από την θεματική περιοχή (μία περιορισμένη σε εύρος και μια γενικότερη). Τέλος πραγματοποιήθηκαν πειράματα όπου ο εκπαιδευμένος αναγνωριστής μερών του λόγου συνδυάστηκε με ένα εκτενές μορφολογικό λεξικό, πετυχαίνοντας την υψηλότερη απόδοση που έχει σημειωθεί στην διεθνή βιβλιογραφία για την αναγνώριση μερών του λόγου στα Ελληνικά.

## 3.5 Το σύνολο των ετικετών

Η επιλογή του συνόλου των ετικετών είναι ένα ενδιαφέρον ζήτημα κατά την σχεδίαση ενός αναγνωριστή μερών του λόγου, αφού το σύνολο των ετικετών είναι άμεσα συνυφασμένο με την χρηστικότητα ενός αναγνωριστή. Στην πλειοψηφία των

συστημάτων επεξεργασίας φυσικής γλώσσας η αναγνώριση μερών του λόγου είναι ίσως η μοναδική μορφολογική ανάλυση που εκτελείται, περιμένοντας από τον αναγνωριστή επιπρόσθετη πληροφορία, πέρα από το μέρος του λόγου, όπως αριθμό, γένος ή ακόμα και πτώση σε κλιτικές γλώσσες όπως η Ελληνική. Για παράδειγμα ένα σύνολο από 36 ετικέτες έχει χρησιμοποιηθεί για την επισημείωση του σώματος κειμένων Penn Treebank [56] για την Αγγλική γλώσσα, ενώ για κλιτικές γλώσσες όπως η Ελληνική έχουν αναφερθεί σύνολα που πλησιάζουν τις 600 ετικέτες, όπως το σύνολο που προέκυψε από την Ελληνική προσαρμογή του PAROLE [57], με 584 ετικέτες.

Πέρα από την χρηστικότητα, το μέγεθος του συνόλου ετικετών είναι μια σημαντική παράμετρος όταν το σύστημα περιλαμβάνει χρήση μηχανικής μάθησης. Όσο πιο εκτεταμένο είναι το σύνολο ετικετών που καλείται ένας αλγόριθμος μηχανικής μάθησης να μάθει, τόσο περισσότερα παραδείγματα από κάθε ετικέτα πρέπει να περιλαμβάνουν τα δεδομένα εκπαίδευσης ώστε να επιτευχθεί το προσδοκώμενο επίπεδο απόδοσης του εκπαιδευμένου συστήματος. Συχνά πρέπει να βρεθεί ο κατάλληλος συμβιβασμός ανάμεσα στο σύνολο των ετικετών και των διαθέσιμων δεδομένων εκπαίδευσης.

Για τους σκοπούς αυτής της διατριβής, δημιουργήθηκε ένα μικρό σύνολο από ετικέτες, ώστε η εκπαίδευση να μην απαιτεί ένα μεγάλο σώμα επισημειωμένων κειμένων. Το σύνολο των ετικετών ακολουθεί το σύνολο των ετικετών του Penn Treebank [56], το οποίο αποτελεί και το βασικό σύνολο πάνω στο οποίο δοκιμάστηκε η μάθηση στηριζόμενη σε κανόνες μετασχηματισμού καθοδηγούμενη από σφάλματα (TBED) για την Αγγλική γλώσσα. Το σύνολο των ετικετών του Penn Treebank περιέχει 36 ετικέτες, οι οποίες περιλαμβάνουν πληροφορία σχετική με κύρια ονόματα, γένος και αριθμό σε ουσιαστικά, αριθμό σε επίθετα, διάφορους τύπους αντωνυμιών (προσωπικές, κτητικές), και διαχωρισμό σε ενεστωτικό ή παρελθοντικό χρόνο στα ρήματα. Το εκτεταμένο σύνολο ετικετών που δημιουργήθηκε για τα Ελληνικά [58] περιλαμβάνει ένα μεγάλο μέρος των ετικετών του Penn Treebank, στις οποίες έχουν προστεθεί πληροφορίες σχετικά με το γένος σε ουσιαστικά, επίθετα και κύρια ονόματα, αριθμός σε επίθετα, και μελλοντικός χρόνος σε ρήματα. Επίσης εξετάστηκε και η προσθήκη πτώσεων σε επίθετα και ουσιαστικά, ή οποία όμως δεν υιοθετήθηκε, λόγω της σημαντικής αύξησης στον αριθμό των ετικετών που θα προκαλούσε. Ο τελικός αριθμός των ετικετών που δημιουργήθηκαν ανέρχεται στις 58, ενώ όλες οι ετικέτες παρουσιάζονται στο Παράρτημα Ι (Πίνακας 48).

### 3.6 Ο κανόνας αρχικοποίησης

Ο κανόνας αρχικοποίησης χρησιμοποιείται από την τεχνική μηχανικής μάθησης TBED για την απόδοση μιας αρχικής ετικέτας σε μια άγνωστη λέξη. Σαν άγνωστη λέξη θεωρείται κάθε λέξη η οποία δεν υπάρχει στο λεξικό που προκύπτει από την διαδικασία εκπαίδευσης, και το οποίο περιέχει όλες τις λέξεις *(λεκτικές μορφές – word forms)* που περιείχαν τα δεδομένα εκπαίδευσης. Ο αρχικός κανόνας της υλοποίησης της τεχνικής TBED από τον Eric Brill για την Αγγλική γλώσσα, υλοποιούσε τον ακόλουθο κανόνα αρχικοποίησης:

ΕΑΝ (η λέξη ξεκινά με κεφαλαίο χαρακτήρα) ΤΟΤΕ
    Κατηγοριοποίησε την λέξη σαν κύριο όνομα (ενικού αριθμού)
ΔΙΑΦΟΡΕΤΙΚΑ
    Κατηγοριοποίησε την λέξη σαν ουσιαστικό (ενικού αριθμού)

Για την Ελληνική γλώσσα, ο κανόνας αυτός τροποποιήθηκε σε:

ΕΑΝ (η λέξη ξεκινά με έναν Αγγλικό χαρακτήρα) ΤΟΤΕ
    Κατηγοριοποίησε την λέξη σαν «ξένη λέξη» (foreign word – FW)
ΔΙΑΦΟΡΕΤΙΚΑ ΕΑΝ (η λέξη ξεκινά με κεφαλαίο Ελληνικό χαρακτήρα) ΤΟΤΕ
    Κατηγοριοποίησε την λέξη σαν κύριο όνομα (ουσιαστικό), αρσενικού γένους

ΔΙΑΦΟΡΕΤΙΚΑ
Κατηγοριοποίησε την λέξη σαν ουσιαστικό θηλυκού γένους.

Ο κανόνας αυτός προέκυψε από στατιστική ανάλυση των σωμάτων κειμένων που περιγράφονται στην επόμενη ενότητα, ώστε να εξαχθούν οι συχνότερες κατηγορίες με βάση τον πρώτο χαρακτήρα κάθε λέξης.

### 3.7 Τροποποίηση των πρωτοτύπων κανόνων

Η υλοποίηση της τεχνικής TBED από τον Eric Brill περιλαμβάνει ένα μικρό σύνολο από προκαθορισμένους, πρότυπους κανόνες, οι οποίοι καθοδηγούν τη διαδικασία αναζήτησης. Οι πρότυποι αυτοί κανόνες είναι ικανοί να περιγράψουν ένα εκτενές σύνολο από γλωσσικά φαινόμενα, τόσο σε μορφολογικό επίπεδο (πρότυπα λεκτικών κανόνων) όσο και σε συντακτικό επίπεδο (πρότυπα κανόνων συμφραζομένων). Ωστόσο η αλλαγή τους δεν είναι εύκολη στη συγκεκριμένη υλοποίηση, αντίθετα με την παραμετροποίησή τους: είναι εύκολη η αλλαγή του «παραθύρου» των κανόνων: στην περίπτωση των λεκτικών κανόνων, το «παράθυρο» αφορά τον μέγιστο αριθμό χαρακτήρων που χαρακτηρίζονται ως πρόθεμα και κατάληξη σε μια λέξη. Στην περίπτωση των κανόνων συμφραζομένων, το «παράθυρο» αφορά τον αριθμό των λέξεων που θα εξεταστούν στο περιβάλλον μιας λέξης. Τα «παράθυρα» αυτά έχουν σαν προκαθορισμένες τιμές τους τρεις χαρακτήρες για την περίπτωση των λεκτικών κανόνων, και τις δύο λέξεις για την περίπτωση των κανόνων συμφραζομένων. Εξετάστηκε τροποποίηση των παραθύρων αυτών σε τέσσερις χαρακτήρες για τους λεκτικούς κανόνες και τρεις λέξεις για τους κανόνες συμφραζομένων. Ωστόσο η μεταβολή της απόδοσης της μεθόδου δεν ήταν σημαντική για τα Ελληνικά (μεταβολή εντός των ορίων του στατιστικού σφάλματος), σε αντίθεση με την σημαντική αύξηση του χώρου αναζήτησης και συνακόλουθα με την αύξηση του χρόνου εκπαίδευσης. Σαν αποτέλεσμα, οι προκαθορισμένες τιμές παραθύρων χρησιμοποιήθηκαν για όλα τα πειράματα που πραγματοποιήθηκαν στα πλαίσια αυτής της διατριβής.

### 3.8 Τα σώματα κειμένων (δεδομένα εκπαίδευσης)

Λόγω του ενδιαφέροντος που παρουσιάζει η εξάρτηση από την θεματική περιοχή των συστημάτων βασισμένων σε μηχανική μάθηση, δύο εντελώς διαφορετικά σώματα κειμένων δημιουργήθηκαν και επισημειώθηκαν χειρωνακτικά με το σύνολο ετικετών που περιγράφηκε στην προηγούμενη ενότητα. Το πρώτο σώμα κειμένων αφορά μια συγκεκριμένη θεματική περιοχή, η οποία θα μπορούσε να περιγραφεί σαν *«επιτυχή γεγονότα διαδοχής διαχείρισης» (succession management events)*, και περιείχε άρθρα ειδήσεων από την Ελληνική εφημερίδα «Διαφημιστική Εβδομάδα»[1] [59]. Το σώμα κειμένων περιέχει κείμενα για προσωπικό το οποίο μετακινείται ή εγκαταλείπει μια εταιρία, καθώς και συγχωνεύσεις/διασπάσεις εταιρειών για την περίοδο από τον Ιανουάριο 1996 έως και τον Δεκέμβριο 1998. Το μέγεθος του σώματος κειμένου είναι περίπου 65.000 λέξεις. Μέρος του σώματος αυτού (περίπου 36.000 λέξεις) επισημειώθηκε χειρωνακτικά με τις ετικέτες για τα μέρη του λόγου. Μια ιδιαιτερότητα του σώματος αυτού είναι ότι περιέχει έναν σημαντικό αριθμό ξένων (κυρίως Αγγλικών) λέξεων, που συχνά αναφέρονται σε ονόματα προσώπων, εταιριών ή θέσεων εργασίας.

Το δεύτερο σώμα κειμένων αφορούσε μια γενικότερη θεματική περιοχή, καθώς αφορά κείμενα που ανήκουν σε διάφορες θεματικές περιοχές. Αυτό το σώμα κειμένων παραχωρήθηκε από το Εργαστήριο Ενσύρματης Τηλεπικοινωνίας, του Τομέα Τηλεπικοινωνιών, του Τμήματος Ηλεκτρολόγων Μηχανικών και Τεχνολογίας

[1] Διαφημιστική Εβδομάδα: http://www.adweek.gr

Υπολογιστών, της Πολυτεχνικής Σχολής του Πανεπιστημίου Πατρών. Το μέγεθος αυτού του σώματος κειμένων είναι περίπου 125.000 λέξεις, ενώ ήταν χειρωνακτικά επισημειωμένο με πληροφορία σχετική με μέρη του λόγου, η οποία προσαρμόστηκε στο σύνολο ετικετών που περιγράφηκε στην ενότητα 3.5.

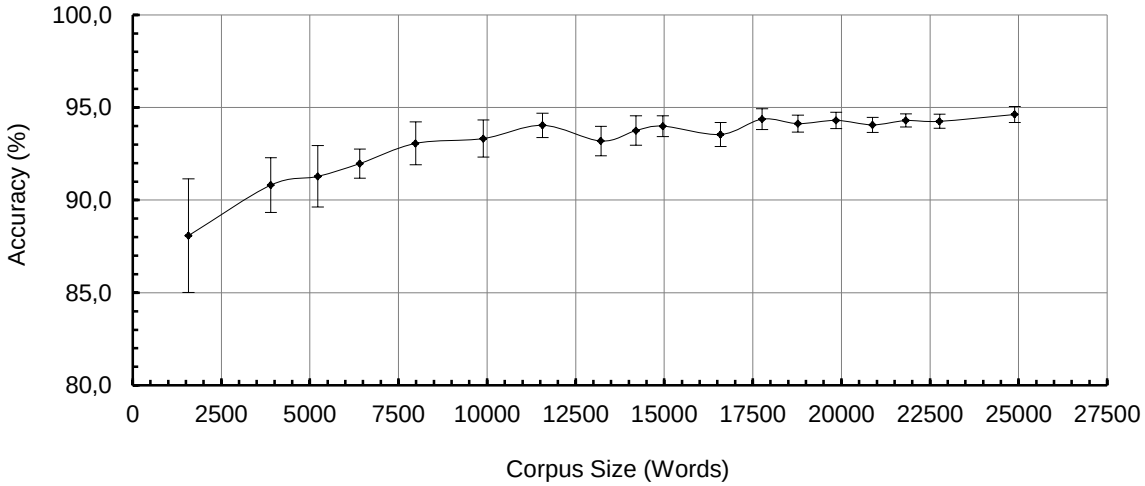
## 3.9 Πειραματική αξιολόγηση και αποτελέσματα

Για την αξιολόγηση της μεθόδου TBED στα Ελληνικά, διεξήχθησαν δύο πειράματα, χρησιμοποιώντας τα δύο σώματα κείμενων που δημιουργήθηκαν για αυτό το σκοπό. Σε κάθε πείραμα, η απόδοση του αναγνωριστή μετράται σε ένα κομμάτι του σώματος κειμένου το οποίο δεν έχει χρησιμοποιηθεί σαν δεδομένα εκπαίδευσης. Για την αποκόμιση μιας ρεαλιστικής και αμερόληπτης εκτίμησης των επιδόσεων της μεθόδου, χρησιμοποιήθηκε *δεκαπλή διασταυρωμένη επικύρωση (10-fold cross validation)* (ενότητα 2.3). Σύμφωνα με αυτή τη μέθοδο αξιολόγησης, το σώμα κειμένων χωρίζεται σε δέκα, ίσου μεγέθους υποσώματα, με το τελικό αποτέλεσμα να είναι ο μέσος όρος της επίδοσης σε δέκα πειράματα. Σε κάθε πείραμα, εννέα από τα δέκα επιμέρους υποσώματα κειμένων χρησιμοποιούνται για την εκπαίδευση του αναγνωριστή μερών του λόγου, ενώ το δέκατο χρησιμοποιείται για την αποτίμηση της απόδοσης (αξιολόγηση).

### 3.9.1 Αναγνώριση μερών του λόγου (γενική θεματική περιοχή)

Στο πρώτο πείραμα αξιολόγησης χρησιμοποιήθηκε το σώμα κειμένων που εκπροσωπούσε την γενικότερη θεματική περιοχή, το οποίο παραχωρήθηκε από το Εργαστήριο Ενσύρματης Τηλεπικοινωνίας του Πανεπιστημίου Πατρών. Το σώμα αυτό ήταν οργανωμένο σαν ένα ενιαίο αρχείο 125.000 λέξεων, όπου κάθε γραμμή αντιστοιχεί σε μία μόνο πρόταση. Κάθε λέξη αυτού του αρχείου έχει επισημειωθεί χρησιμοποιώντας ένα εξαιρετικά πλούσιο και περιγραφικό σύνολο ετικετών, το οποίο περιείχε περισσότερες ετικέτες από τις 58 που παρουσιάζονται στο ΠΑΡΑΡΤΗΜΑ Ι (Πίνακας 48). Από το αρχικό κείμενο, μια νέα έκδοση δημιουργήθηκε, όπου κάθε αρχική ετικέτα έχει αντιστοιχηθεί χειρωνακτικά σε μια ετικέτα από το σύνολο των 58 ετικετών που περιγράφηκε στην ενότητα 3.5. Στη συνέχεια, οι προτάσεις του νέου κειμένου ανακατεύτηκαν με τυχαία σειρά. Ο λόγος για την ανάγκη ανακατέματος ήταν η δομή του κειμένου, το οποίο αποτελείται από μικρές ομάδες προτάσεων από μια θεματική περιοχή: ανακατεύοντας τις προτάσεις μετακινούνται προτάσεις από μια θεματική περιοχή σε διαφορετικές περιοχές του κειμένου.

Η αποτίμηση έγινε με χρήση δεκαπλής διασταυρωμένης επικύρωσης, για διαφορετικά μεγέθη δεδομένων εκπαίδευσης. Τα αποτελέσματα παρουσιάζονται στην Εικόνα 1.



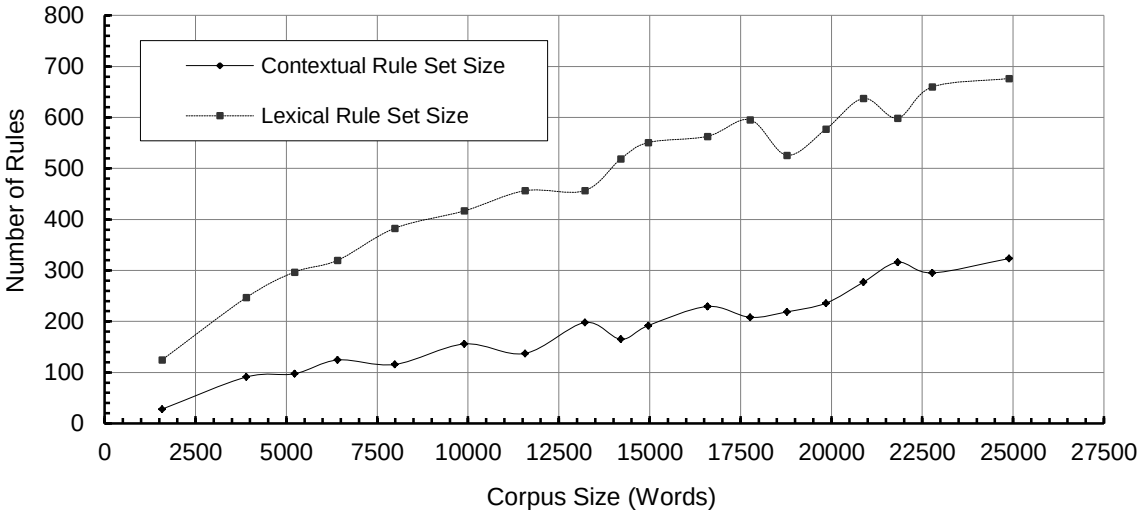

**Εικόνα 1: Η ορθότητα (accuracy) του αναγνωριστή μερών του λόγου TBED, σε συνάρτηση με το μέγεθος των δεδομένων εκπαίδευσης.**

Στο παραπάνω γράφημα ο οριζόντιος άξονας αναπαριστά το μέγεθος των δεδομένων εκπαίδευσης (εκφρασμένο σε αριθμό λέξεων) που χρησιμοποιήθηκε κατά την διαδικασία της εκπαίδευσης, ενώ ο κατακόρυφος άξονας αναπαριστά την *ορθότητα (accuracy)* της αποτίμησης στο σώμα κειμένων ελέγχου (του οποίου το μέγεθος είναι το 1/10 των δεδομένων εκπαίδευσης). Οι γραμμές σφάλματος αντιστοιχούν στην *τυπική απόκλιση (standard deviation)* της ορθότητας, η οποία έχει υπολογιστεί σαν η μέση τιμή από δέκα πειράματα. Όπως είναι αναμενόμενο, η ακρίβεια του αναγνωριστή μερών του λόγου TBED αυξάνει όσο αυξάνονται και τα δεδομένα εκπαίδευσης, με την ορθότητα να σταθεροποιείται γύρω στο 95%, όταν τα δεδομένα εκπαίδευσης ξεπερνούν τις 18.000 λέξεις.

Εκτός από την ορθότητα, εξετάστηκε επίσης και το μέγεθος του παραγόμενου μοντέλου, Ένα άλλο σημείο ενδιαφέροντος, είναι η εξέταση του αριθμού των κανόνων που μαθεύτηκαν. Συνήθως, η υψηλή απόδοση μιας μαθησιακής εργασίας όταν συνδυάζεται με ένα μικρό μέγεθος μοντέλου, αποτελεί ένδειξη της ευρωστίας της μαθησιακής διαδικασίας. Αντίθετα, μεγάλο μέγεθος μοντέλου ενδέχεται να σηματοδοτεί κάποιο πρόβλημα στη μαθησιακή διαδικασία, που ο αλγόριθμος προσπαθεί να ξεπεράσει μέσω *απομνημόνευσης (over-fitting)* των δεδομένων εκπαίδευσης. Με άλλα λόγια, ένας μεγάλος αριθμός κανόνων είναι συνήθως μια ένδειξη ότι ο αλγόριθμος μηχανικής μάθησης προσπαθεί να «απομνημονεύσει» τα δεδομένα εκπαίδευσης, αντί να ανακαλύψει τις συσχετίσεις που διέπουν τα δεδομένα. Καθώς η μέθοδος TBED περιλαμβάνει δύο διακριτά στάδια εκπαίδευσης (εξαγωγή λεκτικών κανόνων και κανόνων συμφραζομένων), τα δύο σύνολα των κανόνων μπορούν να εξεταστούν χωριστά.

Ο αριθμός των παραγόμενων κανόνων σε συνάρτηση με το μέγεθος των δεδομένων εκπαίδευσης εμφανίζεται στην Εικόνα 2. Όπως φαίνεται και από το διάγραμμα, ο αριθμός και των δύο τύπων κανόνων (λεκτικών και κανόνων συμφραζομένων) αυξάνει σχεδόν γραμμικά με το μέγεθος των δεδομένων εκπαίδευσης, με τον αριθμό των λεκτικών κανόνων να είναι σημαντικά μεγαλύτερος από τον αριθμό των κανόνων συμφραζομένων. Αυτός ο μεγάλος αριθμός των λεκτικών κανόνων υποδεικνύει ότι τα πρότυπα των λεκτικών κανόνων (από τα οποία προκύπτουν οι λεκτικοί κανόνες κατά την διαδικασία της εκπαίδευσης) αντιμετωπίζουν μια σχετική δυσκολία, προκειμένου να καλύψουν την μορφολογία της Ελληνικής γλώσσας, με δεδομένο πάντα τον κανόνα

αρχικοποίησης που χρησιμοποιήθηκε. Φαίνεται ότι είναι ευκολότερο να αντλήσει κανείς πληροφορία με βάση γραμματικές και συντακτικές ιδιότητες (κανόνες συμφραζομένων), παρά από την μορφολογία (λεκτικοί κανόνες) στα πλαίσια κλιτικών γλωσσών, όπως τα Ελληνικά.

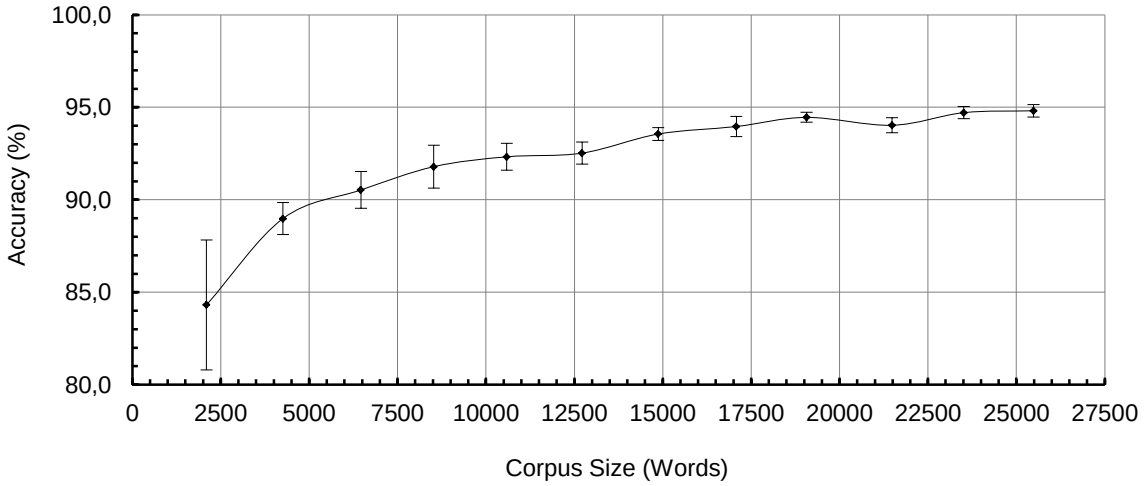



**Εικόνα 2: Ο αριθμός των λεκτικών και συμφραζόμενων κανόνων, σε συνάρτηση με το μέγεθος των δεδομένων εκπαίδευσης.**

Τα παραπάνω πειράματα δείχνουν ότι η ακρίβεια της τεχνικής TBED για την Ελληνική γλώσσα κυμαίνεται περίπου στο 95% όταν εφαρμόζεται σε κείμενα, χωρίς να εξαρτάται ιδιαίτερα από την θεματική περιοχή των κειμένων. Το μέγεθος του παραγόμενου μοντέλου (εκφρασμένο σε αριθμό κανόνων) αυξάνεται σχεδόν γραμμικά με το μέγεθος των δεδομένων εκπαίδευσης, με τον αριθμό των λεκτικών κανόνων να βρίσκεται πάντα σε υψηλότερο επίπεδο από τον αντίστοιχο αριθμό των κανόνων συμφραζομένων.

### 3.9.2 Αναγνώριση μερών του λόγου (συγκεκριμένη θεματική περιοχή)

Στο δεύτερο πείραμα αξιολόγησης χρησιμοποιήθηκε το σώμα κειμένων που αφορά μια συγκεκριμένη θεματική περιοχή, η οποία θα μπορούσε να περιγραφεί σαν *«επιτυχή γεγονότα διαδοχής διαχείρισης» (management succession events)*, και περιείχε άρθρα ειδήσεων από την Ελληνική εφημερίδα «Διαφημιστική Εβδομάδα». Το αρχικό μέγεθος του σώματος είναι περίπου 65.000 λέξεις. Μέρος του σώματος κειμένου (περίπου 36.000 λέξεις) επισημειώθηκαν χειρωνακτικά με το σύνολο των ετικετών που περιγράφηκε στην ενότητα 3.5. Ακολουθήθηκε η ίδια διαδικασία αξιολόγησης όπως και στην περίπτωση του προηγούμενου πειράματος, χρησιμοποιώντας δεκαπλή διασταυρωμένη επικύρωση, και χρησιμοποιώντας διαφορετικά μεγέθη δεδομένων εκπαίδευσης. Τα αποτελέσματα παρουσιάζονται στην Εικόνα 3.



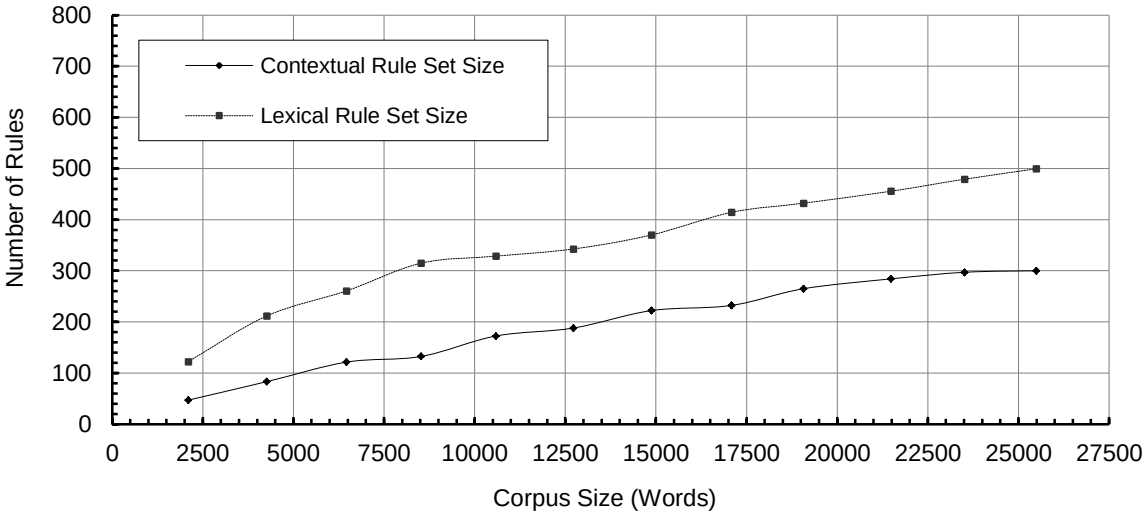

**Εικόνα 3: Η ορθότητα (accuracy) του αναγνωριστή μερών του λόγου TBED, σε συνάρτηση με το μέγεθος των δεδομένων εκπαίδευσης. (Θεματική περιοχή «γεγονότων διαδοχής διαχείρισης»)**

Στο παραπάνω γράφημα ο οριζόντιος άξονας αναπαριστά το μέγεθος των δεδομένων εκπαίδευσης (εκφρασμένο σε αριθμό λέξεων) που χρησιμοποιήθηκε κατά την διαδικασία της εκπαίδευσης, ενώ ο κατακόρυφος άξονας αναπαριστά την *ορθότητα (accuracy)* της αποτίμησης στο σώμα κειμένων ελέγχου (του οποίου το μέγεθος είναι το 1/10 των δεδομένων εκπαίδευσης). Οι γραμμές σφάλματος αντιστοιχούν στην *τυπική απόκλιση (standard deviation)* της ορθότητας, η οποία έχει υπολογιστεί σαν η μέση τιμή από δέκα πειράματα. Η απόδοση του αναγνωριστή μερών του λόγου δεν εμφανίζεται σημαντικά διαφοροποιημένη από το προηγούμενο πείραμα: η ακρίβεια του αναγνωριστή μερών του λόγου TBED αυξάνει όσο αυξάνονται και τα δεδομένα εκπαίδευσης, με την ορθότητα να σταθεροποιείται γύρω στο 95%, όταν τα δεδομένα εκπαίδευσης ξεπερνούν τις 18.000 λέξεις. Μικρές διαφοροποιήσεις υπάρχουν, αλλά βρίσκονται εντός των ορίων του στατιστικού σφάλματος. Το γεγονός ότι σε αυτό το πείραμα η θεματική περιοχή ήταν σαφώς πιο περιορισμένη, και συνεπώς η χρήση της γλώσσας στα κείμενα σαφώς πιο «ελεγχόμενη», δεν φάνηκε να επηρεάζει την απόδοση του αναγνωριστή, ούτε αρνητικά, ούτε θετικά, οδηγώντας στο συμπέρασμα ότι η απόδοση δεν εξαρτάται από την θεματική περιοχή.

Ο αριθμός των παραγόμενων κανόνων σε συνάρτηση με το μέγεθος των δεδομένων εκπαίδευσης εμφανίζεται στην Εικόνα 4. Όπως φαίνεται και από το διάγραμμα, ο αριθμός και των δύο τύπων κανόνων (λεκτικών και κανόνων συμφραζομένων) αυξάνει σχεδόν γραμμικά με το μέγεθος των δεδομένων εκπαίδευσης, όπως συνέβη και με το σώμα κειμένων της ευρύτερης θεματικής περιοχής. Επιπρόσθετα, ο αριθμός των λεκτικών κανόνων είναι σχεδόν ο διπλάσιος σε σχέση με τους κανόνες συμφραζομένων. Ο αριθμός των κανόνων συμφραζομένων εμφανίζεται στα ίδια επίπεδα με τους αντίστοιχους κανόνες του προηγούμενου πειράματος. Ωστόσο, ο αριθμός των λεκτικών κανόνων εμφανίζεται σημαντικά μειωμένος, γεγονός που μπορεί να αποδοθεί στην πιο ελεγχόμενη χρήση της γλώσσας (λιγότερες λεκτικές μορφές) και στην παρουσία λέξεων με λατινικούς χαρακτήρες, οι οποίες είναι εύκολο να αποσαφηνιστούν καθώς επισημειώνονται με την ίδια ετικέτα.

Τα παραπάνω πειράματα δείχνουν ότι η ακρίβεια του αναγνωριστή μερών του λόγου TBED για την Ελληνική γλώσσα είναι περίπου 95%, με ελάχιστη έως μηδαμινή εξάρτηση από την θεματική περιοχή. Σκοπός του πειράματος που περιγράφεται σε αυτή την ενότητα ήταν η εξέταση του κατά πόσο η απόδοση της προσέγγισης θα ήταν σε

υψηλότερα επίπεδα σε μια συγκεκριμένη θεματική περιοχή σε σχέση με μια γενικότερη θεματική περιοχή. Ωστόσο η υπόθεση αυτή δεν επιβεβαιώθηκε πειραματικά.




**Εικόνα 4: Ο αριθμός των λεκτικών και συμφραζόμενων κανόνων, σε συνάρτηση με το μέγεθος των δεδομένων εκπαίδευσης. (Θεματική περιοχή «γεγονότων διαδοχής διαχείρισης»)**

## 3.10 Συνδυασμός μηχανικής μάθησης και μορφολογικού λεξικού

Ένα σημαντικό ερώτημα που προκύπτει ωστόσο, είναι το γιατί δε χρησιμοποιείται ένα ηλεκτρονικό μορφολογικό λεξικό της εκάστοτε γλώσσας, για τη μορφολογική ανάλυση των λέξεων. Η απάντηση είναι απλή και έγκειται στους περιορισμούς που υπεισέρχονται σε μία τέτοια λύση. Καταρχάς, η κατασκευή ενός ηλεκτρονικού λεξικού είναι μία χρονοβόρα, επίπονη και ιδιαίτερα ακριβή διαδικασία, με αποτέλεσμα η διαθεσιμότητα μορφολογικών λεξικών να είναι εξαιρετικά περιορισμένη, και εξαιρουμένης της Αγγλικής γλώσσας, να μην είναι ελεύθερα διαθέσιμα. Επιπρόσθετα, ένα λεξικό περιέχοντας πεπερασμένο πλήθος λέξεων, είναι αδύνατο να καλύψει όλες τις πιθανές λέξεις που μπορεί να εμφανιστούν: ακόμα και αν περιέχει ένα μεγάλο αριθμό λέξεων είναι δύσκολο να καλύψει πλήρως κατηγορίες όπως τα κύρια ονόματα ή τους νέους τεχνικούς όρους. Τέλος, ένα σύστημα που συμβουλεύεται απλά ένα λεξικό χωρίς να λαμβάνει υπόψη του τα συμφραζόμενα της κάθε λέξης, σε πολλές περιπτώσεις δε μπορεί να αποφασίσει για την ετικέτα που πρέπει να αποδοθεί σε μία λέξη, λόγω αμφισημίας των λεκτικών μορφών, κυρίως σε πλούσιες μορφολογικά γλώσσες όπως η Ελληνική.

Η αρχιτεκτονική της μεθόδου TBED καθιστά την συνεργασία της μεθόδου με ένα μορφολογικό λεξικό εξαιρετικά εύκολη. Η μέθοδος TBED χρησιμοποιεί έναν κανόνα αρχικοποίησης (ενότητα 3.6) για να αποδώσει μια αρχική ετικέτα σε κάθε λέξη του κειμένου. Ο κανόνας αυτός μπορεί εύκολα να τροποποιηθεί ώστε να συμβουλεύεται ένα μορφολογικό λεξικό: αν η εκάστοτε λέξη περιέχεται στο μορφολογικό λεξικό, η αρχικοποίηση μπορεί να γίνεται με μια ετικέτα που προέρχεται από το μορφολογικό λεξικό (επιλέγοντας κάποια από τις διαθέσιμες ετικέτες του λεξικού σε περίπτωση αμφισημίας), ενώ στην αντίθετη περίπτωση, όπου η λέξη δεν περιέχεται στο μορφολογικό λεξικό, μπορεί να εφαρμόζεται σε αυτήν ο αρχικός κανόνας αρχικοποίησης, αποδίδοντας στην άγνωστη στο λεξικό λέξη την ετικέτα που θα έπαιρνε αν δεν υπήρχε καθόλου το λεξικό.

### 3.10.1 Πειραματική αξιολόγηση και αποτελέσματα

Για τις ανάγκες της πειραματικής αξιολόγησης, η μέθοδος TBED συνδυάστηκε με ένα μορφολογικό λεξικό για την Ελληνική γλώσσα ( [60], [61]), το οποίο περιέχει περίπου 60.000 λήμματα, τα οποία αντιστοιχούν σε περίπου 710.000 λεκτικές μορφές. Το μορφολογικό λεξικό συμπληρώνει τον κανόνα αρχικοποίησης, ο οποίος χρησιμοποιεί το μέρος του λόγου που επιστρέφει το μορφολογικό λεξικό (επιλέγοντας την πρώτη απάντηση του λεξικού σε περίπτωση αμφισημίας), και την ετικέτα που προκύπτει από τον κανόνα αρχικοποίησης της ενότητας 3.6, εάν η λέξη δεν υπάρχει στο λεξικό.

Σαν σώμα κειμένων αξιολόγησης χρησιμοποιήθηκε ένα σώμα κειμένων που παραχωρήθηκε από το ερευνητικό έργο M-PIRO [62] του προγράμματος IST της Ευρωπαϊκής Ένωσης, ενώ η θεματική περιοχή του σώματος κειμένου ήταν περιγραφές εκθεμάτων αρχαιολογικών μουσείων. Τα κείμενα προέρχονται από ένα σύστημα παραγωγής φυσικής γλώσσας για την Ελληνική γλώσσα, ενώ επισημειώθηκαν χειρωνακτικά με πληροφορία σχετική με τα μέρη του λόγου. Το σώμα κειμένων αξιολόγησης αποτελείται από 441 προτάσεις, οι οποίες περιέχουν 5635 λέξεις (και 643 σημεία στίξης). Αξίζει να σημειωθεί ότι η θεματική περιοχή είναι εντελώς διαφορετική από τις θεματικές περιοχές στις οποίες εκπαιδεύτηκε ο αλγόριθμος TBED. Τα αποτελέσματα της εφαρμογής της μεθόδου TBED, τόσο σε ανεξάρτητη εφαρμογή όσο και σε συνδυασμό με το μορφολογικό λεξικό, εμφανίζονται στον ακόλουθο πίνακα (Πίνακας 3):

| | TBED | Λεξικό | TBED + Λεξικό |
|---|---|---|---|
| Ακρίβεια | 94.03 % | 91.60 % | 97.90 % |
| Ανάκληση | 94.03 % | 91.60 % | 97.90 % |
| F-Measure | 94.03 % | 91.60 % | 97.90 % |

**Πίνακας 3: Αποτελέσματα του υβριδικού αναγνωριστή μερών του λόγου, για την θεματική περιοχή των περιγραφών εκθεμάτων μουσείων.**

## 3.11 Συνεισφορά

Η παρούσα διατριβή προσάρμοσε και εξέτασε την συμπεριφορά της μάθησης στηριζόμενη σε *κανόνες μετασχηματισμού καθοδηγούμενη από σφάλματα (transformation-based error-driven learning – TBED)* για την εργασία της αναγνώρισης μερών του λόγου, για την Ελληνική γλώσσα. Η προσέγγιση που περιγράφεται στο κεφάλαιο αυτό αποτέλεσε την πρώτη εφαρμογή της μεθόδου στην Ελληνική γλώσσα, ενώ η απόδοση της προσέγγισης βρίσκεται στην κορυφή της απόδοσης (μαζί με την προσέγγιση [39]) για την Ελληνική γλώσσα, έχοντας το πλεονέκτημα της μικρότερης απαίτησης στο μέγεθος των δεδομένων εκπαίδευσης. Ενώ η μέθοδος των Δερματά και Κοκκινάκη προσεγγίζει το 95% εκπαιδευόμενη σε ένα σώμα κειμένων 110.000 λέξεων, η προσέγγιση που παρουσιάζεται στο κεφάλαιο αυτό προσεγγίζει το 95% εκπαιδευόμενη σε ένα σώμα κειμένων με μόλις 20.000 λέξεις.

Ένας από τους βασικούς στόχους της παρούσας προσέγγισης ήταν η κατασκευή ενός πρακτικού αναγνωριστή μερών του λόγου, ο οποίος να μπορεί να εξυπηρετήσει τις ανάγκες πραγματικών συστημάτων επεξεργασίας φυσικής γλώσσας. Ο στόχος αυτός επίσης επιτεύχθηκε, καθώς ο αναγνωριστής έχει ήδη χρησιμοποιηθεί σε δεκάδες πρακτικών συστημάτων επεξεργασίας φυσικής γλώσσας, καθώς και από άλλες διδακτορικές διατριβές, τόσο ολοκληρωμένες [22], [63], όσο και εν εξελίξει. Η παρούσα προσέγγιση αποτέλεσε τον πρώτο αναγνωριστή μερών του λόγου που διατέθηκε

ελεύθερα για κάθε χρήση με άδεια χρήσης ανοικτού λογισμικού (LGPL), κάτω από την ονομασία «HBrill», σαν άρθρωμα της ανοικτού κώδικα πλατφόρμας επεξεργασίας φυσικής γλώσσας «Έλλογον» [53]. Ταυτόχρονα, η αρχική δημοσίευση της παρούσας προσέγγισης έχει δεχθεί ήδη μερικές δεκάδες ετεροαναφορές, γεγονός που καταδεικνύει την πρακτική χρήση της προσέγγισης από άλλους ερευνητές.

Τέλος, η επιλεγμένη προσέγγιση μπορεί εύκολα να συνεργαστεί με ένα μορφολογικό λεξικό για την Ελληνική γλώσσα. Μα τέτοια συνεργασία εξετάστηκε, και η μετρηθείσα απόδοση είναι η υψηλότερη που έχει αναφερθεί ποτέ στην διεθνή βιβλιογραφία για την αναγνώριση μερών του λόγου στα Ελληνικά.

# 4. Εξαγωγή Πληροφορίας: Αναγνώριση Ονομάτων Οντοτήτων

Στο κεφάλαιο αυτό θα ασχοληθούμε με την αναγνώριση ονομάτων οντοτήτων, εργασία που αποτελεί ένα από τα κύρια πεδία έρευνας της διατριβής. Στις επόμενες ενότητες θα δοθεί μια γενική εισαγωγή, ορίζοντας την εργασία της αναγνώρισης ονομάτων οντοτήτων και παρουσιάζοντας την αρχιτεκτονική ενός τέτοιου συστήματος. Υπάρχοντα συστήματα αναγνώρισης ονομάτων οντοτήτων που συναντώνται στην διεθνή βιβλιογραφία θα παρουσιαστούν ενώ θα ερευνηθεί ο τρόπος με τον οποίον η μηχανική μάθηση μπορεί να βοηθήσει στην ανάπτυξη τέτοιων συστημάτων με σκοπό να βελτιώσει την προσαρμογή τους σε νέες θεματικές περιοχές και γλώσσες.

## 4.1 Ορισμός προβλήματος

Με τον όρο *αναγνώριση ονομάτων οντοτήτων (named entity recognition and classification – NERC)* περιγράφεται η διαδικασία του εντοπισμού σε κείμενα ονομάτων οντοτήτων και η κατηγοριοποίηση των εντοπισμένων ονομάτων σε κατάλληλες σημασιολογικές κατηγορίες. Το τι είναι οντότητα δεν είναι εύκολο να οριστεί, αφού οι οντότητες σχετίζονται με την πληροφορία που θέλουμε να εντοπίσουμε μέσα στα κείμενα, και ποικίλει ανάλογα με την θεματική περιοχή. Παραδείγματος χάριν, αν η θεματική περιοχή των κειμένων είναι μετακινήσεις στελεχών, πιθανές οντότητες είναι πρόσωπα, τοποθεσίες, εταιρείες, θέσεις εργασίας ή ημερομηνίες, ενώ αν η θεματική περιοχή σχετίζεται με εκτοξεύσεις πυραύλων, τότε πιθανές οντότητες είναι πρόσωπα, τοποθεσίες, μοντέλα πυραύλων, ημερομηνίες ή ακόμη και ουράνια σώματα. Με απλά λόγια, ένας αναγνωριστής ονομάτων οντοτήτων προσπαθεί να εντοπίσει σε κείμενα ονόματα προσώπων, οργανισμών, τοποθεσιών κ.α., αλλά ταυτόχρονα να τα κατηγοριοποιήσει σε κατάλληλες σημασιολογικές κατηγορίες, ανάλογα με το αν περιγράφουν πρόσωπα, οργανισμούς, τοποθεσίες, κ.α., αίροντας πιθανές αμφισημίες.

Η αναγνώριση ονομάτων οντοτήτων πιθανώς να θεωρείται σαν μια εύκολη εργασία, αφού στην πλειονότητα των περιπτώσεων οι οντότητες εμφανίζονται γραμμένες με κεφαλαία γράμματα, τουλάχιστον όσον αφορά το πρώτο γράμμα κάθε λέξης που απαρτίζει μια οντότητα. Ωστόσο, στην πραγματικότητα τα πράγματα είναι αρκετά πιο περίπλοκα από ό,τι εμφανίζονται σε μια τέτοια απλοϊκή προσέγγιση. Σε αρκετές περιπτώσεις, ένας αναγνωριστής ονομάτων οντοτήτων πρέπει να αναγνωρίσει οντότητες σε κείμενα γραμμένα εξ' ολοκλήρου με κεφαλαία γράμματα (όπως π.χ. σε τίτλους ειδήσεων) ή να χειριστεί κείμενα όπου η γραφή με κεφαλαίο το πρώτο γράμμα κάθε λέξης έχει υιοθετηθεί και για στοιχεία που δεν αποτελούν οντότητες, τα οποία ένας αναγνωριστής δεν πρέπει να αναγνωρίσει. Για παράδειγμα, ένας αναγνωριστής μπορεί να κληθεί να επεξεργαστεί κείμενα όπως το ακόλουθο [59]:

> Ο **George McEnee** αναλαμβάνει τη Γενική Διεύθυνση της **Pepsico-Ήβη**, μετακινούμενος από την Εμπορική Διεύθυνση της **Pepsico Τσεχίας**.
> Ο **Νίκος Χαρμαντάς** ανέλαβε την Εμπορική Διεύθυνση της **Misko-Barilla**, στη θέση του **Κων/νου Γραβάνη** που αποχώρησε, μετακινούμενος από την Διεύθυνση Πωλήσεων της **Pepsico-Ήβη**.
> Τα καθήκοντά του στην **Pepsico-Ήβη**, ανέλαβε ο **Ανδρέας Ανδρεάδης**, 30 ετών, μέχρι τώρα Διευθυντής Στρατηγικού Σχεδιασμού της εταιρίας.

**Πίνακας 4: Τμήμα εγγράφου στο οποίο οι λεκτικές μονάδες με τονισμένη γραφή αποτελούν ονόματα οντοτήτων.**

Αν οι επιθυμητές οντότητες είναι τα πρόσωπα και οι εταιρείες, ο αναγνωριστής πρέπει να αγνοήσει τις θέσεις εργασίας έχοντας σαν αποτέλεσμα η πληροφορία από τους κεφαλαίους χαρακτήρες να μην αποτελεί σημαντική ένδειξη. Ταυτόχρονα, ένας αναγνωριστής δεν καλείται μόνο να εντοπίσει ονόματα οντοτήτων αλλά και να τα κατατάξει σε κατηγορίες, δηλαδή να κατηγοριοποιήσει το μέρος του κειμένου “Pepsico-Ήβη” σαν όνομα εταιρείας και το “Νίκος Χαρμαντάς” σαν όνομα προσώπου.

Μία από τις δυσκολότερες αλλά και σημαντικότερες εργασίες ενός αναγνωριστή ονομάτων οντοτήτων είναι η άρση της αμφισημίας. Αμφισημίες προκύπτουν όταν σε ένα συγκεκριμένο τμήμα κειμένου μπορούν να αποδοθούν περισσότερες από μία σημασιολογικές κατηγορίες. Παραδείγματος χάριν, η λέξη “Washington” μπορεί να κατηγοριοποιηθεί είτε σαν πρόσωπο, είτε σαν τοποθεσία, ανάλογα με το περιβάλλον μέσα στο οποίο βρίσκεται η λέξη. Αντίστοιχα, στο ακόλουθο κείμενο, το ποια ονόματα αποτελούν πρόσωπα και ποια όχι πρέπει να επιλεγούν προσεκτικά [59]:

| |
|---|
| Η Πόλυ Φιλιππoπούλου ανέλαβε τη Διεύθυνση Διαφήμισης των Εκδόσεων Κώστα Δραγούνη (περιοδικά ΚΑΙ, Τηλεθεατής και Avantage).<br>Είχε συνεργαστεί στο παρελθόν με τις Ειδικές Εκδόσεις Α. Τερζόπουλου, τον ΔΟΛ και την Ι.Γ. Δραγούνης. |

**Πίνακας 5: Τμήμα εγγράφου με αμφίσημα ονόματα οντοτήτων.**

## 4.2 Βιβλιογραφική επισκόπηση

Τη δεκαετία του 1990 διοργανώθηκαν μια σειρά από συνέδρια, υπό την αιγίδα της κυβερνητικής οργάνωσης των Η.Π.Α. *Defense Advanced Research Projects Agency (DARPA)*. Τα συνέδρια αυτά έγιναν γνωστά με την ονομασία *MUC* (*Message Understanding Conferences*) [7], [8], [19]. Σκοπός των συνεδρίων αυτών ήταν η αξιολόγηση συστημάτων εξαγωγής πληροφορίας κάτω από ένα προσεκτικά ελεγχόμενο περιβάλλον αξιολόγησης. Μετά το πέρας κάθε αξιολόγησης οι συμμετέχοντες παρουσίαζαν το σύστημα με το οποίο συμμετείχαν στην αξιολόγηση. Τα δύο πρώτα συνέδρια MUC ξεκίνησαν, σχεδιάστηκαν και πραγματοποιήθηκαν από την Beth Sundheim υπό την αιγίδα του Αμερικάνικου Ναυτικού και επικεντρώθηκαν στην εξαγωγή πληροφορίας από στρατιωτικά μηνύματα. Τα επόμενα συνέδρια MUC διοργανώθηκαν υπό την αιγίδα του Αμερικανικού ερευνητικού προγράμματος *TIPSTER* [64], και επικεντρώθηκαν στην εξαγωγή πληροφορίας από άρθρα διαφόρων ειδησεογραφικών πρακτορείων. Η συμβολή του προγράμματος TIPSTER ήταν σημαντική για την επιτυχία των συνεδρίων, αφού εκτός από τη διοργάνωση των συνεδρίων, το πρόγραμμα TIPSTER προσέφερε διάφορα εργαλεία αυτοματοποίησης των εργασιών αξιολόγησης, καθώς και έρευνα στον τομέα της μεθοδολογίας αξιολόγησης.

Το συνέδριο MUC-7 απετέλεσε το τελευταίο συνέδριο της σειράς συνεδρίων MUC. Διεξήχθη τον Απρίλιο του 1997 και αποτέλεσε το πιο ολοκληρωμένο συνέδριο, τόσο όσον αφορά τις εργασίες αξιολόγησης των συστημάτων όσο και το πλήθος των συστημάτων που συμμετείχαν στην αξιολόγηση. Για πρώτη φορά η εργασία αναγνώρισης ονομάτων οντοτήτων αξιολογήθηκε σε διαφορετική (αλλά παρεμφερή) θεματική περιοχή από την θεματική περιοχή που χρησιμοποιήθηκε για την εκπαίδευση των συστημάτων. Η θεματική περιοχή των δεδομένων εκπαίδευσης ήταν πτώσεις αεροσκαφών ενώ τα δεδομένα αξιολόγησης είχαν σαν θεματική περιοχή εκτοξεύσεις πυραύλων. Παρόλο που η θεματική περιοχή αξιολόγησης ήταν διαφορετική από εκείνη

των δεδομένων εκπαίδευσης, η απόδοση των συστημάτων που συμμετείχαν στο συνέδριο ήταν πάνω από το 80%, χωρίς να γίνει οποιαδήποτε αλλαγή στα συστήματα λόγω της αλλαγής θεματικής περιοχής. Ταυτόχρονα με το συνέδριο MUC-7 διοργανώθηκε και το δεύτερο συνέδριο *MET (Multilingual Entity Task)* το οποίο είχε σαν αντικείμενο την αξιολόγηση συστημάτων αναγνώρισης ονομάτων οντοτήτων σε διάφορες γλώσσες, εκτός της Αγγλικής. Οι γλώσσες αξιολόγησης του MET-2 ήταν η Ιαπωνική και η Κινεζική. Οι εργασίες αξιολόγησης του MUC-7 [65] ήταν η αναγνώριση ονομάτων οντοτήτων *(Named-Entity Extraction – NE)*, η επίλυση αναφορών *(Co-reference Resolution – CO)*, η πλήρωση σχεδιότυπων *(Template Element filling – TE)*, η συσχέτιση σχεδιότυπων (*Template Relations – TR*) και η δημιουργία των σεναρίων (*Scenario Template – ST*). Στις επόμενες παραγράφους θα παρουσιαστούν συνοπτικά οι εργασίες αυτές [65].

### *Αναγνώριση ονομάτων οντοτήτων (named entity extraction – NE)*

Σαν ονόματα οντοτήτων ορίσθηκαν κύρια ονόματα προσώπων, καθώς και άλλες ενδιαφέρουσες οντότητες, όπως ακρωνύμια, ονόματα εταιριών, κυβερνητικών οργανισμών, ημερομηνίες, ώρες, αριθμητικές εκφράσεις, χρηματικά ποσά, ποσοστά, κ.α. Ο ορισμός των ονομάτων οντοτήτων γίνεται με έμμεσο τρόπο, μέσω της παροχής ενός επισημειωμένου σώματος κειμένου. Η επισημείωση γίνεται με χρήση της SGML, ενώ η χρήση του επισημειωμένου σώματος κειμένου είναι πολλαπλή, περιλαμβάνοντας τον έμμεσο ορισμό των ονομάτων οντοτήτων, τη χρήση μέρους του σώματος κειμένου για την ανάπτυξη του αναγνωριστή ονομάτων οντοτήτων, και τη χρήση ενός άλλου μέρους του σώματος κειμένου για την αποτίμηση του συστήματος. Ένα παράδειγμα επισημειωμένου κειμένου αποτελεί το ακόλουθο Αγγλικό κείμενο (New York Times News Service – MUC-7) [65]:

The **<ENAMEX TYPE="LOCATION">U.K.</ENAMEX>** satellite television broadcaster said its subscriber base grew **<NUMEX TYPE="PERCENT">17.5 percent</NUMEX>** during **<TIMEX TYPE="DATE">the past year</TIMEX>** to 5.35 million.

**Πίνακας 6: Τμήμα εγγράφου επισημειωμένο κατά τα πρότυπα του MUC7 [65].**

Η εργασία της αναγνώρισης οντοτήτων εκτελέστηκε στα Ιαπωνικά και Κινέζικα (MET-2) ταυτόχρονα με τα Αγγλικά (MUC-7).

### *Επίλυση καθορισμού συν-αναφορών (co-reference resolution – CO)*

Η εργασία της *επίλυσης καθορισμού συν-αναφορών* (*co-reference resolution*) σχετίζεται με την ανεύρεση ονομάτων οντοτήτων που αναφέρονται στην ίδια οντότητα καθώς και στη συσχέτιση αντωνυμιών με οντότητες. Στις αξιολογήσεις των συνεδρίων MUC μόνο αναφορές του τύπου "*identity*" λήφθηκαν υπ' όψιν. Το ακόλουθο παράδειγμα από το MUC-7 (New York Times News Service) [65] παρουσιάζει αναφορές μεταξύ της αντωνυμίας "its" και της οντότητας "The U.K. satellite television broadcaster" καθώς και μεταξύ των "its subscriber base" και της τιμής "5.35 million". Η επίλυση καθορισμού συν-αναφορών είναι ο συνδετικός κρίκος μεταξύ της αναγνώρισης οντοτήτων και της επόμενης εργασίας, της πλήρωσης σχεδιότυπων.

| |
|---|
| **The U.K. satellite television broadcaster** said **its** **subscriber base** grew 17.5 percent during the past year to **5.35 million**. |

**Πίνακας 7: Παράδειγμα ισοδυναμίας κλάσεων κατά τα πρότυπα του MUC7 [65].**

### *Πλήρωση σχεδιότυπων (template element filling – TE)*

Τα χαρακτηριστικά των οντοτήτων απεικονίζονται σαν *πεδία (slots)* ενός *σχεδιότυπου (template element)*. Τα πεδία στα συνέδρια MUC περιλαμβάνουν όνομα, τύπο, *περιγραφέα (descriptor)* και κατηγορία. Τα χαρακτηριστικά σε ένα σχεδιότυπο εξυπηρετούν στον περαιτέρω προσδιορισμό της οντότητας, πέρα από το επίπεδο της ονομασίας. Όλα τα *εναλλακτικά ονόματα μιας οντότητας (mentions)* τοποθετούνται στο πεδίο «όνομα». Στο πεδίο «τύπος» τοποθετείται πληροφορία σχετικά με το αν μια οντότητα είναι πρόσωπο, οργανισμός, αντικείμενο (*artifact*) ή τοποθεσία. Όλοι οι σημαντικοί περιγραφείς που εμφανίζονται στο κείμενο τοποθετούνται στο πεδίο του «περιγραφέα». Τέλος, στο πεδίο «κατηγορία» τοποθετούνται πληροφορίες που εξαρτώνται από τον τύπο της οντότητας, όπως πρόσωπο, οργανισμός, τοποθεσία, κα. Για παράδειγμα, πρόσωπα μπορεί να είναι πολίτες, στρατιωτικοί κ.α., οργανισμοί μπορεί να είναι κυβερνητικοί, εταιρείες, μη κερδοσκοπικοί κ.α., ενώ τοποθεσίες μπορεί να είναι πόλεις, επαρχίες, νομοί, χώρες, περιοχές, σώματα νερού (π.χ. ποτάμια, λίμνες, θάλασσες, ωκεανοί, κλπ.), αεροδρόμια, κ.α. Για τις ανάγκες του MUC-7 τα αντικείμενα περιορίστηκαν σε οχήματα, τα οποία μπορούν να ταξιδεύουν στη ξηρά, το νερό ή τον αέρα. Ένα παράδειγμα στοιχείου σχεδιότυπου είναι το ακόλουθο [65]:

| |
|---|
| <ENTITY-9602040136-11> :=<br>ENT_NAME: “Dennis Gillespie”<br>ENT_TYPE: PERSON<br>ENT_DESCRIPTOR: “Capt.” / “the commander of Carrier Air Wing 11”<br>ENT_CATEGORY: PER_MIL |

**Πίνακας 8: Παράδειγμα συμπληρωμένου σχεδιότυπου οντότητας κατά τα πρότυπα του MUC7 [65].**

### *Συσχέτιση σχεδιότυπων (template relations – TR)*

Η εργασία της συσχέτισης σχεδιότυπων προσπαθεί να εντοπίσει σχέσεις μεταξύ οντοτήτων, όπως αντιπροσωπεύονται από συμπληρωμένα σχεδιότυπα. Για τους σκοπούς του MUC-7 οι σχέσεις αυτές περιορίστηκαν στις σχέσεις με οργανισμούς: *υπάλληλος_του (employee_of), προϊόν_του (product_of), τοποθεσία_του (location_of)*. Ωστόσο, η εργασία μπορεί εύκολα να επεκταθεί σε κάθε λογική συσχέτιση μεταξύ τύπων οντοτήτων. Ένα παράδειγμα συσχέτισης σχεδιότυπων από το MUC-7 είναι το ακόλουθο [65]:

```
<EMPLOYEE_OF-9602040136-5> :=
      PERSON:              <ENTITY-9602040136-11>
      ORGANIZATION:        <ENTITY-9602040136-1>
<ENTITY-9602040136-11> :=
      ENT_NAME:            “Dennis Gillespie”
      ENT_TYPE:            PERSON
      ENT_DESCRIPTOR:      “Capt.” / “the commander of Carrier Air Wing 11”
      ENT_CATEGORY:        PER_MIL
<ENTITY-9602040136-1> :=
      ENT_NAME:            “NAVY”
      ENT_TYPE:            ORGANIZATION
      ENT_CATEGORY:        ORG_GOVT
```

**Πίνακας 9: παράδειγμα συσχέτισης οντοτήτων κατά τα πρότυπα του MUC7 [65].**

### *Δημιουργία των σεναρίων (scenario template – ST)*

Ένα σενάριο (*Scenario Template – ST*) κατασκευάζεται έχοντας σαν βάση ένα γεγονός στο οποίο συμμετέχουν οντότητες.

#### 4.2.1 Επιδόσεις συστημάτων στα συνέδρια MUC

Οι εργασίες αξιολόγησης που εκτελέστηκαν για κάθε συνέδριο MUC (από το MUC-3 έως το MUC-7) παρουσιάζονται στον επόμενο πίνακα (Πίνακας 10), ενώ οι μέγιστες αποδόσεις για κάθε εργασία παρουσιάζονται στον πίνακα: Πίνακας 11 [65].

| Συνέδριο | NE | CO | TE | TR | ST | Πολύγλωσσο |
|---|---|---|---|---|---|---|
| **MUC-3** | ☒ | ☒ | ☒ | ☒ | ☑ | ☒ |
| **MUC-4** | ☒ | ☒ | ☒ | ☒ | ☑ | ☒ |
| **MUC-5** | ☒ | ☒ | ☒ | ☒ | ☑ | ☑ |
| **MUC-6** | ☑ | ☑ | ☑ | ☒ | ☑ | ☒ |
| **MUC-7** | ☑ | ☑ | ☑ | ☑ | ☑ | ☒ |
| **MET-1** | ☑ | ☒ | ☒ | ☒ | ☒ | ☑ |
| **MET-2** | ☑ | ☒ | ☒ | ☒ | ☒ | ☑ |

**Πίνακας 10: Οι εργασίες αξιολόγησης των συνεδρίων MUC-3 έως MUC-7.**

Από τον πίνακα (Πίνακας 11) παρατηρούμε ότι τα συστήματα που συμμετείχαν συνέδρια MUC (και κυρίως εκείνα του MUC-6), πέτυχαν πολύ υψηλές επιδόσεις στην εργασία της αναγνώρισης ονομάτων οντοτήτων (με την μετρική F-Measure να προσεγγίζει το 97%), στην θεματική περιοχή στην οποία αξιολογήθηκαν («γεγονότα επιτυχημένης διαδοχής διαχείρισης» – management succession events), επίδοση η οποία πλησιάζει την ανθρώπινη για αυτή την εργασία. Η αναγνώριση ονομάτων οντοτήτων (NE) αποτελεί μια από τις δύο επί μέρους εργασίας που ενδιαφέρουν την παρούσα διατριβή και εξετάζεται στο παρόν κεφάλαιο. Η δεύτερη επί μέρους εργασία είναι η πλήρωση σχεδιότυπων (TE), η οποία εξετάζεται στο κεφάλαιο 6.

| Συνέδριο | NE | CO | TE | TR | ST | Πολύγλωσσο |
|---|---|---|---|---|---|---|
| **MUC-3** | | | | | R < 50%<br>P < 70% | |
| **MUC-4** | | | | | F < 56% | |
| **MUC-5** | | | | | EJV F < 53%<br>EME F < 50% | JJV F < 64%<br>JME F < 57% |
| **MUC-6** | F < 97% | R < 63%<br>P < 72% | F < 80% | | F < 57% | |
| **MUC-7** | F < 94% | F < 62% | F < 87% | F < 76% | F < 51% | |
| **MET-1** | C F < 85%<br>J F < 93%<br>S F < 94% | | | | | |
| **MET-2** | C F < 91%<br>J F < 87% | | | | | |

R = Ανάκληση, P = Ακρίβεια, F = F-Measure με ίσο βάρος μεταξύ ανάκλησης/ακρίβειας,
E = Αγγλικά, C = Κινέζικα, J = Ιαπωνικά, S = Ισπανικά,
JV = Θεματική περιοχή “Joint Venture”, ME = Θεματική περιοχή “Microelectronics”

**Πίνακας 11: Μέγιστα αποτελέσματα για κάθε εργασία αξιολόγησης (MUC-3 έως MUC-7).**

### 4.2.2 Προσεγγίσεις των συστημάτων του MUC-7

#### Η χειρωνακτική προσέγγιση

Τα περισσότερα συστήματα που συμμετείχαν στο συνέδριο MUC-7 εντάσσονται στην κατηγορία των «χειρωνακτικά κατασκευασμένων» συστημάτων, με την έννοια ότι βασίζονται στην γνώση και την διαίσθηση των σχεδιαστών τους, όπως έχει αποτυπωθεί σε χειρωνακτικά κατασκευασμένους κανόνες και γραμματικές. Τυπικός εκπρόσωπος της κατηγορίας αυτής (ο οποίος εμφάνισε την υψηλότερη απόδοση μεταξύ όλων των χειρωνακτικών συστημάτων στο συνέδριο MUC-6) είναι το σύστημα «Proteus», του πανεπιστημίου της Νέας Υόρκης [66]. Το σύστημα αναγνώρισης ονομάτων οντοτήτων βασίζεται σε ένα μεγάλο αριθμό κανόνων, οι οποίοι εξαρτώνται από τα συμφραζόμενα, όπως οι ακόλουθοι [67]:

1. «Τίτλος» + «λέξη που αρχίζει από κεφαλαίο γράμμα» ⇒ «Τίτλος» + «Όνομα προσώπου»
2. «Όνομα ημερολογιακού μήνα» + «αριθμός $< 32$» ⇒ «Ημερομηνία»
3. “from” + «Ημερομηνία» + “to” + «Ημερομηνία» ⇒ «Ημερομηνία»

Αν και οι παραπάνω κανόνες θα λειτουργήσουν σε αρκετές περιπτώσεις σωστά (π.χ. ο πρώτος κανόνας θα αναγνωρίσει επιτυχώς ονόματα όπως “Mr. Jones” ή “Gen. Schwarzkopf”, ενώ ο δεύτερος θα αναγνωρίσει επιτυχώς ημερομηνίες όπως “February 28” ή “July 15”), εντούτοις θα οδηγήσουν και σε μερικές λάθος ανιχνεύσεις, όπως “Mrs. Field’s Cookies” (εταιρία κατασκευής μπισκότων – κανόνας 1) και “Long March 3” (ονομασία Κινεζικού πυραύλου – κανόνας 2).

Πέρα από την απόδοση, σημαντικά είναι και τα θέματα του κόστους και της σταθερότητας των αποτελεσμάτων. Η συγγραφή τόσο ειδικευμένων κανόνων απαιτεί εργασία αρκετών μηνών από ειδικούς με σημαντική εμπειρία στην υπολογιστική γλωσσολογία, ενώ ταυτόχρονα απαιτεί και προσήλωση σε μια μόνο θεματική περιοχή. Για παράδειγμα, στο συνέδριο MUC-7, το δεύτερο καλύτερο σύστημα, από την

IsoQuest [68], ήταν επίσης βασισμένο σε χειρωνακτικά κατασκευασμένους κανόνες, ενώ εμφάνισε μια απόδοση ίση με 91.6 (F-Measure), ενώ το επίσης χειρωνακτικό σύστημα FACILE [69] είχε σημαντικά χειρότερη απόδοση (F-Measure = 81.91). Αν και τα δύο συστήματα είχαν σημαντικές ομοιότητες στην προσέγγισή τους, η επένδυση χρόνου για την συγγραφή κανόνων ήταν εντελώς διαφορετική. Και τα δύο συστήματα προσαρμόστηκαν στα δεδομένα του MUC-7 μέσα σε ένα ημερολογιακό μήνα, όμως πάνω από το 90% των κανόνων του συστήματος NetOwl της IsoQuest προήρθε από το βασικό τους, εμπορικό, σύστημα, το οποίο αναπτυσσόταν τουλάχιστον δύο χρόνια. Αν και η απόδοση των χειρωνακτικά κατασκευασμένων συστημάτων έδειξε ότι η κατασκευή αποδοτικών αναγνωριστών ονομάτων οντοτήτων βασισμένων σε κανόνες είναι εφικτή, εντούτοις τέτοιες προσεγγίσεις εμφανίζουν ένα σημαντικό αριθμό μειονεκτημάτων, όπως:

- η κατασκευή τους είναι δαπανηρή, και απαιτούν την εμπειρία εξειδικευμένου προσωπικού (κυρίως από την περιοχή της υπολογιστικής γλωσσολογίας),
- η προσαρμογή σε νέες θεματικές περιοχές είναι επίσης μια χειρωνακτική διαδικασία,
- η προσαρμογή σε νέες γλώσσες είναι ακόμα πιο δύσκολη, αφού συχνά όλοι οι κανόνες πρέπει να ξαναγραφούν, και
- η απόδοση εξαρτάται σε μεγάλο βαθμό από την εμπειρία των εμπλεκόμενων ειδικών, αλλά και από το ποσό της εργασίας που έχει επενδυθεί στο σύστημα.

### Αυτοματοποιημένη προσέγγιση

Αναζητώντας λύσεις που περιορίζουν τα παραπάνω προβλήματα, οδηγηθήκαμε στην αξιοποίηση της μηχανικής μάθησης. Συστήματα βασισμένα σε μηχανική μάθηση έκαναν την εμφάνισή τους στο τελευταίο συνέδριο της σειράς συνεδρίων MUC, το MUC-7, σε μια προσπάθεια να μειώσουν τη συμβολή ειδικών στην ανάπτυξη ανάλογων συστημάτων. Μια από τις πρώτες προσεγγίσεις αφορούσε την χρήση *δέντρων αποφάσεων (decision trees)* [70], και συγκριμένα του αλγορίθμου C4.5 [25]. Ακολουθώντας μια προσέγγιση μηχανικής μάθησης με επίβλεψη, η χρήση ειδικών μετατέθηκε από την κατασκευή κανόνων, στην επισημείωση κειμένων, τα οποία αποτελούσαν τα παραδείγματα εκπαίδευσης του αλγορίθμου μηχανικής μάθησης. Ωστόσο, η επισημείωση είναι μια σαφώς ευκολότερη εργασία από την χειρωνακτική κατασκευή κανόνων. Έρευνες [71] έδειξαν ότι είναι δυνατή η επισημείωση δεκάδων χιλιάδων λέξεων εντός μιας ανθρωπομέρας. Η απόδοση του συστήματος ήταν ικανοποιητική (F-Measure = 79.49 % για την Ιαπωνική γλώσσα), όμως κύριο πλεονέκτημα του συστήματος ήταν η προσαρμοστικότητά του.

Το σύστημα IdentiFinder [72] αποτελεί ένα σύστημα αναγνώρισης ονομάτων οντοτήτων το οποίο βασίζεται ολοκληρωτικά σε μηχανική μάθηση, και συγκεκριμένα σε *κρυφά μοντέλα Markov (Hidden Markov Models – HMM)*. Αποτελώντας εμπορικό προϊόν της εταιρίας BBN, κατέλαβε την τρίτη θέση όσον αφορά την απόδοση στην εργασία αναγνώρισης ονομάτων σε σχέση με τα υπόλοιπα συστήματα (F-Measure = 90.44 %). Τα HMM αποδίδουν σε κάθε λέξη μία από τις επιθυμητές κατηγορίες (πρόσωπο, οργανισμός, τοποθεσία κ.α.) ή τον χαρακτηρισμό “NOT-A-NAME”. Η έκδοση του συστήματος που συμμετείχε στο MUC-7 χρησιμοποιούσε μια σειρά χαρακτηριστικών για κάθε λέξη, τα οποία περιελάμβαναν πληροφορία σχετική με αριθμητικές εκφράσεις, με το είδος των χαρακτήρων που απαρτίζουν τη λέξη (κεφαλαίοι – πεζοί) καθώς και με το αν η λέξη περιέχεται σε λίστες σημαντικών λέξεων (π.χ. προσδιοριστές εταιρειών). Για την εκπαίδευση του συστήματος χρησιμοποιήθηκε ένα αρκετά μεγάλο σώμα κειμένων, σημαντικά μεγαλύτερο από το επίσημο σώμα κειμένου του MUC-7. Συγκεκριμένα, το σύστημα εκπαιδεύτηκε σε κείμενα συνολικού μεγέθους 790,000

λέξεων, επισημειωμένων με περίπου 65,500 ονόματα οντοτήτων. Η θεματική περιοχή ήταν η ίδια με εκείνη του επίσημου σώματος εκπαίδευσης του MUC-7 (πτώσεις αεροσκαφών), με τις ειδήσεις να προέρχονται από το ίδιο ειδησεογραφικό πρακτορείο (NYT). Ενδεικτικά αναφέρεται ότι το μέγεθος του επίσημου σώματος εκπαίδευσης του MUC-7 ήταν 90,000 λέξεις.

### Υβριδικές προσεγγίσεις

Το πιο επιτυχημένο σύστημα στην αναγνώριση ονομάτων οντοτήτων στο συνέδριο MUC-7, ήταν αναμφισβήτητα το σύστημα LTG, του πανεπιστημίου του Εδιμβούργου [73]. Το σύστημα LTG περιέχει ένα υβριδικό σύστημα αναγνώρισης ονομάτων οντοτήτων, στο οποίο στάδια βασισμένα σε χειρωνακτικά κατασκευασμένες γραμματικές εναλλάσσονται με στάδια βασισμένα σε μηχανική μάθηση, και συγκεκριμένα σε αλγόριθμο *μέγιστης εντροπίας (maximum entropy)*. Το σύστημα LTG αποτελείται από μια σειρά γενικών εργαλείων τα οποία επικοινωνούν μεταξύ τους χρησιμοποιώντας τη γλώσσα XML. Τα εργαλεία αυτά μπορούν να προσαρμοστούν σε διαφορετικές περιοχές τροποποιώντας τις γραμματικές τις οποίες συμβουλεύονται κατά την διάρκεια της λειτουργίας τους.

Το σύστημα LTG ακολουθεί την τυπική δομή ενός συστήματος αναγνώρισης ονομάτων οντοτήτων (ενότητα 4.3). Το στάδιο της γλωσσικής προ-επεξεργασίας αποτελείται από έναν αναγνωριστή λεκτικών μονάδων (*lttok*), έναν αναγνωριστή προτάσεων (*ltstop*) και μια μονάδα απόδοσης μερών του λόγου (*ltpos*). Ο αναγνωριστής λεκτικών μονάδων χρησιμοποιεί χειρωνακτικά κατασκευασμένους κανόνες για τον προσδιορισμό των λεκτικών μονάδων (λέξεων). Ο αναγνωριστής προτάσεων βασίζεται σε μηχανική μάθηση (*maximum entropy*) και είναι υπεύθυνος τόσο για τον εντοπισμό των προτάσεων όσο και για την άρση της αμφισημίας του κατά πόσο μία τελεία αποτελεί μέρος σύντμησης, τέλος πρότασης ή και τα δύο. Τέλος, ή απόδοση μερών του λόγου τελείται με χρήση μηχανικής μάθησης και πιο συγκεκριμένα με τη χρήση ενός κρυφού μοντέλου Markov του οποίου η αρχικοποίηση γίνεται με την χρήση ενός τρι-γραμμικού (*trigram*) μοντέλου μέγιστης εντροπίας, το οποίο αποδίδει σε κάθε λέξη το πιο πιθανό μέρος του λόγου. Η μονάδα απόδοσης μερών του λόγου τέλος είναι ικανή να αποδώσει μέρη του λόγου ακόμα και σε άγνωστες λέξεις, ιδιότητα που αποδείχθηκε καθοριστική για τον εντοπισμό ονομάτων οντοτήτων στο εν λόγω σύστημα.

| Κανόνας | Κατηγορία | Παράδειγμα |
|---|---|---|
| Xxxx+ is a? JJ* PROF | PERS | Yuri Gromov is a former director |
| PERSON-NAME is a? JJ* REL | PERS | John White is beloved brother |
| Xxxx+, a JJ* PROF, | PERS | White, a retired director, |
| Xxxx+ ,? whose REL | PERS | Nunberg, whose stepfather |
| Xxxx+ himself | PERS | White himself |
| Xxxx+, DD+, | PERS | White, 33, |
| shares of Xxxx+ | ORG | shares of Eagle |
| PROF of/at/with Xxxx+ | ORG | director of Trinity Motors |
| in/at LOC | LOC | in Washington |
| Xxxx+ area | LOC | Beribidjan area |

**Πίνακας 12: Παραδείγματα "σίγουρων κανόνων". Σαν Xxxx+ σημειώνεται μια σειρά από λέξεις που αρχίζουν με κεφαλαίο χαρακτήρα, σαν DD ένας αριθμός, σαν PROF ένα επάγγελμα (director, manager, analyst κ.α.), σαν REL μια σχέση (sister, nephew κ.α.), σαν JJ* μια σειρά από κανένα ή περισσότερα επίθετα, σαν LOC μια γνωστή τοποθεσία, σαν PERSON-NAME ένα έγκυρο όνομα προσώπου αναγνωρισμένο από γραμματική ονομάτων προσώπων. [73]**

Το υποσύστημα του λεξικού εμφανίζεται με μειωμένη σημαντικότητα/χρήση στο σύστημα αυτό, αφού χρησιμοποιείται αποκλειστικά για τον εντοπισμό πιθανών ονομάτων οντοτήτων, αφήνοντας τις τελικές αποφάσεις στο υποσύστημα της γραμματικής. Το λεξικό αποτελείται από ονόματα προσώπων, οργανισμούς και τοποθεσίες. Αντίθετα, ιδιαίτερη βαρύτητα έχει δοθεί στο υποσύστημα της γραμματικής. Αρχικά, χειρωνακτικά κατασκευασμένες γραμματικές χρησιμοποιούνται για την αναγνώριση χρονικών και αριθμητικών οντοτήτων, οι οποίες θεωρήθηκαν απλούστερες στον εντοπισμό τους από τις υπόλοιπες οντότητες. Για τον εντοπισμό των υπόλοιπων οντοτήτων χρησιμοποιούνται τόσο χειρωνακτικά κατασκευασμένες γραμματικές όσο και μηχανική μάθηση, ενώ η διαδικασία της αναγνώρισης μπορεί να χωριστεί στα ακόλουθα πέντε στάδια [73]:

1. «Σίγουροι κανόνες» (*sure-fire rules*): Οι κανόνες αυτοί είναι χειρωνακτικά κατασκευασμένοι και βασίζονται ολοκληρωτικά στα συμφραζόμενα, δηλαδή ενεργοποιούνται μόνο όταν πιθανό όνομα οντότητας περιέχεται σε συγκεκριμένο περιβάλλον. Οι κανόνες αυτοί βασίζονται κυρίως σε *προσδιοριστές ονομάτων εταιριών (company designators)* όπως τα “Ltd.”, “Inc.”, κ.α., τίτλους προσώπων (“Mr.”, “Dr.”, “Sen.”) ή σε περιβάλλοντα όπως αυτά που απεικονίζονται στον πίνακα (Πίνακας 12).
2. «Μερικό ταίριασμα 1» (*partial match 1*): Στο στάδιο αυτό πραγματοποιείται ένα πιθανοτικό μερικό ταίριασμα οντοτήτων που έχουν ήδη αναγνωριστεί. Από κάθε αναγνωρισμένο όνομα οντότητας δημιουργούνται εναλλακτικές ονομασίες, χρησιμοποιώντας τις λέξεις που απαρτίζουν το αρχικό όνομα, διατηρώντας όμως τη σειρά των λέξεων. Για παράδειγμα, από το όνομα οντότητας “Lockheed Martin Production” θα δημιουργηθούν τα ονόματα “Lockheed Martin Production”, “Lockheed Martin”, “Lockheed Production”, “Martin Production”, “Lockheed” και “Martin”, όλες οι εμφανίσεις των οποίων στο κείμενο θα χαρακτηριστούν σαν πιθανές οντότητες. Στην συνέχεια, ένα σύστημα μηχανικής μάθησης (μέγιστης εντροπίας) καλείται να προσδιορίσει αν κάθε πιθανό όνομα οντότητας είναι σωστό ή όχι, χρησιμοποιώντας πληροφορία σχετικά με το περιβάλλον της πιθανής οντότητας, τη θέση της στην πρόταση, το εάν η πιθανή ονομασία περιέχει κεφαλαίους χαρακτήρες ή εάν η ονομασία έχει χρησιμοποιηθεί αλλού στο κείμενο γραμμένη μόνο με πεζούς χαρακτήρες.
3. «Χαλαρότεροι κανόνες» (*rule relaxation*): Στο στάδιο αυτό εφαρμόζονται χειρωνακτικά κατασκευασμένοι κανόνες, οι οποίοι έχουν ασθενέστερη εξάρτηση από το περιβάλλον, ενώ χρησιμοποιούν σε σημαντικό βαθμό τα αποτελέσματα των προηγούμενων σταδίων. Βασιζόμενο στην υπόθεση ότι αρκετά ονόματα οντοτήτων έχουν ήδη αναγνωριστεί με αρκετή ακρίβεια, στο στάδιο αυτό εφαρμόζονται πιο γενικοί κανόνες με σκοπό την αναγνώριση των υπολοίπων οντοτήτων. Για παράδειγμα, αν λέξη που έχει χαρακτηριστεί από το λεξικό σαν μικρό κύριο όνομα ακολουθείται από λέξεις που ξεκινούν με κεφαλαίο χαρακτήρα, τότε το όνομα αναγνωρίζονται σαν όνομα προσώπου, ανεξάρτητα από το περιβάλλον στο οποίο εμφανίζεται.
4. «Μερικό ταίριασμα 2» (*partial match 2*): Έχοντας εξαντλήσει όλους τους πόρους του (λεξικό και γραμματική), το σύστημα εκτελεί ένα ακόμα στάδιο μερικού ταιριάσματος, πανομοιότυπο με το “μερικό ταίριασμα 1”.
5. “Αναγνώριση σε τίτλους” (*title assignment*): Έχοντας ολοκληρώσει την αναγνώριση ονομάτων οντοτήτων στο κυρίως κείμενο, το σύστημα προσπαθεί να αναγνωρίσει ονόματα οντοτήτων σε περιοχές γραμμένες μόνο με κεφαλαίους χαρακτήρες, όπως ο τίτλος του κειμένου. Η αναγνώριση αυτή βασίζεται κυρίως σε ταίριασμα μεταξύ ονομάτων οντοτήτων που αναγνωρίστηκαν στο κείμενο και εμφανίσεών τους στους τίτλους.

Το σύστημα LTG συμμετείχε στο συνέδριο MUC-7, επιτυγχάνοντας μάλιστα την καλύτερη απόδοση στην εργασία αναγνώρισης ονομάτων σε σχέση με τα υπόλοιπα συστήματα (F-Measure = 93.39 %). Τέλος, στον επόμενο πίνακα (Πίνακας 13) παρουσιάζεται η απόδοση του συστήματος μετά την ολοκλήρωση κάθε σταδίου του συστήματος για τα κείμενα αποτίμησης του MUC-7.

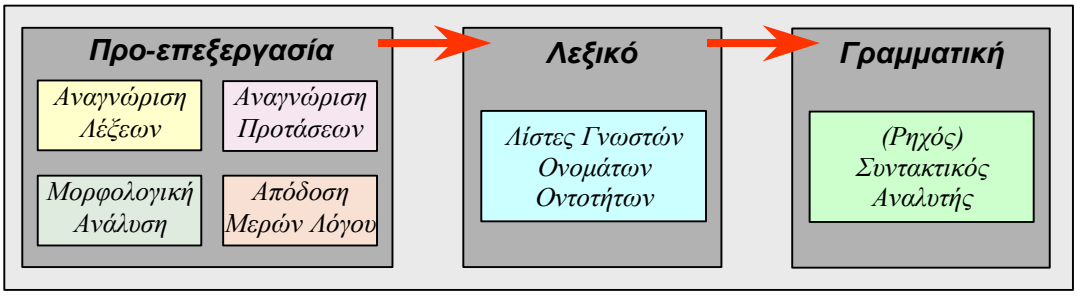
| Στάδιο | Οργανισμοί (%) | Πρόσωπα (%) | Τοποθεσίες (%) |
|---|---|---|---|
| "Σίγουροι κανόνες" | R: 42 P: 98 | R: 40 P: 99 | R: 36 P: 96 |
| "Μερικό ταίριασμα 1" | R: 75 P: 98 | R: 80 P: 99 | R: 69 P: 93 |
| "Χαλαρότεροι κανόνες" | R: 83 P: 96 | R: 90 P: 98 | R: 86 P: 93 |
| "Μερικό ταίριασμα 2" | R: 85 P: 96 | R: 93 P: 97 | R: 88 P: 93 |
| "Αναγνώριση σε τίτλους" | R: 91 P: 95 | R: 95 P: 97 | R: 95 P: 93 |

**Πίνακας 13: Η απόδοση του συστήματος LTG έχοντας ολοκληρώσει διάφορα στάδια του συστήματος. R = ανάκτηση (recall) P = ακρίβεια (precision).**

## 4.3 Τυπική Αρχιτεκτονική ενός συστήματος αναγνώρισης ονομάτων οντοτήτων

Ένα τυπικό σύστημα αναγνώρισης ονομάτων οντοτήτων (*NERC*) αποτελείται από τρία υποσυστήματα (Εικόνα 5):

- το υποσύστημα της προ-επεξεργασίας
- ένα λεξικό, και
- μια γραμματική.




**Εικόνα 5: Υποσυστήματα ενός τυπικού συστήματος αναγνώρισης ονομάτων οντοτήτων.**

Το πρώτο υποσύστημα είναι υπεύθυνο για την διεξαγωγή στοιχειωδών γλωσσολογικών εργασιών, όπως η αναγνώριση λέξεων και προτάσεων, η απόδοση μερών του λόγου σε λέξεις, και ίσως η μορφολογική ανάλυση. Το δεύτερο υποσύστημα (λεξικό) είναι υπεύθυνο για τον εντοπισμό και την κατηγοριοποίηση *γνωστών* ονομάτων οντοτήτων, που συνήθως περιέχονται σε *λίστες γνωστών ονομάτων (gazetteers)*. Τέλος, το τρίτο υποσύστημα (*γραμματική αναγνώρισης ονομάτων οντοτήτων*) είναι υπεύθυνο για τον τελικό έλεγχο των αποτελεσμάτων του λεξικού, την άρση τυχόν αμφισημιών, καθώς και τον εντοπισμό και κατηγοριοποίηση ονομάτων οντοτήτων που δεν περιέχονται στο λεξικό. Τόσο το λεξικό όσο και η γραμματική είναι πόροι που εξαρτώνται σημαντικά από την θεματική περιοχή.

Ένα σημαντικό πρόβλημα που συνοδεύει τα συστήματα αναγνώρισης ονομάτων οντοτήτων αφορά την αναβάθμισή τους με την πάροδο του χρόνου, αλλά και τη μεταφερσιμότητά τους σε νέες θεματικές περιοχές. Συνήθως, οι οντότητες μεταβάλλονται με την πάροδο του χρόνου, καθώς νέες προστίθενται ή υπάρχουσες αποκτούν νέα σημασία, καθιστώντας την αναβάθμιση ενός συστήματος επιτακτική

ανάγκη. Για παράδειγμα, έχοντας το κείμενο "... Λονδίνο 2012 ..." ένα μη επαρκώς ενημερωμένο σύστημα θα αναγνωρίσει μόνο τη λέξη "Λονδίνο" σαν τοποθεσία και ίσως αποτύχει στο να αναγνωρίσει τον οργανισμό "Λονδίνο 2012", μια οντότητα που δημιουργήθηκε μετά την ανάπτυξη του συστήματος. Η μεταφερσιμότητα σχετίζεται με την προσαρμογή ενός υπάρχοντος συστήματος σε μια νέα θεματική περιοχή ή γλώσσα. Έχοντας για παράδειγμα ένα υπάρχον σύστημα για μια θεματική περιοχή όπως είναι οι χρηματο-οικονομικές ειδήσεις, να είναι δυνατή η τροποποίησή του ώστε να μπορεί να λειτουργεί με ειδήσεις σχετικά με εκτοξεύσεις πυραύλων. Και οι δύο αυτές διαδικασίες είναι αρκετά χρονοβόρες και αρκετές φορές αδύνατες χωρίς την συμβολή ειδικών.

Η μεταφερσιμότητα σε μια νέα γλώσσα, αποτελεί μια ακόμα μεγαλύτερη πρόκληση, σε σχέση με την μεταφορά σε μια νέα θεματική περιοχή. Η μεγαλύτερη δυσκολία εμφανίζεται στην προσαρμογή του υποσυστήματος της γραμματικής, καθώς κάθε γλώσσα έχει τις δικές της ιδιομορφίες και τους δικούς της τρόπους έκφρασης. Στα πλαίσια αυτής της διδακτορικής διατριβής αναπτύχθηκε ένα συμβατικό σύστημα αναγνώρισης ονομάτων οντοτήτων, το οποίο χρησιμοποιεί χειρωνακτικά κατασκευασμένους κανόνες στο υποσύστημα της γραμματικής. Αν και εκτός των στόχων της διατριβής, το σύστημα αυτό αποτέλεσε μια ευκαιρία να αξιολογηθούν στην πράξη τα προβλήματα της χειρωνακτικής κατασκευής ενός αναγνωριστή ονομάτων οντοτήτων, αλλά και της προσαρμογής του σε μια νέα γλώσσα. Το χειρωνακτικά κατασκευασμένο σύστημα, το οποίο περιγράφεται στο ΠΑΡΑΡΤΗΜΑ ΙΙ, βασίστηκε σε ένα αντίστοιχο σύστημα για την Αγγλική γλώσσα, οι γραμματικοί κανόνες του οποίου προσαρμόστηκαν για την Ελληνική γλώσσα. Ωστόσο, ένα μεγάλο μέρος των αρχικών κανόνων για τα Αγγλικά, δεν κατέστη δυνατό να μεταφερθούν στην Ελληνική γλώσσα, καθώς η χρήση προσδιοριστών (όπως τα Ελληνικά αντίστοιχα των "Mr.", "Mrs.", "Ltd.", "Co.", κλπ.) δεν χρησιμοποιούνται τόσο συχνά στα Ελληνικά, ενώ η εύρεση αξιόπιστων προτύπων συμφραζομένων (για την μετατροπή τους σε γραμματικούς κανόνες) είναι σημαντικά δυσκολότερη για τα Ελληνικά.

## 4.4 Η προτεινόμενη προσέγγιση

Η χρήση μεθόδων μηχανικής μάθησης έχει προταθεί για την αντιμετώπιση των προβλημάτων της αναγνώρισης ονομάτων οντοτήτων σαν ένας εναλλακτικός τρόπος ανάπτυξης τέτοιων συστημάτων. Σκοπός της χρήσης μηχανικής μάθησης είναι είτε η υποβοήθηση της διαδικασίας αναβάθμισης των υποσυστημάτων που εξαρτώνται από την θεματική περιοχή ή τη γλώσσα, είτε η πλήρης αντικατάσταση αυτών των υποσυστημάτων. Καθώς η παρούσα διδακτορική διατριβή έχει σαν κύριο στόχο την αντιμετώπιση του προβλήματος της αναβάθμισης και προσαρμογής αυτών των συστημάτων, ερευνήθηκαν διάφορα σενάρια χρήσης της μηχανικής μάθησης. Πιο συγκεκριμένα, χρησιμοποιήθηκε μηχανική μάθηση για:

- Την ολοκληρωτική αντικατάσταση του υποσυστήματος της γραμματικής.
- Την προσαρμογή/εμπλουτισμό του υποσυστήματος του λεξικού.
- Την ενημέρωση ενός συστήματος αναγνώρισης ονομάτων οντοτήτων, μέσω της κατασκευής ενός συστήματος ικανού να ελέγχει την έξοδο ενός συστήματος αναγνώρισης ονομάτων οντοτήτων, με σκοπό την σηματοδότηση κατά πόσον το τελευταίο πρέπει να ενημερωθεί.

## 4.5 Αντικατάσταση του υποσυστήματος της γραμματικής

Το υποσύστημα της γραμματικής αποτελεί το δυσκολότερο υποσύστημα στο να δημιουργηθεί ή να προσαρμοστεί σε νέες θεματικές περιοχές και ακόμα και γλώσσες, καθιστώντας την αντικατάστασή του μέσω μηχανικής μάθησης περισσότερο από θεμιτή. Στην ενότητα αυτή παρουσιάζονται δύο διαφορετικές προσεγγίσεις, που στόχο έχουν την αντικατάσταση του υποσυστήματος της γραμματικής εξ' ολοκλήρου με τεχνικές

μηχανικής μάθησης, μετατρέποντας την χειρωνακτική συγγραφή γραμματικών κανόνων σε επισημείωση παραδειγμάτων σε ένα σώμα κειμένου, μια ευκολότερη διαδικασία.

Η πρώτη προσέγγιση (προσέγγιση Α) εξετάζει συγκριτικά δύο διαφορετικούς αλγορίθμους στην εργασία της αναγνώρισης ονομάτων οντοτήτων. Ο πρώτος αλγόριθμος που εξετάζεται αφορά ένα δέντρο αποφάσεων, ενώ ο δεύτερος ένα νευρωνικό δίκτυο. Η απόδοση και των δύο αλγόριθμων αξιολογείται σε δύο σώματα κειμένων, χειρωνακτικά επισημειωμένα, για την Αγγλική και Ελληνική γλώσσα. Η δεύτερη προσέγγιση (προσέγγιση Β) αφορά τον συνδυασμό αλγορίθμων μηχανικής μάθησης, έχοντας σαν σκοπό την συνδυασμένη αξιοποίηση συστημάτων κατηγοριοποίησης τόσο σε επίπεδο λέξης, όσο και σε επίπεδο φράσης. Το σύστημα που υλοποιεί την δεύτερη προσέγγιση αποτιμάται επίσης σε δύο γλώσσες (Αγγλικά και Ελληνικά) με την βοήθεια χειρωνακτικά επισημειωμένων σωμάτων κειμένων, ενώ η απόδοσή του για την Αγγλική γλώσσα συγκρίνεται με αντίστοιχο σύστημα κατασκευασμένο από το Πανεπιστήμιο του Εδιμβούργου, Ηνωμένο Βασίλειο.

## 4.6 Προσέγγιση Α: μηχανική μάθηση στην αναγνώριση ονομάτων οντοτήτων

Η χειρωνακτική προσαρμογή συστημάτων *NERC* σε μια νέα θεματική περιοχή ή σε μια νέα γλώσσα είναι μια ιδιαιτέρως χρονοβόρα διαδικασία, και σε μερικές περιπτώσεις αδύνατη, λόγω της έλλειψης ειδικών. Κατά συνέπεια, η αυτόματη απόκτηση/προσαρμογή των απαιτούμενων πόρων από σώματα κειμένων είναι ιδιαίτερα επιθυμητή. Οι τεχνικές μηχανικής μάθησης είναι ταξινομημένες σε δύο ευρείες κατηγορίες: επιβλεπόμενες και χωρίς επίβλεψη. Οι επιβλεπόμενες τεχνικές μάθησης απαιτούν την ύπαρξη παραδειγμάτων εκπαίδευσης, τα οποία έχουν επισημειωθεί με τις κατάλληλες κατηγορίες. Αντίθετα, οι τεχνικές χωρίς επίβλεψη υποθέτουν ότι η σωστή ταξινόμηση των παραδειγμάτων εκπαίδευσης δεν είναι γνωστή εκ των προτέρων, και ταξινομούν τα παραδείγματα σύμφωνα με μια μετρική ομοιότητας. Οι επιβλεπόμενες τεχνικές είναι πιο δαπανηρές από τις τεχνικές χωρίς επίβλεψη, λόγο της προσπάθειας που απαιτείται για την επισημείωση των δεδομένων εκπαίδευσης. Ωστόσο, η επιπλέον πληροφορία που προστίθεται στα δεδομένα, οδηγεί συχνά σε συστήματα με καλύτερη απόδοση. Τα συστήματα Nymble [72], Alembic [74], [75], RoboTag [76], και AutoLearn [77] είναι παραδείγματα συστημάτων που εκμεταλλεύονται επιβλεπόμενες τεχνικές μηχανικής μάθησης. Αντίθετα, το σύστημα που αναπτύχθηκε για την Ιταλική γλώσσα [78] είναι ένα παράδειγμα συστήματος που εκμεταλλεύεται μάθηση χωρίς επίβλεψη. Οι αλγόριθμοι μηχανικής μάθησης μπορούν να ταξινομηθούν περαιτέρω σύμφωνα με την αναπαράσταση μοντέλου που χρησιμοποιούν σε *συμβολικούς (symbolic)* και *στοχαστικούς ή αριθμητικούς (subsymbolic or numeric)*. Οι συμβολικές μέθοδοι χρησιμοποιούν μια αναπαράσταση με διακριτά σύμβολα, ενώ οι στοχαστικές μέθοδοι χρησιμοποιούν συνεχείς αριθμητικές τιμές στην αναπαράστασή τους.

Η προσέγγιση που παρουσιάζεται σε αυτή την ενότητα, εξετάζει το πρώτο στόχο της παραγράφου 4.4, δηλαδή της ολοκληρωτικής αντικατάστασης του υποσυστήματος της γραμματικής. Παρουσιάζονται δύο διαφορετικές μέθοδοι αναγνώρισης ονομάτων οντοτήτων, βασισμένων σε μηχανική μάθηση: η πρώτη μέθοδος αφορά έναν επιβλεπόμενο συμβολικό αναγνωριστή ονομάτων οντοτήτων, βασισμένο στον αλγόριθμο εκμάθησης δέντρων αποφάσεων C4.5 [25], ενώ η δεύτερη μέθοδος βασίζεται σε ένα νευρωνικό δίκτυο. Το νευρικό δίκτυο που χρησιμοποιήθηκε ήταν τυπικό feed-forward πολυστρωματικό perceptron, με ένα *κρυμμένο επίπεδο απόφασης (hidden layer)*. Τα δέντρα αποφάσεων είναι χαρακτηριστικοί εκπρόσωποι των συμβολικών τεχνικών μηχανικής μάθησης, ενώ τα νευρωνικά δίκτυα είναι χαρακτηριστικοί εκπρόσωποι των αριθμητικών αλγορίθμων μηχανικής μάθησης.

Ιδιαίτερη έμφαση δίνεται στην αναπαράσταση των παραδειγμάτων εκπαίδευσης για τις δύο μεθόδους. Η συμβολική φύση των δέντρων απόφασης τα καθιστά καταλληλότερα για εργασίες γλωσσικής τεχνολογίας με συμβολικά κειμενικά δεδομένα. Από την άλλη, τα νευρωνικά δίκτυα απαιτούν την κωδικοποίηση των δεδομένων σε αριθμητικά διανύσματα χαρακτηριστικών, η οποία μπορεί να είναι προβληματική εάν θέλουμε να διατηρήσουμε όλες τις αρχικές πληροφορίες. Σε αυτές τις προσεγγίσεις, χρησιμοποιούμε μια απλή συμβολική αναπαράσταση, καθώς επίσης και μια αριθμητική αναπαράσταση, η οποία αγνοεί τη σειρά των λέξεων μέσα σε ένα όνομα οντότητας, προκειμένου να μειωθεί η *διαστατικότητα (dimensionality)* του προβλήματος σε διαχειρίσιμα επίπεδα. Η μέθοδος εκμάθησης δέντρων αποφάσεων αξιολογείται και με τις δύο αναπαραστάσεις, οδηγώντας σε ενδιαφέροντα συμπεράσματα σχετικά με την μη χρησιμοποίηση της σειράς των λέξεων στην αριθμητική αναπαράσταση.

### 4.6.1 Μηχανική μάθηση και αναπαράσταση γνώσης

Ένα από τα προβλήματα που εμφανίζονται κατά την εφαρμογή μηχανικής μάθησης, είναι φυσικά η εύρεση μιας κατάλληλης αναπαράστασης. Με δεδομένο ότι η πλειονότητα των αλγορίθμων μηχανικής μάθησης απαιτούν σαν είσοδο διανύσματα χαρακτηριστικών *σταθερού μήκους*, η εφαρμογή τους σε προβλήματα επεξεργασίας φυσικής γλώσσας δεν είναι αυτονόητη, αφού σε αυτή την περιοχή τίποτα δεν είναι σταθερό σε μήκος. Οι προτάσεις σαφώς δεν έχουν σταθερά μήκη, αλλά ακόμα και μικρότερα τμήματα, όπως τα ονόματα οντοτήτων, κυμαίνονται από μία ή δύο λέξεις, μέχρι τις δέκα ή δώδεκα, όπως για παράδειγμα οι θέσεις εργασίας σε αγγελίες ευρέσεως εργασίας. Μια λύση σε αυτό το πρόβλημα, είναι η επιλογή ενός ορίου, ενός μέγιστου μήκους, και η αποδοχή ότι το σύστημα δεν θα καταφέρει να επεξεργαστεί σωστά περιπτώσεις που υπερβαίνουν αυτό το όριο. Μια τέτοια παραδοχή ακολουθήθηκε στο πλαίσιο αυτής της προσέγγισης, βασιζόμενη και στην ιδιότητα των ονομάτων οντοτήτων να αποτελούν πάντα *ονοματικές φράσεις (noun phrases)*: στην πλειονότητα τους, τα ονόματα οντοτήτων αποτελούνται από ουσιαστικά, επίθετα, ξένες λέξεις και συνδέσμους, συστατικά που απαρτίζουν επίσης τις ονοματικές φράσεις.

Η ιδέα πίσω από την προτεινόμενη αναπαράσταση είναι απλή: με δεδομένο ότι τα ονόματα οντοτήτων είναι ονοματικές φράσεις, εφαρμόζεται στα κείμενα ένας *ρηχός συντακτικός αναλυτής (shallow syntactic parser)*, ο οποίος στοχεύει στην αναγνώριση φράσεων, κυρίως ονοματικών (*noun phrase chunker – NP chunker*). Από τη στιγμή που τα ονόματα οντοτήτων ταυτίζονται με μερικές από τις εξαγόμενες φράσεις, το πρόβλημα μετατίθεται στον διαχωρισμό των φράσεων σε ονόματα οντοτήτων ή μη, δηλ. σε ένα πρόβλημα κατηγοριοποίησης το οποίο μπορεί να μοντελοποιηθεί ευκολότερα ώστε να λυθεί μέσω μηχανικής μάθησης. Θέτοντας ένα όριο στο μέγιστο μήκος των λέξεων (π.χ. 10 λέξεις) που μπορούν να απαρτίσουν ένα όνομα οντότητας, και καθορίζοντας και των αριθμό των χαρακτηριστικών ανά λέξη (π.χ. δύο χαρακτηριστικά), το πρόβλημα της κατηγοριοποίησης ονοματικών φράσεων σε ονόματα οντοτήτων μπορεί να αναπαρασταθεί με διανύσματα σταθερού μήκους (20 χαρακτηριστικών στην συγκεκριμένη περίπτωση).

Συνεπώς, η συγκριμένη προσέγγιση (προσέγγιση Β) μπορεί να χωριστεί σε δύο στάδια επεξεργασίας:

- Στάδιο 1: Προσδιορισμός όλων των ονοματικών φράσεων σε ένα σώμα κειμένων, με την χρήση κατάλληλου *ρηχού συντακτικού αναλυτή (NP chunker)*. Ο αναλυτής αυτός πρέπει να είναι σε θέση να αναγνωρίσει τις στοιχειώδεις ονοματικές φράσεις, αλλά και φράσεις πιο σύνθετες, οι οποίες έχουν προκύψει από την συνένωση άλλων ονοματικών φράσεων.

- Στάδιο 2: Σημασιολογική κατηγοριοποίηση των ονοματικών φράσεων. Μόλις έχουν προσδιοριστεί όλες οι ονοματικές φράσεις, κάθε φράση κατηγοριοποιείται σε μια κατάλληλη κατηγορία (π.χ. όνομα προσώπου, οργανισμού, τοποθεσίας, κ.α.), ή σε μια ειδική κατηγορία που δηλώνει ότι η φράση δεν είναι όνομα οντότητας (ώστε να ικανοποιηθεί η διπλή φύση του συστήματος αναγνώρισης ονομάτων οντοτήτων, δηλαδή ο προσδιορισμός και η ταξινόμηση των ονομάτων οντοτήτων).

Όπως έχει ήδη αναφερθεί, η επιλογή της αναπαράστασης των δεδομένων ένα κρίσιμο ζήτημα στην εφαρμογή της μηχανικής μάθησης. Και οι δύο από τους αλγορίθμους που χρησιμοποιήθηκαν, απαιτούν τα δεδομένα να παρασχεθούν μέσω διανυσμάτων χαρακτηριστικών σταθερού μήκους. Αυτή η αναπαράσταση απαιτεί τα δεδομένα (στην περίπτωσή μας οι ονοματικές φράσεις) για να κωδικοποιηθούν σε ένα καθορισμένου μήκους διάνυσμα, με τιμές από ένα συγκεκριμένο σύνολο χαρακτηριστικών. Στην αναπαράσταση που σχεδιάστηκε και θα παρουσιαστεί στις επόμενες ενότητες, εστιάσαμε κυρίως σε δύο ζητήματα:

- Πώς να αναπαρασταθεί μια μεμονωμένη λέξη, με τρόπο κατάλληλο για το πρόβλημα αναγνώρισης και ταξινόμησης που μας ενδιαφέρει.
- Πώς να συνδυαστούν αναπαραστάσεις μεμονωμένων λέξεων ώστε να διαμορφωθεί ένα καθορισμένου μήκους διάνυσμα χαρακτηριστικών, το οποίο να αντιπροσωπεύει μια ολόκληρη ονοματική φράση.

Η λύσεις που υιοθετήθηκαν διαφέρουν ως ένα σημείο για τους δύο αλγορίθμους, λόγω του διαφορετικού τύπου εισόδου που απαιτούν. Εντούτοις είναι βασισμένες στο ίδιο είδος πληροφορίας. Στο πλαίσιο αυτής της προσέγγισης, επιλέξαμε την κωδικοποίηση κάθε μεμονωμένης λέξης με δύο χαρακτηριστικά: το μέρος του λόγου της λέξης, και μιας ετικέτας που προέρχεται από το λεξικό, εάν τέτοια πληροφορία είναι διαθέσιμη. Αξίζει να σημειωθεί ότι η ίδια η λέξη ή μέρος αυτής (λήμμα, θέμα, κλπ) δεν συμπεριλαμβάνεται στο διάνυσμα σαν χαρακτηριστικό.

Οι αλγόριθμοι εκμάθησης απαιτούν επίσης όλα τα διανύσματα χαρακτηριστικών για να είναι ενός σταθερού, καθορισμένου μήκους. Προκειμένου να επιτευχθεί αυτή η απαίτηση, περιορίσαμε το μήκος των ονοματικών φράσεων σε 10 λέξεις, δηλ. φράσεις με αριθμό λέξεων μεγαλύτερο από 10 θεωρούνται ότι δεν είναι ονόματα οντοτήτων. Ταυτόχρονα, κάθε διάνυσμα περιλαμβάνει και μερικές λέξεις από το περιβάλλον της ονοματικής φράσης. Αυτή η πληροφορία συμφραζομένων ορίστηκε στις δύο λέξεις δεξιά και αριστερά από την υπό εξέταση ονοματική φράση.

### 4.6.2 Σύστημα NERC βασισμένο σε δέντρα αποφάσεων

Το πρώτο σύστημα αναγνώρισης ονομάτων οντοτήτων είναι βασισμένο σε έναν γενικής χρήσης συμβολικό αλγόριθμο μηχανικής μάθησης, γνωστό με την ονομασία C4.5. Ο C4.5 είναι ένας επιβλεπόμενος αλγόριθμος μάθησης, που εκτελεί επαγωγική εξαγωγή των δέντρων απόφασης, δηλ., κατασκευάζει δέντρα απόφασης από δεδομένα εκπαίδευσης. Ο C4.5 χρησιμοποιεί μια *εξαντλητική (greedy)* αναζήτηση *«αναρρίχησης λόφου» (hill-climbing)* στο χώρο των πιθανών δέντρων απόφασης, στοχεύοντας να κατασκευάσει ένα που εξηγεί τα δεδομένα καλά. Εκτελεί αυτήν την αναζήτηση με τη μέθοδο του *αναδρομικού χωρισμού (recursive partitioning)* των δεδομένων εκπαίδευσης: αρχίζοντας με το πλήρες σύνολο δεδομένων, επιλέγει ένα χαρακτηριστικό που διαχωρίζει καλύτερα τα παραδείγματα (διανύσματα χαρακτηριστικών) στις διάφορες κατηγορίες, όπως π.χ. οργανισμούς, πρόσωπα ή τοποθεσίες. Η ποιότητα του διαχωρισμού αξιολογείται από μια μετρική, βασισμένη στην *αμοιβαία πληροφορία (mutual information)* [79]. Η ίδια διαδικασία εφαρμόζεται κατ' επανάληψη σε κάθε υποσύνολο, όπου επιλέγεται ένα διαφορετικό χαρακτηριστικό για την περαιτέρω

κατάτμηση των δεδομένων εκπαίδευσης. Αυτός ο συνεχής διαχωρισμός οδηγεί σε όλο και περισσότερο «καθαρότερα» υποσύνολα, δηλ., σύνολα που περιέχουν πολλά παραδείγματα μιας κατηγορίας, π.χ. πρόσωπο, αλλά λίγα από άλλες κατηγορίες. Η διαδικασία ολοκληρώνεται όταν ικανοποιείται ένα κριτήριο τερματισμού. Στην απλούστερη περίπτωση, αυτό το κριτήριο απαιτεί απολύτως καθαρά υποσύνολα, όπου κάθε φύλλο του δέντρου περιέχει μόνο παραδείγματα μιας κατηγορίας. Όμως, αυτό το κριτήριο είναι μη ρεαλιστικό για πραγματικά προβλήματα και οδηγεί στην *απομνημόνευση* των δεδομένων εκπαίδευσης από το δέντρου απόφασης *(overfitting)*. Προκειμένου να αποφευχθεί αυτό το πρόβλημα, ο C4.5 ενσωματώνει μια μέθοδο *περικοπής (pruning)*, η οποία κατασκευάζει ένα πιο γενικό δέντρο απόφασης, επιτρέποντας λίγα «υπολείμματα» στα υποσύνολα των φύλλων.

### 4.6.3 Σύστημα NERC βασισμένο σε νευρωνικά δίκτυα

Το δεύτερο σύστημα αναγνώρισης ονομάτων οντοτήτων είναι βασισμένο σε ένα νευρωνικό δίκτυο. Τα *τεχνητά νευρωνικά δίκτυα (artificial neural networks – ANNs)* είναι υπολογιστικά συστήματα η αρχιτεκτονική και η λειτουργία των οποίων είναι βασισμένη στην τρέχουσα γνώση μας για τα βιολογικά νευρικά συστήματα. Αναλογικά με αυτά τα συστήματα, τα ANNs αποτελούνται από ένα σύνολο κατάλληλα τοποθετημένων απλών στοιχείων επεξεργασίας (κόμβων), που αντιπροσωπεύουν τους νευρώνες. Κάθε κόμβος λαμβάνει τα σήματα από ένα σταθερό σύνολο άλλων κόμβων, μέσω συνδέσεων που αποκαλούνται συνάψεις, και καθορίζει την ενεργοποίησή του σαν συνάρτηση αυτών των σημάτων και της δύναμης (βάρος) των συνάψεων. Η απάντηση κάθε κόμβου είναι συνάρτηση των ερεθισμάτων του. Διαφορετικά μοντέλα ANN μπορούν να προκύψουν ανάλογα με την διασύνδεση των στοιχείων επεξεργασίας.

Ο τύπος ANN που χρησιμοποιείται σε αυτήν την προσέγγιση είναι το πολυστρωματικό feed-forward δίκτυο (FNN), το οποίο αποτελείται από ένα επίπεδο εισόδου, ένα ενδιάμεσο ή «κρυμμένο» επίπεδο από κόμβους και ένα επίπεδο από τους κόμβους εξόδου, με κάθε κόμβο που λαμβάνει είσοδο μόνο από κόμβους του προηγούμενου επιπέδου. Ένα σημαντικό θεωρητικό αποτέλεσμα σχετικά με τα FNN, είναι η θεωρητική καθιέρωση ότι αυτά τα δίκτυα μπορούν να *προσεγγίσουν οποιαδήποτε συνάρτηση (global approximators)* [80]. Δεδομένου ότι ένας ικανοποιητικός αριθμός κρυμμένων κόμβων συμπεριλαμβάνεται στη αρχιτεκτονική του δικτύου, υπάρχει ένα σύνολο από βάρη συνάψεων, με το οποίο ένα FNN μπορεί να προσεγγίσει μια αυθαίρετα περίπλοκη, μη γραμμική συνάρτηση μεταξύ των εισόδων και των εξόδων, μέσω του συνδυασμού στοιχειωδών συναρτήσεων εντός των κόμβων.

Ένα σημαντικό θέμα στην ερευνητική περιοχή των νευρωνικών δικτύων είναι η μεθοδολογία εύρεσης των βαρών στις συνάψεις. Ένας δημοφιλής αλγόριθμος εκπαίδευσης είναι ο αλγόριθμος «back-propagation», ο οποίος ελαχιστοποιεί το σφάλμα των κόμβων εξόδου σε σχέση με την επιθυμητή συνάρτηση, χρησιμοποιώντας *κλιμακωτή κάθοδο (gradient descent)*. Πρόσφατα, μια νέα οικογένεια αλγορίθμων έχει προταθεί (*Algorithms for Learning Efficiently using Constrained Optimisation – ALECO*) [81]. Αυτοί οι αλγόριθμοι είναι βασισμένοι στις αρχές της μη γραμμικής θεωρίας προγραμματισμού (*principles of non-linear programming theory*) και ενσωματώνουν στο φορμαλισμό τους πρόσθετες πληροφορίες σχετικά με τις ιδιότητες μάθησης των FNNs, υπό τη μορφή μη γραμμικών περιορισμών. Η παραλλαγή που χρησιμοποιήθηκε περιγράφεται λεπτομερώς στην εργασία [82]. Ο αλγόριθμος είναι γρηγορότερος από την *ανάστροφη μετάδοση (back propagation)* και τις παραλλαγές του, ενώ εμφανίζει καλές ιδιότητες σύγκλισης σε δύσκολα προβλήματα συγκριτικής μέτρησης επιδόσεων. Ταυτόχρονα, είναι καλά ταιριαγμένος για την επίλυση των προβλημάτων μεγάλης κλίμακας και έχει αποδειχθεί [82] ότι επιτυγχάνει ικανοποιητική γενίκευση σε προβλήματα ταξινόμησης.

### 4.6.4 Συμβολικό διάνυσμα χαρακτηριστικών

Για το συμβολικό αλγόριθμο μάθησης, η κωδικοποίηση των διανυσμάτων χαρακτηριστικών ήταν προφανής. Το διάνυσμα χαρακτηριστικών αποτελείται από 26 χαρακτηριστικά (δύο χαρακτηριστικά για κάθε μια από τις εννέα λέξεις μιας ονοματικής φράσης, συν τα οχτώ χαρακτηριστικά των τεσσάρων λέξεων πριν και μετά την φράση). Στην περίπτωση όπου μια φράση είναι κοντύτερη από εννέα λέξεις, στα μη χρησιμοποιούμενα χαρακτηριστικά δίδεται μια ειδική τιμή (“?”), η οποία μεταχειρίζεται ως ετικέτα έλλειψης τιμής από τον αλγόριθμο C4.5. Η έλλειψη πληροφορίας από το λεξικό αντιμετωπίζεται διαφορετικά: αποδίδεται μια πρόσθετη ετικέτα (“NOTAG”) στις άγνωστες στο λεξικό λέξεις. Σαν παράδειγμα του τρόπου με τον οποίο οι φράσεις κωδικοποιούνται σε διανύσματα χαρακτηριστικών, εξετάστε την ακόλουθη φράση:

… of the ***Securities and Exchange Commission*** in the …

όπου η ονοματική φράση (που σε αυτή την περίπτωση αντιπροσωπεύει έναν οργανισμό) παρουσιάζεται με πλάγια, έντονη γραφή. Το διάνυσμα που αντιστοιχεί σε αυτήν την φράση, είναι το ακόλουθο:

```
[IN, NOTAG, DT, NOTAG, NNP, org_key+organisation, CC,
organisation, NNP, organisation, NNP, org_base+organisation,
?, ?, ?, ?, ?, ?, ?, ?, ?, ?, IN, NOTAG, DT, NOTAG]
```

όπου οι ετικέτες μέρους του λόγου πρέπει να ερμηνευθούν ως εξής:

- `IN`: πρόθεση, `DT`: άρθρο, `NNP`: κύριο όνομα ουσιαστικό, `CC`: σύνδεσμος.

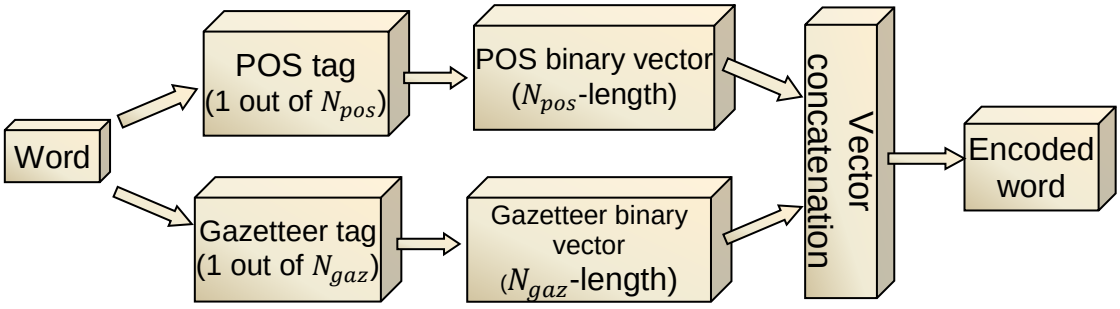
Οι ετικέτες του λεξικού που εμφανίζονται στο παραπάνω παράδειγμα είναι: `organisation`, `org_key`, `org_base` και `NOTAG`. Η φράση «*Securities and Exchange Commission*» εμφανίζεται στον κατάλογο οργανισμών του λεξικού, και συνεπώς όλες οι λέξεις που την απαρτίζουν χαρακτηρίζονται με την ετικέτα `organisation`. Αξίζει να σημειωθεί ότι περισσότερες από μια ετικέτες του λεξικού μπορούν να αποδοθούν σε μια λέξη, σηματοδοτώντας την παρουσία της λέξης σε περισσότερους από έναν καταλόγους του λεξικού, όπως στην περίπτωση της λέξης «*Securities*», η οποία είναι και `org_key` και μέρος ενός `organisation`. Οι πολλαπλές ετικέτες συνενώνονται με το σύμβολο «+» στον φορμαλισμό που χρησιμοποιούμε. Τέλος, τα σύμβολα «?» συμβολίζουν λέξεις που δεν υπάρχουν, όπως έχει ήδη εξηγηθεί.

### 4.6.5 Αριθμητικό διάνυσμα χαρακτηριστικών

Οι γλωσσική πληροφορία κωδικοποιείται φυσικότερα με συμβολική μορφή, σε αντίστοιχες αναπαραστάσεις. Πράγματι, τόσο τα μέρη του λόγου όσο και οι ετικέτες του λεξικού δεν έχουν καμία αριθμητική ιδιότητα: για παράδειγμα, δεν υπάρχει κανένας απλός τρόπος να διαταχτεί ένα ρήμα και ένα επίθετο. Από την άλλη, τα νευρωνικά δίκτυα δέχονται σαν είσοδο ένα μονοδιάστατο, καθορισμένου μήκους διάνυσμα, με πραγματικούς αριθμούς σαν τιμές των χαρακτηριστικών. Συνεπώς, μια κατάλληλη κωδικοποίηση των πληροφοριών των ετικετών σε ένα αριθμητικό διάνυσμα δεν είναι τόσο απλή όσο σε ένα συμβολικό. Εντούτοις, η επιλογή της αναπαράστασης είναι σημαντική για την απόδοση του νευρωνικού δικτύου, συνεπώς δόθηκε ιδιαίτερη έμφαση στην δημιουργία μιας αναπαράστασης με νόημα.

Για την διατήρηση της συμβολικής πληροφορίας όσο το δυνατόν περισσότερο, χωρίς την εισαγωγή οποιονδήποτε παραδοχών σχετικά με τις ετικέτες, υποθέτουμε ότι οι ετικέτες είναι ορθογώνιες *η μια στην άλλη*. Κατά συνέπεια, για να κωδικοποιήσουμε μια λέξη σε διανυσματική μορφή, προχωρήσαμε με τον ακόλουθο τρόπο:

- Ας θεωρήσουμε ότι $N_{pos}$ είναι ο αριθμός των πιθανών διαφορετικών ετικετών των μέρων του λόγου. Υποθέτοντας ότι οι ετικέτες αυτές είναι ανεξάρτητες, μπορούμε να διαμορφώσουμε ένα χώρο «μερών του λόγου», $N_{pos}$ διαστάσεων, με μια διάσταση για κάθε μέρος του λόγου. Ως εκ τούτου, κάθε ετικέτα μέρους του λόγου μπορεί να κωδικοποιηθεί ως ένα διάνυσμα μήκους $N_{pos}$, με όλες τις διαστάσεις καθορισμένες στο 0, εκτός από την μία που αντιστοιχεί στην συγκεκριμένη ετικέτα μέρους του λόγου που ανήκει η λέξη, η οποία λαμβάνει την τιμή 1. Επιπλέον, το σύνολο με τα $N_{pos}$ διανύσματα που κωδικοποιούν τις ετικέτες μερών του λόγου, μπορεί να θεωρηθεί ως ορθογώνια βάση για τον χώρο των «μερών του λόγου».
- Η ίδια θεώρηση μπορεί να μεταφερθεί και στην περίπτωση των ετικετών του λεξικού. Ως εκ τούτου, $\mathrm{N_{gaz}}$ διανύσματα, διάστασης $\mathrm{N_{gaz}}$, ορθογώνια το ένα στο άλλο, μπορούν να αναπαραστήσουν $\mathrm{N_{gaz}}$ ανεξάρτητες ετικέτες λεξικού.
- Μια μεμονωμένη λέξη μπορεί να αναπαρασταθεί με το συνδυασμό των διανυσμάτων κωδικοποίησης των χαρακτηριστικών της. Το πλήρες διάνυσμα διαμορφώνεται από την αλληλουχία ενός διανύσματος που κωδικοποιεί το μέρος του λόγου, και ενός διανύσματος που κωδικοποιεί τις κατηγορίες του λεξικού. Ως εκ τούτου, κάθε λέξη αντιπροσωπεύεται από ένα διάνυσμα $N$ διαστάσεων, όπου $N = N_{pos} + N_{gaz}$, το οποίο μπορεί να θεωρηθεί ως ένα σημείο σε έναν χώρο διάσταση $N$.

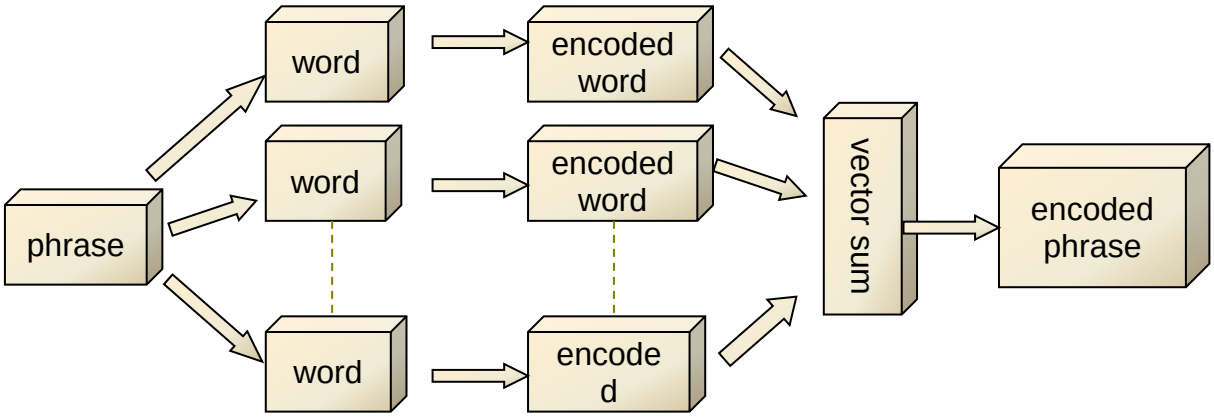



**Εικόνα 6: Κωδικοποίηση μιας λέξης σε αριθμητικό διάνυσμα.**

Για την συγκεκριμένη εργασία κατηγοριοποίησης, περισσότερες από μια λέξεις θα δίνονται σαν είσοδο στον ταξινομητή. Για να αναπαρασταθεί μια ομάδα λέξεων $W$, κάποιος θα μπορούσε εύκολα να σκεφτεί την διαδοχική σύνδεση (αλληλουχία) των κωδικοποιημένων $W$ λέξεων, καταλήγοντας σε ένα διάνυσμα με $W \cdot N$ διαστάσεις. Ωστόσο, μια τέτοια αναπαράσταση έχει δύο μειονεκτήματα:

- Το συνολικό μήκος του διανύσματος τείνει να γίνει πολύ μεγάλο. Στην περίπτωσή μας, θα ήταν περίπου 650 χαρακτηριστικά, αν υποθέσουμε 3 κατηγορίες εξόδου (οι οποίες πρέπει να κωδικοποιηθούν και αυτές, καταλήγοντας σε μια διάσταση $W \cdot N \cdot 3$).
- Στις περισσότερες περιπτώσεις, δεν υπάρχουν όλες οι λέξεις. Πράγματι, οι ονοματικές φράσεις μπορούν να έχουν έναν ποικίλο αριθμό λέξεων, ωστόσο δεν μπορούμε να χρησιμοποιήσουμε διανύσματα ποικίλου μήκους. Η κωδικοποίηση χαρακτηριστικών που λείπουν δεν είναι τόσο εύκολη χωρίς περεταίρω αύξηση της διαστατικότητας. Ταυτόχρονα, η χρήση μιας σταθερής, «ουδέτερης» τιμής, για παράδειγμα 0.5, μπορεί να είναι μια λύση σε αυτό το πρόβλημα. Εντούτοις, θεωρούμε ότι αυτή η λύση είναι αμφίβολη, δεδομένου ότι μπορεί να ερμηνευτεί σαν μια λέξη που περιέχει «λίγο από όλα».

Για να παρακάμψουμε αυτά τα μειονεκτήματα, αντί της *διαδοχικής σύνδεσης* των διανυσμάτων των λέξεων, τα *προσθέτουμε*. Ως εκ τούτου, κάθε ομάδα λέξεων $W$ αντιπροσωπεύεται από ένα διάνυσμα το οποίο παραμένει μήκους $N$, το οποίο είναι και μικρότερο από ότι στην περίπτωση της σύνδεσης, και σταθερό. Φυσικά, η πρόσθεση των διανυσμάτων οδηγεί σε διανύσματα, με τις τιμές χαρακτηριστικών μεγαλύτερες από 1. Για παράδειγμα, εάν μια ονοματική φράση έχει μεταξύ άλλων, δύο λέξεις που έχουν κατηγοριοποιηθεί σαν επίθετα, η διάσταση που αντιστοιχεί στα επίθετα θα λάβει την τιμή 2. Αξίζει να σημειωθεί ότι με την πρόσθεση των διανυσμάτων, αντί για την σύνδεσή τους, *η πληροφορία σχετικά με την σειρά των λέξεων χάνεται*. Αυτή είναι μια σημαντική διαφορά μεταξύ των δύο αναπαραστάσεων. Ωστόσο, θεωρήσαμε ότι η απώλεια της σειράς εντός της ονοματικής φράσης είναι ίσως αποδεκτή, αλλά θα ήταν περισσότερο φρόνιμο να υπάρξει τουλάχιστον μια διάκριση μεταξύ των λέξεων εντός της φράσης και των λέξεων του περιβάλλοντος της φράσεως. Συνεπώς, το τελικό διάνυσμα είναι μια αλληλουχία τριών διανυσμάτων με μήκος $N$, ένα για κάθε ομάδα λέξεων (πριν, εντός, και μετά την ονοματική φράση). Χρησιμοποιώντας αυτήν την μέθοδο, το διάνυσμα έχει ένα σταθερό μήκος περίπου 150 διαστάσεων. Η διανυσματική διαστατικότητα είναι ακόμα αρκετά μεγάλη, το οποίο είναι αναντίρρητα ένα μειονέκτημα για την εκπαίδευση ενός νευρωνικού δικτύου. Εντούτοις, είναι ένας λογικός συμβιβασμός μεταξύ μιας λογικής αναπαράστασης και μιας καθορισμένου μήκους αριθμητικής διανυσματικής μορφής.




**Εικόνα 7: Κωδικοποίηση μιας φράσης σε αριθμητικό διάνυσμα.**

### 4.6.6 Πειραματική αξιολόγηση και αποτελέσματα

Έχοντας κωδικοποιήσει τις προσδιορισμένες ονοματικές φράσεις στις δύο διαφορετικές αναπαραστάσεις, πραγματοποιήθηκαν δύο τύποι πειραμάτων. Ο σκοπός των δύο πειραμάτων ήταν να εξεταστεί η συμπεριφορά κάθε μεθόδου στην εργασία της αναγνώρισης ονομάτων οντοτήτων *(NERC)* συνολικά, καθώς επίσης και σε κάθε μια από τις δευτερεύουσες εργασίες του: αναγνώριση και κατηγοριοποίηση ονομάτων οντοτήτων. Σαν πρώτο πείραμα, το NERC αντιμετωπίστηκε σαν κατηγοριοποίηση σε τρεις κατηγορίες. Οι κατηγορίες αυτές είναι: πρόσωπο, οργανισμός και όχι-οντότητα. Κατά συνέπεια, κάθε αλγόριθμος καλείται να εκτελέσει και τις δύο εργασίες του NERC ταυτόχρονα, δηλ., του προσδιορισμού και της ταξινόμησης των φράσεων σε οντότητες. Το δεύτερο πείραμα διαιρείται σε δύο στάδια. Στο πρώτο στάδιο, χρησιμοποιούνται μόνο δύο κατηγορίες: όνομα οντότητας (NE) και όχι-NE. Στο δεύτερο στάδιο, μόνο οι ονοματικές φράσεις που είναι NE χρησιμοποιούνται σαν δεδομένα εκπαίδευσης, ενώ οι κατηγορίες που ζητούνται είναι δύο: πρόσωπο και οργανισμός. Σε κάθε πείραμα, χρησιμοποιήθηκε *δεκαπλή διασταυρωμένη επικύρωση (10-fold cross validation)*, προκειμένου να εξαχθεί μια αμερόληπτη εκτίμηση της απόδοσης του συστήματος σε νέα δεδομένα.

Οι μετρικές που επιλέχτηκαν για την αξιολόγηση της απόδοσης, είναι εκείνα που χρησιμοποιούνται τυπικά για την αξιολόγηση συστημάτων γλωσσικής τεχνολογίας: η ανάκληση και η ακρίβεια. Τέλος, σαν βάση για τη σύγκριση των αποτελεσμάτων μπορούμε να χρησιμοποιήσουμε την απόδοση συστημάτων χειρωνακτικά κατασκευασμένων. Τα αποτελέσματα των συστημάτων MITOS [16], [83] (το οποίο αποτελεί βελτίωση του συστήματος που παρουσιάζεται στο ΠΑΡΑΡΤΗΜΑ ΙΙ με μεγαλύτερο λεξικό και περισσότερους γραμματικούς κανόνες) και VIE [84] NERC, όπως έχουν αξιολογηθεί στα σώματα κειμένων της Καπα TEL[2] και του συνεδρίου MUC-6 αντίστοιχα, παρουσιάζονται στον πίνακα [83]: Πίνακας 14.

| | **MITOS NERC System** | | **VIE NERC System** | |
|---|---|---|---|---|
| | ***Πρόσωπα*** | ***Οργανισμοί*** | ***Πρόσωπα*** | ***Οργανισμοί*** |
| ***Ανάκληση*** | 76.50 % | 84.20 % | 84.97 % | 69.25 % |
| ***Ακρίβεια*** | 87.50 % | 89.80 % | 92.50 % | 83.42 % |
| ***F-Measure*** | 81.63 % | 86.91 % | 88.57 % | 75.68 % |

**Πίνακας 14: Η απόδοση των χειρωνακτικά κατασκευασμένων συστημάτων αναγνώρισης ονομάτων οντοτήτων.**

#### Πείραμα 1: Κατηγοριοποίηση σε: πρόσωπο – οργανισμό – όχι-NE

Στο πρώτο πείραμα οι δύο αλγόριθμοι μηχανικής καλούνται να αξιολογηθούν στην εργασία του NERC, περιλαμβάνοντας τόσο την αναγνώριση όσο και την κατηγοριοποίηση των ονομάτων οντοτήτων. Οι κατηγορίες που καλούνται να εντάξουν τις ονοματικές φράσεις είναι τρεις: πρόσωπο, οργανισμός και όχι-NE, δεδομένου ότι θελήσαμε να εξετάσουμε τη συμπεριφορά κάθε αλγορίθμου εκμάθησης και στα δύο μέρη του στόχου NERC ταυτόχρονα, δηλ., ο προσδιορισμός και η ταξινόμηση των

[2] Καπα TEL: http://www.kapatel.gr/

φράσεων σε NE. Τα αποτελέσματα και για τους δύο αλγορίθμους παρουσιάζονται στον ακόλουθο πίνακα (Πίνακας 15):

| | **Ελληνικά** | | **Αγγλικά** | |
|---|---|---|---|---|
| | ***Πρόσωπα*** | ***Οργανισμοί*** | ***Πρόσωπα*** | ***Οργανισμοί*** |
| ***DT Ανάκληση*** | 59.68 % | 72.69 % | 90.40 % | 87.85 % |
| ***NN Ανάκληση*** | 60.29 % | 75.67 % | 90.61 % | 86.21 % |
| ***DT-N Ανάκληση*** | 59.86 % | 80.30 % | 89.35 % | 86.98 % |
| ***DT Ακρίβεια*** | 80.45 % | 77.60 % | 93.23 % | 80.43 % |
| ***NN Ακρίβεια*** | 79.36 % | 79.00 % | 94.73 % | 83.33 % |
| ***DT-N Ακρίβεια*** | 83.79 % | 79.10 % | 93.05 % | 86.43 % |
| ***DT F-Measure*** | 68.53 % | 75.06 % | 91.79 % | 83.98 % |
| ***NN F-Measure*** | 68.52 % | 77.30 % | 92.62 % | 84.75 % |
| ***DT-N F-Measure*** | 69.83 % | 79.70 % | 91.16 % | 86.70 % |

**Πίνακας 15: Αποτελέσματα για το πείραμα με τις τρεις κατηγορίες (πρόσωπο – οργανισμός – όχι-NE). (DT: δέντρο απόφασης με τη συμβολική αντιπροσώπευση (ενότητα 4.6.4), NN: νευρικό δίκτυο με την αριθμητική αντιπροσώπευση (ενότητα 4.6.5), DT-N: δέντρο απόφασης με την αριθμητική αντιπροσώπευση).**

Όσον αφορά τα πειράματα για την ελληνική γλώσσα, συγκρίνοντας τα αποτελέσματα των εκπαιδευμένων ταξινομητών (Πίνακας 15) με τα αποτελέσματα του χειρωνακτικά κατασκευασμένου συστήματος NERC που παρουσιάζεται στον πίνακα (Πίνακας 14), μπορούμε να καταλήξουμε στο συμπέρασμα ότι και τα δύο συστήματα μηχανικής μάθησης αποδίδουν χειρότερα από το χειρωνακτικά κατασκευασμένο σύστημα. Αυτή η αποτυχία αποδίδεται πρώτιστα στη γλωσσική φάση προ-επεξεργασίας και κυρίως στο τμήμα προσδιορισμού των ονοματικών φράσεων. Ο χρησιμοποιούμενος *αναγνωριστής ονοματικών φράσεων (NP chunker)* δεν παρέχει πληροφορίες σχετικά με ονοματικές φράσεις που μπορούν να περιέχονται μέσα σε μια αναγνωρισμένη, μεγαλύτερη ονοματική φράση. Συνεπώς, στα δεδομένα εκπαίδευσης υπάρχουν μόνο πληροφορίες για τις πιο σύνθετες και μεγάλες σε αριθμό λέξεων ονοματικές φράσεις, και όχι για τις ονοματικές φράσεις που εσωκλείονται μέσα σε αυτές τις αναγνωρισμένος ονοματικές φράσεις, όπως συμβαίνει με τον Αγγλικό αναγνωριστή ονοματικών φράσεων. Επιπλέον, το γεγονός ότι το ελληνικό σώμα κειμένων περιείχε μερικά ορθογραφικά λάθη (συμπεριλαμβανομένων Αγγλικών χαρακτήρων που χρησιμοποιούνται ως Ελληνικοί χαρακτήρες και αντίστροφα) έχει μειώσει την απόδοση του αναγνωριστή μερών του λόγου, καθώς επίσης και την απόδοση του λεξικού.

Συγκρίνοντας την απόδοση των δύο αλγορίθμων μηχανικής μάθησης με τα αποτελέσματα του χειρωνακτικού συστήματος VIE, μπορούμε να συμπεράνουμε ότι τόσο η ανάκληση όσο και η ακρίβεια για τα πρόσωπα και τους οργανισμούς είναι σε υψηλότερα επίπεδα από το σύστημα VIE. Η βελτίωση στην ανάκληση για τα πρόσωπα και για τους οργανισμούς είναι μεγαλύτερες από τις αντίστοιχες αυξήσεις στην ακρίβεια, που καταδεικνύει ότι τα συστήματα μάθησης είναι σε θέση να προσδιορίσουν και να ταξινομήσουν σωστά περισσότερα ονόματα οντοτήτων από το χειρωνακτικά κατασκευασμένο σύστημα. Η μόνη εξαίρεση είναι η ακρίβεια (*DT Precision*) για τους οργανισμούς του συστήματος δέντρων απόφασης, όπου η αύξηση στην ανάκληση συνοδεύεται με μια μικρή πτώση στην ακρίβεια.

Εάν συγκρίνουμε την απόδοση του δέντρου απόφασης με την απόδοση του νευρωνικού δικτύου, μπορούμε να συμπεράνουμε ότι και τα δύο συστήματα επιτυγχάνουν συγκρίσιμη απόδοση, με το νευρωνικό δίκτυο να υπερέχει ελαφρώς από το δέντρο

απόφασης. Ωστόσο, η πιο ενδιαφέρουσα ιδιότητα του νευρωνικού συστήματος NERC, είναι ότι πέτυχε υψηλότερη επίδοση από το σύστημα δέντρου απόφασης, έχοντας *λιγότερη πληροφορία σαν είσοδο*, καθώς δεν διατηρείται η σειρά των λέξεων. Όπως αναλύθηκε στην ενότητα 4.6.5, η πληροφορία σχετικά με την σειρά των λέξεων που δίνεται ως δεδομένα εισόδου στον αλγόριθμο C4.5, χάνεται κατά τη διάρκεια της κατασκευής των διανυσμάτων εκπαίδευσης για το νευρωνικό δίκτυο. Αυτό είναι ένα απροσδόκητο αποτέλεσμα, δεδομένου ότι η αρχική πεποίθησή μας ήταν ότι η σειρά των λέξεων είναι σημαντική για την αναγνώριση ονομάτων οντοτήτων. Προκειμένου να εξετάσουμε περαιτέρω εάν αυτή η βελτιωμένη απόδοση οφείλεται στον αλγόριθμο μηχανικής μάθησης ή στην χρησιμοποιούμενη αναπαράσταση των δεδομένων, διενεργήσαμε πρόσθετα πειράματα. Σε αυτά τα πειράματα, επανεκπαιδεύσαμε τον C4.5 με τα ίδια διανύσματα που χρησιμοποιήθηκαν για την εκπαίδευση του νευρωνικού δικτύου (αριθμητική αναπαράσταση). Από τα αποτελέσματα (που παρουσιάζονται επίσης στον πίνακα (Πίνακας 15) – ομάδα αποτελεσμάτων DT-N) είναι εμφανές ότι η απόδοση του C4.5 έχει βελτιωθεί. Πιο συγκεκριμένα, η ανάκληση για τα πρόσωπα και τους οργανισμούς, καθώς επίσης και η ακρίβεια για τα πρόσωπα έχουν μειωθεί ελαφρώς (λιγότερο από 1% πτώση για κάθε μετρική) αλλά η ακρίβεια για τους οργανισμούς έχει αυξηθεί σημαντικά (~ 6%). Αυτό είναι ένα ενδιαφέρον αποτέλεσμα, δεδομένου ότι τα δέντρα απόφασης σχεδιάζονται πρώτιστα προκειμένου να χρησιμοποιηθούν με συμβολικά χαρακτηριστικά και όχι αριθμητικά, αλλά και απροσδόκητο, δεδομένου ότι η αριθμητική αναπαράσταση δεν περιέχει πληροφορίες σχετικά με τη σειρά των λέξεων. Αυτά τα αποτελέσματα δείχνουν ότι αυτή η μείωση πληροφορίας απλοποίησε το στόχο εκμάθησης, βελτιώνοντας κατά συνέπεια την απόδοση της ταξινόμησης για το συγκεκριμένο πρόβλημα.

### Πείραμα 2: Στάδιο 1: Κατηγοριοποίηση σε: NE – όχι-NE

Στο πρώτο στάδιο του δεύτερου πειράματος, θελήσαμε να εξετάσουμε τη συμπεριφορά κάθε αλγορίθμου μάθησης στον προσδιορισμό των φράσεων ονομάτων οντοτήτων. Τα αποτελέσματα παρουσιάζονται στον πίνακα Πίνακας 16 και για το τους δύο αναγνωριστές μηχανικής μάθησης.

| | Greek corpus | English corpus |
|---|---|---|
| ***DT Ανάκληση*** | 75.36 % | 94.31 % |
| ***NN Ανάκληση*** | 78.37 % | 96.14 % |
| ***DT-N Ανάκληση*** | 82.23 % | 93.53 % |
| ***DT Ακρίβεια*** | 79.34 % | 91.26 % |
| ***NN Ακρίβεια*** | 81.69 % | 97.40 % |
| ***DT-N Ακρίβεια*** | 81.94 % | 93.60 % |
| ***DT F-Measure*** | 77.30 % | 92.76 % |
| ***NN F-Measure*** | 79.99 % | 96.76 % |
| ***DT-N F-Measure*** | 82.08 % | 93.56 % |

**Πίνακας 16: Αποτελέσματα για το πείραμα με τις δύο κατηγορίες (NE – όχι-NE). (DT: δέντρο απόφασης με τη συμβολική αντιπροσώπευση (ενότητα 4.6.4), NN: νευρικό δίκτυο με την αριθμητική αντιπροσώπευση (ενότητα 4.6.5), DT-N: δέντρο απόφασης με την αριθμητική αντιπροσώπευση).**

Από τα αποτελέσματα που παρουσιάζονται στον πίνακα (Πίνακας 16), είναι εμφανές ότι το αναγνωριστής ονομάτων οντοτήτων νευρωνικού δικτύου αποδίδει καλύτερα από τον αναγνωριστή ονομάτων οντοτήτων του δέντρου απόφασης. Αν και οι δύο αλγόριθμοι επιτυγχάνουν συγκρίσιμη ανάκληση, το νευρωνικό σύστημα επιτυγχάνει υψηλότερη ακρίβεια, η οποία πλησιάζει το 97.5 %. Αυτό είναι μια σημαντική βελτίωση από το 91.3 % που επιτυγχάνεται από το δέντρο απόφασης. Ωστόσο, η εξέταση της απόδοσης και των δύο μεθόδων ακριβώς στα ίδια δεδομένα, οδηγεί στο συμπέρασμα ότι η απόδοση των δύο μεθόδων είναι παρόμοια.

### Πείραμα 2: Στάδιο 2: Κατηγοριοποίηση σε: πρόσωπο – οργανισμό

Τέλος, στο δεύτερο στάδιο του δεύτερου πειράματος θελήσαμε να αξιολογήσουμε τους δύο αλγορίθμους στην κατηγοριοποίηση NE. Τα αποτελέσματα παρουσιάζονται στον ακόλουθο πίνακα (Πίνακας 17) και για τους δύο ταξινομητές.

Όπως συνέβη και στο προηγούμενο πείραμα, και σε αυτήν την περίπτωση ο νευρωνικός ταξινομητής ξεπερνά τον ταξινομητή δέντρων απόφασης, αν και οι διαφορές μεταξύ των δύο αλγορίθμων δεν είναι τόσο μεγάλες όπως στο προηγούμενο πείραμα. Αλλά πάλι εάν αξιολογήσουμε τους δύο αλγορίθμους στο ίδιο σχήμα αναπαράστασης, η απόδοσή τους γίνεται ουσιαστικά όμοια.

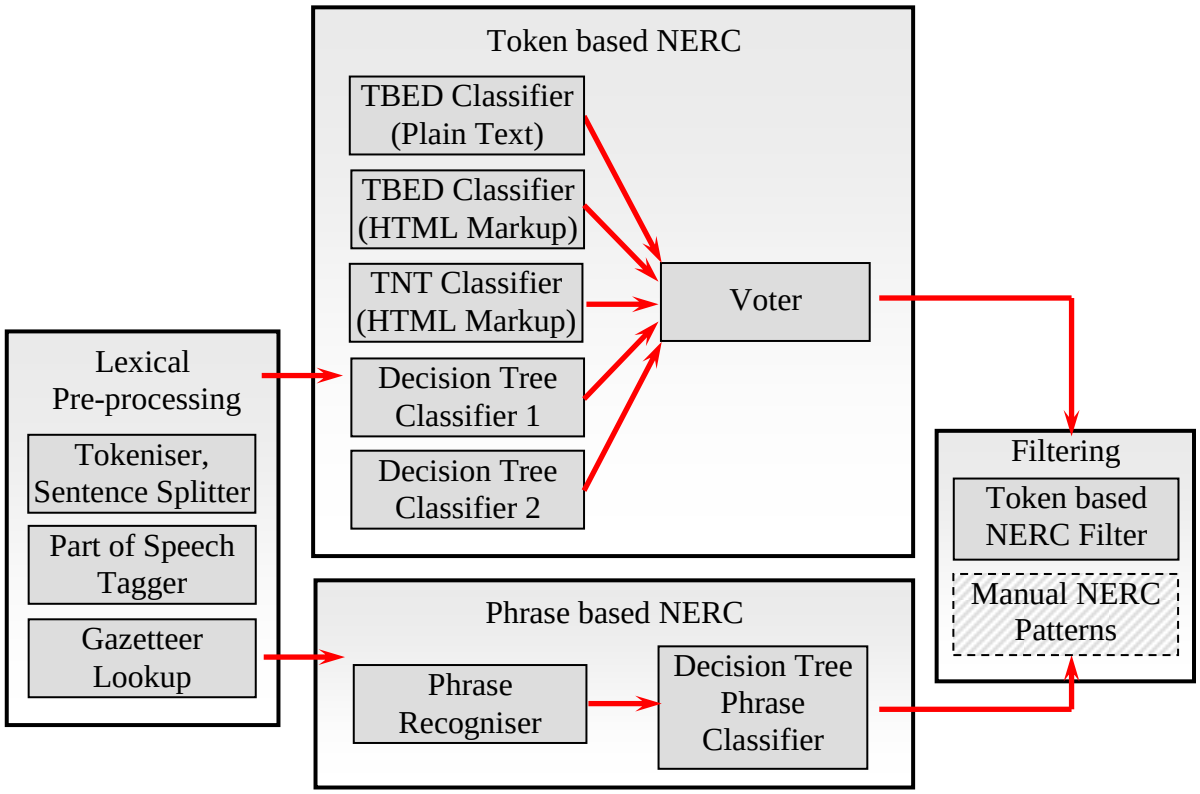
| | **Greek corpus** | | **English corpus** | |
|---|---|---|---|---|
| | ***Πρόσωπα*** | ***Οργανισμοί*** | ***Πρόσωπα*** | ***Οργανισμοί*** |
| ***DT Ανάκληση*** | 69.08 % | 98.99 % | 91.70 % | 97.48 % |
| ***NN Ανάκληση*** | 78.42 % | 98.39 % | 95.17 % | 97.81 % |
| ***DT-N Ανάκληση*** | 74.53 % | 98.54 % | 94.88 % | 96.87 % |
| ***DT Ακρίβεια*** | 89.08 % | 96.42 % | 96.43 % | 93.54 % |
| ***NN Ακρίβεια*** | 85.29 % | 97.45 % | 97.62 % | 95.12 % |
| ***DT-N Ακρίβεια*** | 86.46 % | 97.02 % | 96.16 % | 95.95 % |
| ***DT F-Measure*** | 77.81 % | 97.68 % | 94.00 % | 95.47 % |
| ***NN F-Measure*** | 81.71 % | 97.92 % | 96.38 % | 96.45 % |
| ***DT-N F-Measure*** | 80.05 % | 97.77 % | 95.51 % | 96.41 % |

**Πίνακας 17: Αποτελέσματα για το πείραμα με τις δύο κατηγορίες (πρόσωπο – οργανισμός). (DT: δέντρο απόφασης με τη συμβολική αντιπροσώπευση (ενότητα 4.6.4), NN: νευρικό δίκτυο με την αριθμητική αντιπροσώπευση (ενότητα 4.6.5), DT-N: δέντρο απόφασης με την αριθμητική αντιπροσώπευση).**

## 4.7 Προσέγγιση Β: συνδυασμός συστημάτων κατηγοριοποίησης σε επίπεδο λέξεων και φράσεων

Η δεύτερη προτεινόμενη προσέγγιση *(ML-HNERC)* συνδυάζει δύο διαφορετικούς τύπους πληροφορίας, υλοποιώντας δύο συστήματα κατηγοριοποίησης. Το πρώτο σύστημα κατηγοριοποίησης κατηγοριοποιεί μεμονωμένες λέξεις, χρησιμοποιώντας πληροφορίες από την ίδια την λέξη, αλλά και από ένα μικρό περιβάλλον γειτονικών λέξεων. Το σύστημα κατηγοριοποίησης αυτό χρησιμοποιεί τέσσερις διαφορετικούς αλγορίθμους μηχανικής μάθησης και τέσσερις διαφορετικές αναπαραστάσεις, σε πέντε διαφορετικούς ταξινομητές, οι οποίοι συνεργάζονται μέσω ενός συστήματος ψηφοφορίας, ώστε να προκύψει το τελικό αποτέλεσμα. Το δεύτερο σύστημα κατηγοριοποίησης αποτελείται από ένα σύστημα κατηγοριοποίησης το οποίο

αναγνωρίζει και κατηγοριοποιεί φράσεις (δηλ. σύνολα από διαδοχικές λέξεις), τα οποία είναι ονόματα οντοτήτων. Η αρχιτεκτονική του συστήματος της προτεινόμενης προσέγγισης παρουσιάζεται στην ακόλουθη εικόνα (Εικόνα 8):




**Εικόνα 8: Η αρχιτεκτονική του συστήματος ML-HNERC (δεύτερη προτεινόμενη προσέγγιση).**

Το σύστημα αναγνώρισης ονομάτων οντοτήτων ML-HNERC μπορεί να χωριστεί σε τέσσερα σημαντικά υποσυστήματα. Το πρώτο υποσύστημα είναι αρμόδιο για την εκτέλεση της λεκτικής προ-επεξεργασίας, όπως ο χωρισμός του κειμένου σε λέξεις και προτάσεις, η επισημείωση των λέξεων με μέρη του λόγου, και ο χαρακτηρισμός λέξεων σαν μέρος πιθανού ονόματος οντότητας, μέσω *λιστών γνωστών ονομάτων οντοτήτων (gazetteer lists)*. Την καρδιά του συστήματος ML-HNERC αποτελούν τα δύο υποσυστήματα που είναι επιφορτισμένα με την αναγνώριση ονομάτων οντοτήτων, τόσο σε επίπεδο *λέξης (token-based NERC)*, όσο και σε επίπεδο *φράσης (phrase-based NERC)*. Τέλος, το τελευταίο υποσύστημα φιλτράρει τα αποτελέσματα του υποσυστήματος αναγνώρισης ονομάτων οντοτήτων βασισμένο σε λέξεις, και εφαρμόζει επιπρόσθετους (χειρωνακτικά κατασκευασμένους) κανόνες αναγνώρισης ονομάτων οντοτήτων, αν υπάρχουν διαθέσιμοι. Οι χειρονακτικοί αυτοί κανόνες είναι συνήθως πρότυπα που εντοπίζουν χρονικές εκφράσεις (όπως ημερομηνίες), και εντοπίζουν ονόματα οντοτήτων μόνο αν δεν έχουν ήδη αναγνωριστεί από τα προηγούμενα στάδια.

### 4.7.1 Γλωσσική προ-επεξεργασία

Η γλωσσική προ-επεξεργασία αποτελείται από έναν αναγνωριστή λέξεων και προτάσεων *(tokeniser and sentence splitter)*, έναν αναγνωριστή μερών του λόγου *(part of speech tagger)*, και έναν αναγνωριστή γνωστών ονομάτων οντοτήτων, βασισμένο σε στατικές λίστες ονομάτων οντοτήτων *(gazetteer list lookup)*. Όλοι αυτοί οι αναγνωριστές είναι αρθρώματα της πλατφόρμας επεξεργασίας φυσικής γλώσσας «Έλλογον» [53], με τον αναγνωριστή μερών του λόγου για την Ελληνική γλώσσα να είναι η προσέγγιση που

παρουσιάστηκε στο κεφάλαιο 3. Ο αναγνωριστής γνωστών ονομάτων οντοτήτων είναι υπεύθυνος για τον εντοπισμό και τον χαρακτηρισμό σαν πιθανά ονόματα οντοτήτων, φράσεων που περιέχονται σε λίστες γνωστών ονομάτων οντοτήτων. Οι λίστες αυτές εξαρτώνται τόσο από την θεματική περιοχή, αλλά και από την γλώσσα. Για τις ανάγκες της πειραματικής αποτίμησης που παρουσιάζεται σε αυτή την ενότητα, οι λίστες αυτές εξάχθηκαν από τα χειρωνακτικά επισημειωμένα σώματα κειμένων για την Αγγλική και Ελληνική γλώσσα, κρατώντας μόνο ονόματα οντοτήτων που εμφανίζονται τουλάχιστον τρεις φορές στο κάθε σώμα κειμένου, χαρακτηρισμένο με τον ίδιο τύπο οντότητας (π.χ. όνομα προσώπου, οργανισμού, τοποθεσίας, κλπ.).

### 4.7.2 Αναγνώριση ονομάτων οντοτήτων σε επίπεδο λέξης

Το πρώτο υποσύστημα που είναι επιφορτισμένο με την αναγνώριση ονομάτων οντοτήτων, ο *αναγνωριστής ονομάτων οντοτήτων σε επίπεδο λέξης (token-based NERC)*, επισημειώνει κάθε λέξη του κειμένου με μια *ετικέτα (tag)*, ανάλογα αν η λέξη αποτελεί την πρώτη λέξη ενός ονόματος οντότητας, αν αποτελεί μέρος ενός ονόματος οντότητας, ή αν δεν περιέχεται σε όνομα οντότητας. Αυτός ο τύπος επισημείωσης φράσεων με επισημείωση λέξεων με ετικέτες από τρία σύνολα είναι γνωστός στην βιβλιογραφία σαν $\{I, O, B\}$ [85], ενώ κάθε σύνολο από τα τρία σύνολα ετικετών, περιέχει τόσες ετικέτες όσο και οι τύποι των ονομάτων οντοτήτων. Το συνολικό σύνολο των ετικετών εξαρτάται φυσικά από την θεματική περιοχή, η οποία συνήθως ορίζει και τους τύπους των οντοτήτων που θα αναγνωριστούν. Η θεματική περιοχή που επιλέχθηκε αφορά αγγελίες προσφοράς εργασίας, με το πλήρες σύνολο των ετικετών του αναγνωριστή ονομάτων οντοτήτων σε επίπεδο λέξης να εμφανίζεται στον ακόλουθο πίνακα (Πίνακας 18):

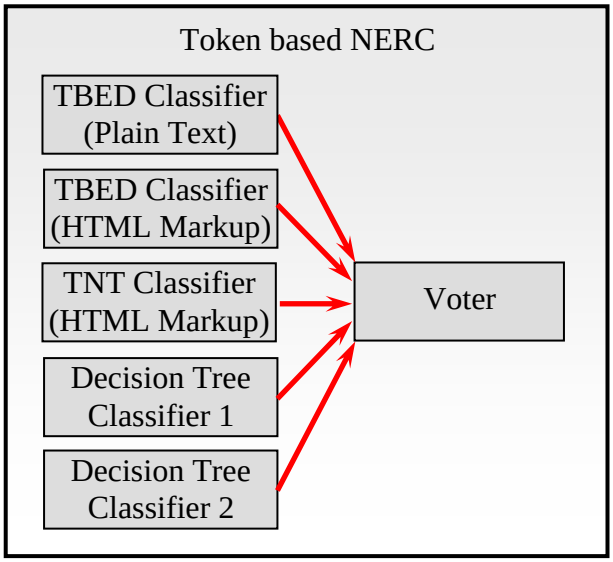
| Τύπος | 1η Λέξη | Υπόλοιπες Λέξεις |
|---|---|---|
| **Οντότητες (NE)** | S:COUNTRY | COUNTRY |
| | S:EDU_TITLE | EDU_TITLE |
| | S:JOB_TITLE | JOB_TITLE |
| | S:LANGUAGE | LANGUAGE |
| | S:MUNICIPALITY | MUNICIPALITY |
| | S:ORGANIZATION | ORGANIZATION |
| | S:REGION | REGION |
| | S:S/W | S/W |
| **Χρονικές εκφράσεις (TIMEX)** | S:DATE | DATE |
| | S:DURATION | DURATION |
| **Αριθμητικές εκφράσεις (NUMEX)** | S:MONEY | MONEY |
| **Όροι (TERMS)** | S:ORG_UNIT | ORG_UNIT |
| | S:SCHEDULE | SCHEDULE |
| **Λέξη εκτός ονόματος οντότητας** | no | no |

**Πίνακας 18: Το πλήρες σύνολο ετικετών του αναγνωριστή ονομάτων οντοτήτων σε επίπεδο λέξης *(token-based NERC)*.**

Η αναμενόμενη έξοδος του αναγνωριστή ονομάτων οντοτήτων σε επίπεδο λέξης είναι η επισημείωση κάθε λέξης με μια ετικέτα από τις ετικέτες του παραπάνω πίνακα (Πίνακας 18), ανάλογα με το αν η λέξη αποτελεί ή όχι μέρος ενός ονόματος οντότητας. Αν η λέξη

δεν αποτελεί μέρος ονόματος οντότητας, αναμένεται να επισημειωθεί με την ετικέτα “no”. Διαφορετικά, αν η λέξη αποτελεί την πρώτη λέξη ενός ονόματος οντότητας, αναμένεται να επισημειωθεί με ετικέτα του τύπου “S:τύπος-ονόματος-οντότητας”, ενώ εάν η λέξη αποτελεί μέρος ενός ονόματος οντότητας και δεν είναι η πρώτη του λέξη, αναμένεται να επισημειωθεί με ετικέτα του τύπου “τύπος-ονόματος-οντότητας”.

Ο αναγνωριστής ονομάτων οντοτήτων σε επίπεδο λέξης αποτελείται από τον συνδυασμό πέντε ταξινομητών που λειτουργούν σε επίπεδο λέξης, αξιοποιώντας διαφορετικούς αλγόριθμους μηχανικής μάθησης, ή διαφορετικές αναπαραστάσεις εισόδου. Η έξοδος των πέντε ταξινομητών, που έργο τους είναι η ανεξάρτητη επισημείωση κάθε λέξης με μια ετικέτα από το σύνολο του πίνακα (Πίνακας 18), συγχωνεύεται με την βοήθεια ενός απλού συστήματος ψηφοφορίας: κάθε λέξη λαμβάνει την ετικέτα στην οποία συμφωνούν οι περισσότεροι ταξινομητές, ή μιας τυχαίας σε περίπτωση ισοβαθμίας. Η αρχιτεκτονική του αναγνωριστή ονομάτων οντοτήτων σε επίπεδο λέξης παρουσιάζεται στην ακόλουθη εικόνα (Εικόνα 9):



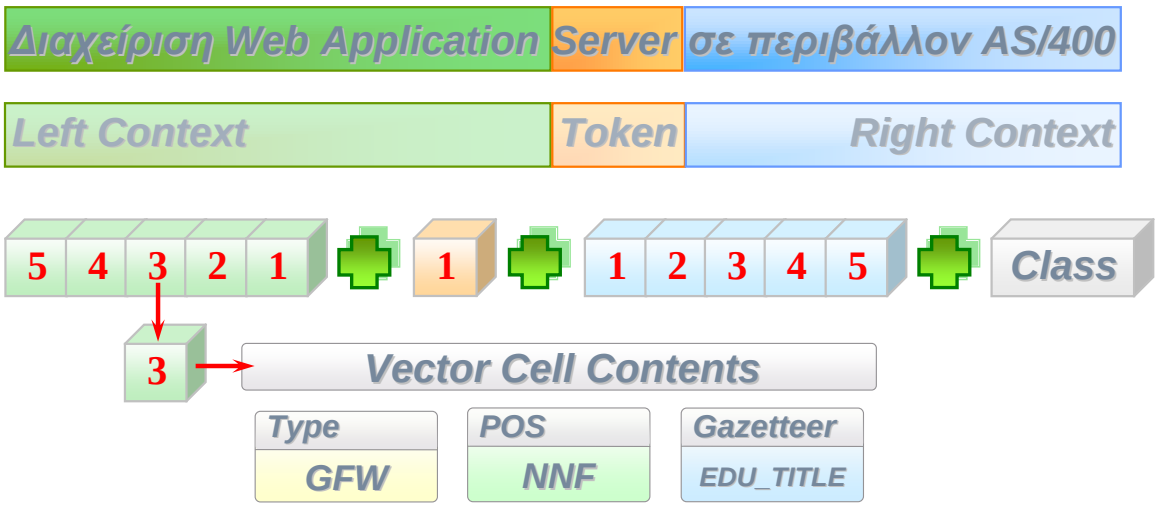

**Εικόνα 9: Η αρχιτεκτονική του αναγνωριστή ονομάτων οντοτήτων σε επίπεδο λέξης *(token-based NERC)*.**

Οι ταξινομητές που απαρτίζουν τον αναγνωριστή ονομάτων οντοτήτων σε επίπεδο λέξης, είναι οι ακόλουθοι:

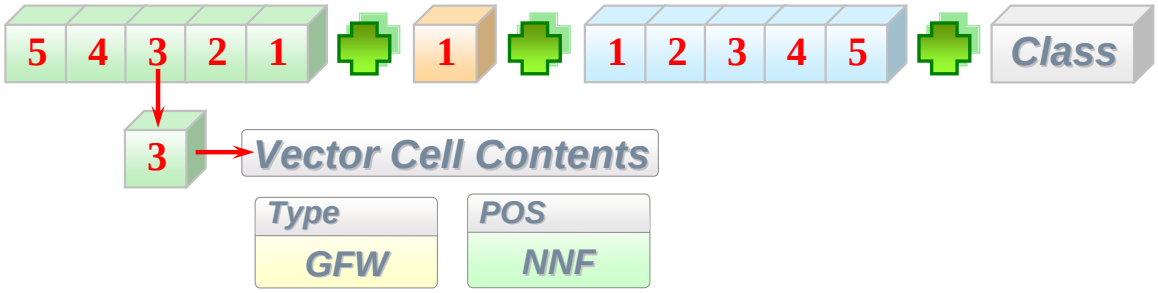
- TBED Classifier (Plain Text): ο ταξινομητής αυτός χρησιμοποιεί μάθηση στηριζόμενη σε *κανόνες μετασχηματισμού καθοδηγούμενη από σφάλματα (transformation-based error-driven learning – TBED)* [27], για να επισημειώσει κάθε λέξη με μια ετικέτα του πίνακα (Πίνακας 18). Αυτός ο αλγόριθμος μηχανικής μάθησης επιλέχθηκε κυρίως για την δυνατότητά του να ταξινομεί λέξεις τόσο χρησιμοποιώντας την μορφολογία της λέξης (εσωτερική πληροφορία), όσο και την κατηγοριοποίηση γειτονικών λέξεων (πληροφορία περιβάλλοντος). Η είσοδος αυτού το ταξινομητή αποτελεί απλό κείμενο (δηλ. κείμενο χωρίς τις ετικέτες HTML σε περίπτωση ιστοσελίδας).
- TBED Classifier (Html Markup): ο ταξινομητής αυτός είναι παρεμφερής με τον προηγούμενο, με την έννοια ότι χρησιμοποιεί τον ίδιο αλγόριθμο μηχανικής μάθησης. Ωστόσο η είσοδος είναι διαφορετική, αφού στην περίπτωση ιστοσελίδων, χρησιμοποιούνται και οι HTML ετικέτες, χωρίς τα χαρακτηριστικά κάθε HTML ετικέτας και τους χαρακτήρες “<” και “>”. Το κίνητρο πίσω από αυτόν

τον ταξινομητή, είναι η αξιοποίηση επιπρόσθετης πληροφορίας που μπορεί να περιέχεται από της HTML ετικέτες, αν αυτές υπάρχουν.

- TNT Classifier (Html Markup): ο ταξινομητής αυτός χρησιμοποιεί τον αλγόριθμο του Viterbi [86] σε μοντέλα Markov δεύτερης τάξεως (όπως υλοποιείται από τον αλγόριθμο TNT) [38], για να επισημειώσει κάθε λέξη με μια ετικέτα του πίνακα (Πίνακας 18). Ο αλγόριθμος αυτός χρησιμοποιεί τόσο την ίδια την λέξη, αλλά και το περιβάλλον κάθε λέξης, κυρίως με την μορφή των ετικετών που αποδόθηκαν σε προηγούμενες λέξεις. Η είσοδος επίσης περιλαμβάνει τις ετικέτες HTML, αν αυτές υπάρχουν, στην ίδια μορφή με τον προηγούμενο ταξινομητή (TBED Classifier (Html Markup)).
- Decision Tree Classifier 1: ο τέταρτος ταξινομητής χρησιμοποιεί τον αλγόριθμο μηχανικής μάθησης C4.5 [25], ο οποίος υλοποιεί δέντρα αποφάσεων. Η είσοδος του ταξινομητή αυτού είναι ένα διάνυσμα σταθερού μήκους, το οποίο περιλαμβάνει χαρακτηριστικά της λέξης προς κατηγοριοποίηση, αλλά ταυτόχρονα περιλαμβάνει και χαρακτηριστικά από τις πέντε προηγούμενες, και τις πέντε επόμενες λέξεις. Σε αντίθεση με τους τρεις ταξινομητές που προηγήθηκαν, οι οποίοι δέχονται σαν είσοδο τις ίδιες τις λέξεις, ο ταξινομητής αυτός δεν χρησιμοποιεί τις ίδιες τις λέξεις. Αντίθετα, κάθε λέξη αναπαριστάται από τρία χαρακτηριστικά. Το πρώτο χαρακτηριστικό αποτελεί μια ετικέτα η οποία περιγράφει το είδος των χαρακτήρων που απαρτίζουν την λέξη: αν είναι κεφαλαίοι οι πεζοί, αν περιέχονται Ελληνικοί ή Αγγλικοί χαρακτήρες, αν μόνο ο πρώτος χαρακτήρας είναι κεφαλαίος, αν περιέχονται νούμερα, ή αν η λέξη είναι σύντμηση (δηλ. περιέχει τον χαρακτήρα της τελείας). Το δεύτερο χαρακτηριστικό περιγράφει το μέρος του λόγου της λέξης, χρησιμοποιώντας το σύνολο ετικετών που περιγράφηκε στο κεφάλαιο 3 και παρουσιάζεται στο ΠΑΡΑΡΤΗΜΑ Ι. Τέλος, το τρίτο χαρακτηριστικό είναι η ετικέτα που δόθηκε στην λέξη από τον αναγνωριστή γνωστών ονομάτων οντοτήτων *(gazetteer list lookup)*, κατά το στάδιο της προ-επεξεργασίας. Μια γραφική αναπαράσταση της αναπαράστασης του διανύσματος δίνεται στην Εικόνα 10.
- Decision Tree Classifier 2: ο ταξινομητής αυτός είναι παρεμφερής με τον προηγούμενο (Decision Tree Classifier 1), με την έννοια ότι χρησιμοποιεί τον ίδιο αλγόριθμο μηχανικής μάθησης. Επίσης, παρεμφερής είναι και η είσοδος, αφού χρησιμοποιεί την ίδια αναπαράσταση, όπου κάθε λέξη (είτε η λέξη προς κατηγοριοποίηση είτε οι πέντε λέξεις πριν και μετά από αυτήν) αναπαριστάται από ένα σύνολο χαρακτηριστικών, εκφρασμένα σαν ένα διάνυσμα σταθερού μήκους, το οποίο περιέχει σαν τιμές σύμβολα. Η διαφορά με τον προηγούμενο ταξινομητή βρίσκεται στο σύνολο των χαρακτηριστικών που περιγράφουν μια λέξη, το οποίο είναι μειωμένο, με το τρίτο χαρακτηριστικό (την ετικέτα από τον αναγνωριστή γνωστών ονομάτων οντοτήτων *(gazetteer list lookup)*) να απουσιάζει. Μια γραφική αναπαράσταση της αναπαράστασης του διανύσματος δίνεται στην Εικόνα 11. Το κίνητρο πίσω από αυτήν την επιλογή είναι ότι αυτός ο ταξινομητής δεν έχει καμία εξάρτηση από τις λίστες γνωστών ονομάτων, οπότε ίσως λειτουργήσει καλύτερα σε ονόματα οντοτήτων που δεν περιέχονται (ολόκληρα ή μέρη τους) στις λίστες.
- Voter: αυτό το υποσύστημα εφαρμόζει μια απλή πλειοψηφική ψηφοφορία, προκειμένου να αποφασίσει για την τελική ταξινόμηση κάθε λέξης σε μια από τις ετικέτες του πίνακα (Πίνακας 18). Δεδομένου ότι κάθε λέξη έχει εξεταστεί και έχει ταξινομηθεί από καθέναν από τους πέντε ταξινομητές, αυτό το υποσύστημα αποφασίζει την τελική ταξινόμηση επιλέγοντας απλά την συχνότερη ετικέτα.




**Εικόνα 10: Η αναπαράσταση του διανύσματος εισόδου του τέταρτου ταξινομητή (Decision Tree Classifier 1).**

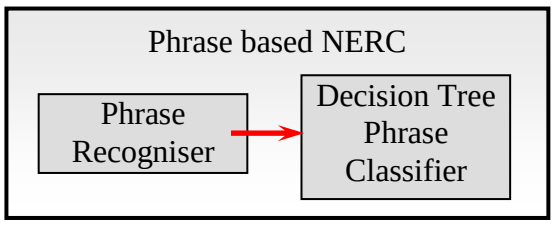



**Εικόνα 11: Η αναπαράσταση του διανύσματος εισόδου του πέμπτου ταξινομητή (Decision Tree Classifier 2).**

### 4.7.3 Αναγνώριση ονομάτων οντοτήτων σε επίπεδο φράσης

Ο αναγνωριστής ονομάτων οντοτήτων σε επίπεδο φράσης υιοθετεί μια διαφορετική προσέγγιση από τον αναγνωριστή ονομάτων οντοτήτων σε επίπεδο λέξης. Ο αναγνωριστής ονομάτων οντοτήτων σε επίπεδο λέξης εξετάζει και επισημειώνει κάθε λέξη, είτε σαν αρχή ονόματος οντότητας, σαν μέρος ονόματος οντότητας, ή σαν λέξη που δεν αποτελεί μέρος ονόματος οντότητας. Ένα πιθανό πρόβλημα με την προσέγγιση αυτή, είναι ότι όλες οι λέξεις ενός ονόματος οντότητας πρέπει να επισημειωθούν σωστά για να αναγνωριστεί ολόκληρο το όνομα της οντότητας. Για παράδειγμα, ακόμα και αν ο χαρακτήρας “/” αναγνωριστεί λανθασμένα στο όνομα οντότητας “AS/400 Software Engineers / Developers”, το όνομα οντότητας μπορεί να αναγνωριστεί εν μέρη, ή να μην αναγνωριστεί καθόλου. Ο αναγνωριστής ονομάτων οντοτήτων σε επίπεδο φράσης προσπαθεί να αντιμετωπίσει προβλήματα αυτού του τύπου, εξετάζοντας και επισημειώνοντας ολόκληρες φράσεις οι οποίες μπορεί να είναι ονόματα οντοτήτων.

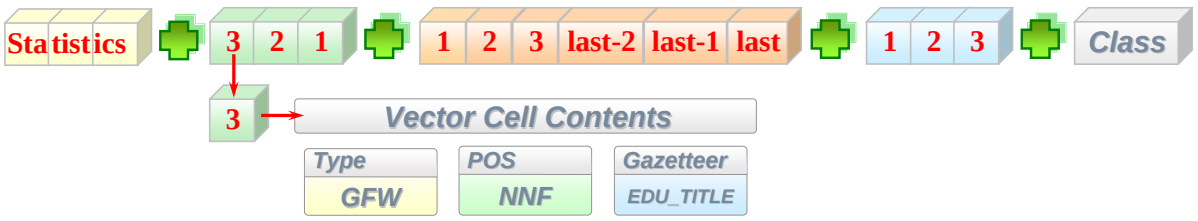
Η αρχιτεκτονική του αναγνωριστή ονομάτων οντοτήτων σε επίπεδο φράσης παρουσιάζεται στην ακόλουθη εικόνα (Εικόνα 12):




**Εικόνα 12: Η αρχιτεκτονική του αναγνωριστή ονομάτων οντοτήτων σε επίπεδο φράσης *(phrase-based NERC)***

Αποτελείται από δύο σημαντικά υπο-συστήματα. Σαν πρώτο βήμα, αναγνωρίζονται όλες οι φράσεις οι οποίες μπορούν να αποτελέσουν πιθανό όνομα οντότητας. Συνήθως, η εργασία αυτή εκτελείται με την βοήθεια ενός συντακτικού αναλυτή, ικανού να αναγνωρίζει *ονοματικές φράσεις (noun phrases)*. Ωστόσο, η χρησιμοποίηση ενός συντακτικού αναλυτή με κατάλληλη γραμματική, θα περιόριζε την ικανότητα του συστήματος ML-HNERC να προσαρμόζεται σε νέες θεματικές περιοχές, και φυσικά γλώσσες, καθώς η προσαρμογή γραμματικών δεν είναι μια τετριμμένη διαδικασία. Αντ' αυτού, επιλέξαμε την αυτόματη εξαγωγή μιας *κανονικής γραμματικής (regular grammar)*, η οποία περιγράφει όλα τα ονόματα οντοτήτων του σώματος κειμένου που χρησιμοποιείται για την εκπαίδευση του συστήματος ML-HNERC, χρησιμοποιώντας τα μέρη του λόγου των εμπλεκόμενων λέξεων. Η εξαχθείσα γραμματική μετατρέπεται σε ένα *αυτόματο (automaton)*, ικανό να αναγνωρίζει τις επιθυμητές φράσεις σε κείμενα, με την βοήθεια του προγράμματος ανοιχτού λογισμικού "GNU Flex". Ένα σημαντικό πλεονέκτημα αυτής της προσέγγισης είναι ότι η εργασία της αναγνώρισης φράσεων που αποτελούν πιθανός ονόματα οντοτήτων, είναι ότι είναι υπολογιστικά γρήγορη και *εύρωστη (robust)*. Ωστόσο, καθώς δεν γίνεται κάποια γενίκευση στην εξαχθείσα γραμματική, ο αναγνωριστής φράσεων μπορεί δυνητικά να αναγνωρίσει λιγότερες φράσεις από μια χειρωνακτικά ανεπτυγμένη γραμματική σε νέα σώματα κειμένων, πέρα από το σώμα κειμένου της εκπαίδευσης.

Από την στιγμή που έχουν αναγνωριστεί οι φράσεις που αποτελούν πιθανά ονόματα οντοτήτων, το δεύτερο υποσύστημα καλείται να αναγνωρίσει ποιες από τις φράσεις αποτελούν ονόματα οντότητας, και να αποδώσει μια σημασιολογική ετικέτα σε κάθε φράση, ανάλογα με τον τύπο της οντότητας (π.χ. όνομα προσώπου, οργανισμού, τοποθεσίας, τίτλος εργασίας, κλπ.). Η επιλογή και επισημείωση γίνεται με την βοήθεια ενός ταξινομητή, που χρησιμοποιεί τον αλγόριθμο μηχανικής μάθησης C4.5 [25], ο οποίος υλοποιεί δέντρα αποφάσεων. Ο ταξινομητής αυτός χρησιμοποιεί σαν είσοδο διανύσματα σταθερού μήκους, χρησιμοποιώντας την αναπαράσταση που εμφανίζεται στην Εικόνα 13.




**Εικόνα 13: Η αναπαράσταση ενός διανύσματος του υποσυστήματος κατηγοριοποίησης φράσεων (Decision Tree Phrase Classifier).**

Το διάνυσμα περιλαμβάνει κάποια στατιστικά στοιχεία της φράσης που αποτελεί πιθανό όνομα οντότητας, όπως:

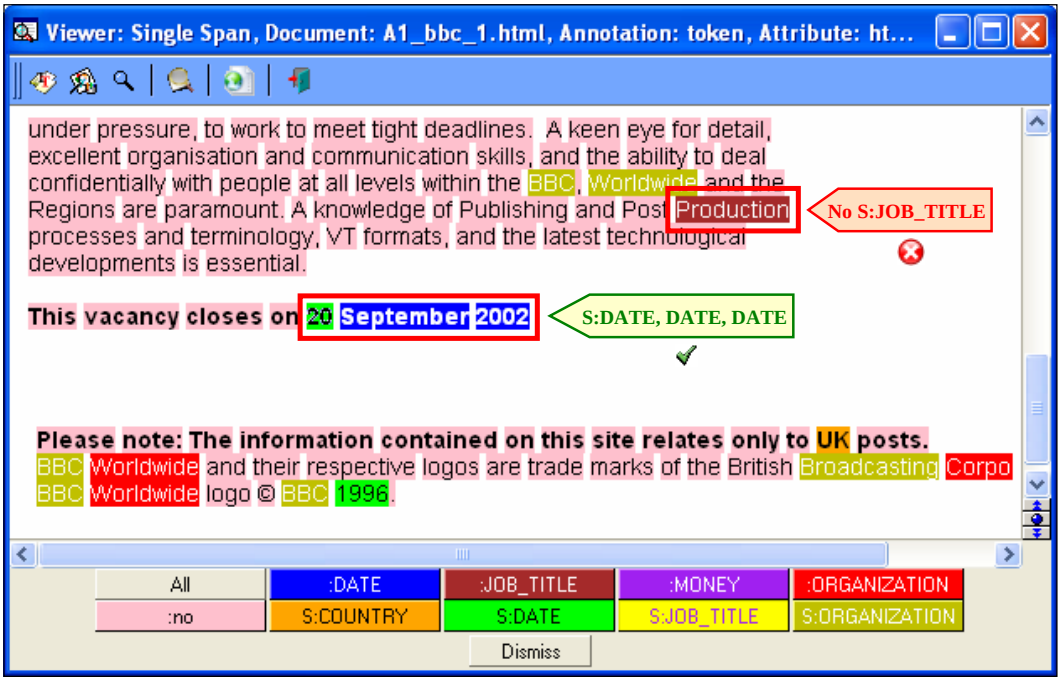
- Το μήκος της φράσης, σε λέξεις (αριθμός λέξεων).
- Την τοποθεσία της φράσης εντός της πρότασης που την περιέχει (αριθμός λέξεων από την αρχή της πρότασης),
- Την συχνότερη ετικέτα που έχει αποδοθεί από το στάδιο της προ-επεξεργασίας στις λέξεις που απαρτίζουν την φράση, και συγκεκριμένα από τον *αναγνωριστή γνωστών ονομάτων οντοτήτων (gazetteer list lookup)*.
- Τις συχνότητες εμφάνισης των ετικετών οι οποίες περιγράφουν το είδος των χαρακτήρων που απαρτίζουν κάθε λέξη, Οι ετικέτες αυτές αποδίδονται από τον αναγνωριστή λέξεων της πλατφόρμας επεξεργασίας φυσικής γλώσσας «Έλλογον» [53], και είναι οι ίδιες ετικέτες που περιγράφηκαν και στο σύστημα αναγνώρισης ονομάτων οντοτήτων σε επίπεδο λέξης, στους ταξινομητές " Decision Tree Classifier 1" και " Decision Tree Classifier 2" (ταξινομητές 4 και 5 αντίστοιχα).

Μετά τα χαρακτηριστικά, στο διάνυσμα αναπαριστάται πληροφορία σχετικά με τις τρεις αρχικές και τις τρεις τελικές λέξης της φράσης (επιτρέποντας επικαλύψεις σε φράσεις με μήκος μικρότερο από 6 λέξεις), καθώς και πληροφορία από τρεις λέξεις πριν και μετά την φράση (οι οποίες περιβάλλουν την φράση). Αυτή η μορφή αναπαράστασης έχει σαν στόχο να μετατρέψει μια φράση μεταβλητού μήκους, σε ένα διάνυσμα σταθερού μήκους. Σταθερό μήκος σημαίνει ότι μπορούμε να αναπαραστήσουμε μόνο μέρος μιας φράσης που ξεπερνά το μέγιστο μήκος που ορίζεται από το μήκος του διανύσματος (το οποίο έχουμε επιλέξει εμπειρικά να είναι έξι λέξεις). Επίσης, η επιλεγμένη αναπαράσταση κωδικοποιεί πληροφορία λέξεων και από τα 2 άκρα κάθε φράσης, σε μια προσπάθεια να αντιπροσωπευτεί καλύτερα η φράση, και να είναι περισσότερο συνεπής με τις λέξεις που βρίσκονται στο άμεσο περιβάλλον της. Κάθε λέξη από τις έξι λέξεις (μέγιστο) που χαρακτηρίζουν κάθε φράση, αναπαριστάται από τρία χαρακτηριστικά. Το πρώτο χαρακτηριστικό αποτελεί μια ετικέτα η οποία περιγράφει το είδος των χαρακτήρων που απαρτίζουν την λέξη: αν είναι κεφαλαίοι οι πεζοί, αν περιέχονται Ελληνικοί ή Αγγλικοί χαρακτήρες, αν μόνο ο πρώτος χαρακτήρας είναι κεφαλαίος, αν περιέχονται νούμερα, ή αν η λέξη είναι σύντμηση (δηλ. περιέχει τον χαρακτήρα της τελείας). Το δεύτερο χαρακτηριστικό περιγράφει το μέρος του λόγου της λέξης, χρησιμοποιώντας το σύνολο ετικετών που περιγράφηκε στο κεφάλαιο 3 και παρουσιάζεται στο ΠΑΡΑΡΤΗΜΑ Ι. Τέλος, το τρίτο χαρακτηριστικό είναι η ετικέτα που δόθηκε στην λέξη από τον αναγνωριστή γνωστών ονομάτων οντοτήτων *(gazetteer list lookup)*, κατά το στάδιο της προ-επεξεργασίας. (Η αναπαράσταση αυτή είναι η ίδια με τον ταξινομητή "Decision Tree Classifier 1", του συστήματος αναγνώρισης ονομάτων οντοτήτων σε επίπεδο λέξης). Τέλος, με τον ίδιο τρόπο (μέσω των ιδίων τριών χαρακτηριστικών) αναπαριστώνται και οι λέξεις από το άμεσο περιβάλλον της φράσης (τρείς λέξεις πριν και τρεις λέξεις μετά την κάθε φράση).

### 4.7.4 Φιλτράρισμα (filtering)

Η κύρια λειτουργία του υπο-συστήματος *φιλτραρίσματος (filtering)*, είναι η μετατροπή των επισημειώσεων του αναγνωριστή ονομάτων οντοτήτων σε επίπεδο λέξης (όπου κάθε λέξη έχει επισημειωθεί ανεξάρτητα από τις υπόλοιπες) σε ονόματα οντοτήτων, τα οποία μπορεί να περιλαμβάνουν από μία έως έναν οποιοδήποτε αριθμό λέξεων (η θεματική περιοχή που επιλέχθηκε για την αποτίμηση του συστήματος ML-HNERC, οι αγγελίες προσφοράς εργασίας, εμφανίζει ονόματα οντοτήτων που ξεπερνούν τις 12 λέξεις, κυρίως σε ονομασίες θέσεων εργασίας). Ανακαλώντας την περιγραφή του αναγνωριστή ονομάτων οντοτήτων σε επίπεδο λέξης, το υποσύστημα αυτό επισημειώνει κάθε λέξη με μια ετικέτα, ανάλογα αν η λέξη αποτελεί αρχή ονόματος οντότητας, μέρος ονόματος οντότητας, ή λέξη εκτός ονόματος οντότητας. Το υποσύστημα φιλτραρίσματος εξετάζει όλες τις επισημειώσεις των λέξεων για να εντοπίσει

*συνεπείς (consistent)* ακολουθίες ετικετών. Μια συνεπής ακολουθία ετικετών αποτελεί μια λέξη επισημειωμένη με την ετικέτα “S:<TYPE>”, που πιθανώς ακολουθείται από μια ή περισσότερες λέξεις επισημειωμένες με την ετικέτα “<TYPE>”. Για παράδειγμα, η Εικόνα 14 απεικονίζει την έξοδο του δεύτερου ταξινομητή (TBED Classifier (Html Markup)) του συστήματος αναγνώρισης ονομάτων οντοτήτων σε επίπεδο λέξης. Η φράση “20 September 2002” έχει χαρακτηριστεί επιτυχώς σαν όνομα χρονικής οντότητας (“TIMEX”), τύπου ημερομηνίας (“DATE”), αφού εντοπίστηκε η συνεπής ακολουθία ετικετών “S:DATE”, “DATE” και “DATE”. Αντίθετα, το κείμενο “Production” δεν χαρακτηρίστηκε σαν όνομα οντότητας, αφού επισημειώθηκε με την ετικέτα “JOB_TITLE”, αλλά η αμέσως προηγούμενη λέξη δεν έχει επισημειωθεί με την ετικέτα “S:JOB_TITLE”.




**Εικόνα 14: Παράδειγμα της εξόδου του συστήματος αναγνώρισης ονομάτων οντοτήτων σε επίπεδο λέξης.**

Μια δεύτερη αρμοδιότητα του υπο-συστήματος φιλτραρίσματος είναι η εφαρμογή χειρωνακτικά κατασκευασμένων κανόνων (με την μορφή *κανονικών εκφράσεων – regular expressions*), που έχουν σαν στόχο τον εντοπισμό περισσότερων ονομάτων οντοτήτων. Οι κανόνες αυτοί (αν υπάρχουν), εφαρμόζονται σε ολόκληρο το κείμενο, μαζί με τις ετικέτες HTML αν τέτοιες υπάρχουν, και επισημείωνουν αυστηρά περιοχές που δεν έχουν ήδη επισημειωθεί (από το υπόλοιπο σύστημα αναγνώρισης ονομάτων οντοτήτων που προηγείται). Οι υπάρχοντες κανόνες αφορούν μόνο τον εντοπισμό ημερομηνιών, τόσο για την Αγγλική, όσο και την Ελληνική γλώσσα.

### 4.7.5 Πειραματική αξιολόγηση και αποτελέσματα

#### Σώμα κειμένου αξιολόγησης

Για την αποτίμηση του συστήματος ML-HNERC συγκεντρώθηκε ένα σώμα κειμένων από 41 διαφορετικούς Ελληνικούς ιστότοπους εταιριών, στους ιστότοπους των οποίων υπάρχουν αγγελίες προσφοράς εργασίας. Το σώμα κειμένων αποτελείται από 50 ιστοσελίδες, οι οποίες περιέχουν 128 αγγελίες προσφοράς εργασίας, τα ονόματα οντοτήτων των οποίων επισημειώθηκαν χειρωνακτικά, στα πλαίσια ενός ερευνητικού

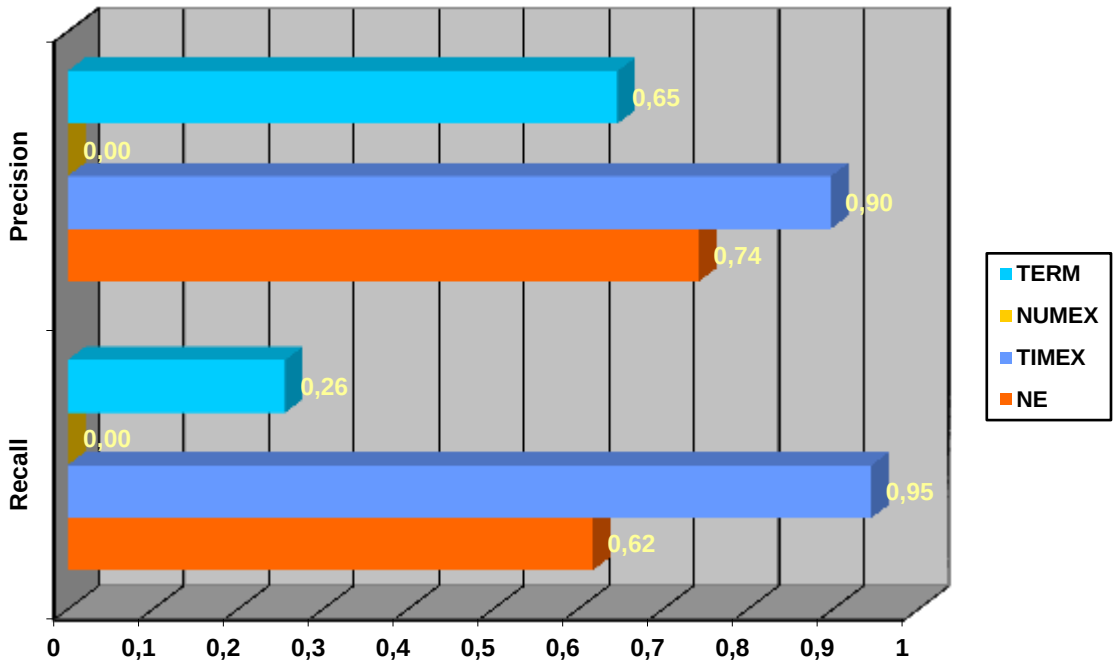
έργου με την ονομασία CROSSMARC [87], [88]. Τα χαρακτηριστικά του σώματος κειμένων αποτίμησης παρουσιάζονται στον ακόλουθο πίνακα (Πίνακας 19):

| | |
|---|---|
| Ιστότοποι | 41 |
| Σελίδες | 50 |
| Αγγελίες προσφοράς εργασίας | 128 |
| Ονόματα οντοτήτων (σύνολο) | 847 |
| Αριθμητικές εκφράσεις (NUMEX) | 0 |
| Χρονικές εκφράσεις (TIMEX) | 48 |
| Όροι (TERM) | 41 |

**Πίνακας 19: Τα χαρακτηριστικά του σωμάτων κειμένων αποτίμησης (θεματική περιοχή: αγγελίες προσφοράς εργασίας).**

## Πειραματική αξιολόγηση για την Ελληνική γλώσσα

Το σύστημα αναγνώρισης ονομάτων οντοτήτων ML-HNERC αποτιμήθηκε για τα Ελληνικά, χρησιμοποιώντας το σώμα κειμένων που περιγράφεται στην προηγούμενη παράγραφο. Τα αποτελέσματα της αξιολόγησης παρουσιάζονται στο ακόλουθο γράφημα (Εικόνα 15), ενώ τα αναλυτικά αποτελέσματα για κάθε τύπο οντότητας παρουσιάζονται στον πίνακα (Πίνακας 20):




**Εικόνα 15: Αποτελέσματα του συστήματος ML-HNERC για την Ελληνική γλώσσα.**

Από τα αποτελέσματα είναι εμφανές ότι το σύστημα αναγνώρισης ονομάτων οντοτήτων ML-HNERC εμφανίζει μια μικρή ροπή προς την ακρίβεια, σε βάρος της ανάκλησης. Παρουσιάζοντας ακρίβεια γύρω στο 75 % με την ανάκληση να βρίσκεται στο 62 %, τα αποτελέσματα είναι ικανοποιητικά, για ένα πλήρως αυτοματοποιημένο σύστημα για την Ελληνική γλώσσα, το οποίο δεν αξιοποιεί κανέναν λεκτικό πόρο, εκτός από έναν

αναγνωριστή ονομάτων οντοτήτων για την Ελληνική γλώσσα. Από τα αναλυτικά αποτελέσματα που εμφανίζονται στον πίνακα (Πίνακας 20), είναι εμφανές ότι ο τύπος ονομάτων οντοτήτων «τίτλος θέσης» ("JOB_TITLE") είναι εκείνος με την χαμηλότερη απόδοση (με την μετρική F-Measure να βρίσκεται στο 50 %), ενώ ταυτόχρονα τα ονόματα αυτού του τύπου αποτελούν το 15 %, οδηγώντας σε μείωση την απόδοση ολόκληρου του συστήματος αναγνώρισης ονομάτων οντοτήτων ML-HNERC για τα Ελληνικά. Ο λόγος για αυτή την μειωμένη απόδοση μπορεί να αποδοθεί στην μεγάλη ελευθερία με την οποία περιγράφονται οι τίτλοι των θέσεων εργασίας, καθώς οι αγγελίες που εμφανίζονται στους σχετικούς ιστόχωρους ευρέσεως εργασίας δεν φαίνεται να ακολουθούν κάποιους κανόνες σχετικά με τις ονομασίες των θέσεων εργασίας. Ταυτόχρονα είναι αρκετά περιγραφικά, κάνοντας των εντοπισμός τους δυσκολότερο. Για παράδειγμα, το σύστημα ML-HNERC απέτυχε να αναγνωρίσει τον τίτλο "Διαχείριση Web application server σε περιβάλλον AS/400", αναγνωρίζοντας μόνο τμήμα αυτού ("Διαχείριση Web application server"), το οποίο προσμετράτε φυσικά σαν λάθος του συστήματος ML-HNERC. Ταυτόχρονα, δεν είναι απίθανο η ίδια θέση να περιγράφεται με διαφορετική ονομασία, ακόμα και αν η θέση προέρχεται από την ίδια εταιρία, αυξάνοντας την δυσκολία αναγνώρισης, αφού το ίδιο όνομα οντότητας μπορεί να εκφραστεί με διαφορετικούς τρόπους.

| | Ακρίβεια | Ανάκληση | F-Measure | Σωστά | Επισημειωμένα | Απάντηση |
|---|---|---|---|---|---|---|
| Όλοι οι τύποι | 0.753 | 0.621 | 0.681 | 562 | 905 | 746 |
| **NE** | 0.743 | 0.618 | 0.675 | 498 | 806 | 670 |
| COUNTRY | 0.800 | 0.889 | 0.842 | 8 | 9 | 10 |
| EDU_TITLE | 0.746 | 0.602 | 0.667 | 50 | 83 | 67 |
| JOB_TITLE | 0.562 | 0.453 | 0.502 | 63 | 139 | 112 |
| LANGUAGE | 0.985 | 0.928 | 0.955 | 64 | 69 | 65 |
| MUNICIPALITY | 0.889 | 0.769 | 0.825 | 40 | 52 | 45 |
| ORGANIZATION | 0.787 | 0.366 | 0.500 | 37 | 101 | 47 |
| REGION | 0.667 | 0.286 | 0.400 | 2 | 7 | 3 |
| S/W | 0.729 | 0.676 | 0.702 | 234 | 346 | 321 |
| **TIMEX** | 0.898 | 0.946 | 0.922 | 53 | 56 | 59 |
| DATE | 0.964 | 0.900 | 0.931 | 27 | 30 | 28 |
| DURATION | 0.839 | 1.000 | 0.912 | 26 | 26 | 31 |
| **NUMEX** | 0.000 | 0.000 | 0.000 | 0 | 0 | 0 |
| MONEY | 0.000 | 0.000 | 0.000 | 0 | 0 | 0 |
| **TERM** | 0.647 | 0.256 | 0.367 | 11 | 43 | 17 |
| ORG_UNIT | 0.647 | 0.275 | 0.386 | 11 | 40 | 17 |
| SCHEDULE | 0.000 | 0.000 | 0.000 | 0 | 3 | 0 |

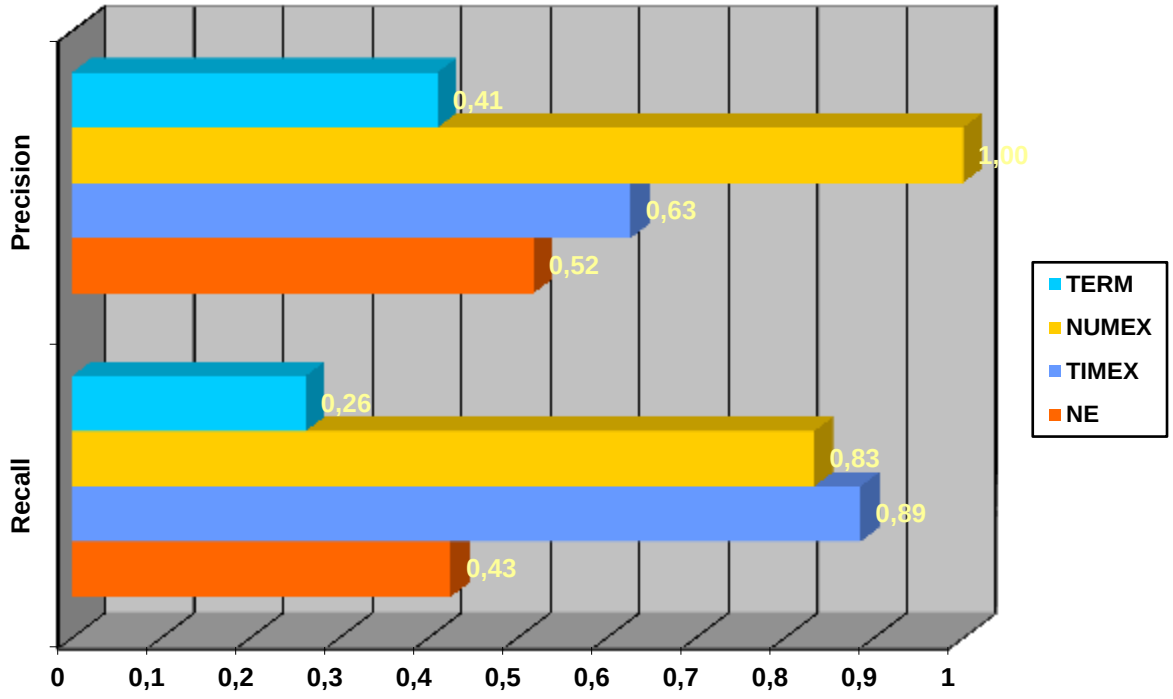
**Πίνακας 20: Αναλυτικά αποτελέσματα (ανά τύπο οντότητας) του συστήματος ML-HNERC για την Ελληνική γλώσσα.**

Παρεμφερή απόδοση με του τίτλους θέσεως εργασίας εμφανίζουν και τα ονόματα εταιριών ("ORGANIZATION"), όμως ο λόγος της μειωμένης απόδοσης είναι εντελώς διαφορετικός. Επειδή τα σώματα κειμένων, τόσο της εκπαίδευσης όσο και της αποτίμησης, συλλέχθηκαν από ιστότοπους εταιριών, οι αγγελίες αρκετές φορές δεν

περιέχουν καθόλου το όνομα της εταιρίας που προσφέρει την θέση, ενώ αρκετές φορές το όνομα της εταιρίας εμφανίζεται σε γενικές περιοχές, όπως σημειώσεις για θέματα πνευματικών δικαιωμάτων. Μην έχοντας αρκετά παραδείγματα στα δεδομένα εκπαίδευσης σχετικά με ονόματα εταιριών, το σύστημα ML-HNERC εμφανίζει αυξημένη δυσκολία στον εντοπισμό τους σε νέες ιστοσελίδες, στην συγκεκριμένη θεματική περιοχή.

### Πειραματική αξιολόγηση για την Αγγλική γλώσσα

Πέρα από την Ελληνική γλώσσα, το σύστημα ML-HNERC αξιολογήθηκε και στην Αγγλική γλώσσα, στην ίδια θεματική περιοχή (αγγελίες προσφοράς εργασίας), χρησιμοποιώντας πάλι σώματα κειμένων (τόσο εκπαίδευσης όσο και αποτίμησης) που παραχωρήθηκαν από το έργο CROSSMARC. Τα αποτελέσματα της αξιολόγησης παρουσιάζονται στο ακόλουθο γράφημα (Εικόνα 16), ενώ τα αναλυτικά αποτελέσματα για κάθε τύπο οντότητας παρουσιάζονται στον πίνακα (Πίνακας 21):




**Εικόνα 16: Αποτελέσματα του συστήματος ML-HNERC για την Αγγλική γλώσσα.**

Τα αποτελέσματα της αξιολόγησης δείχνουν ότι υπάρχει μια πτώση στην συνολική απόδοση του συστήματος ML-HNERC, όταν αυτό μεταφέρεται στην Αγγλική γλώσσα, με την απόδοση να πέφτει από το 68 % (F - Measure για την Ελληνική γλώσσα) στο 52 % (F -Measure για την Αγγλική γλώσσα). Παρατηρούμε την ίδια δυσκολία στον εντοπισμό των τίτλων θέσεων εργασίας, μόνο που στην Αγγλική γλώσσα οι τίτλοι θέσεων εργασίας είναι σημαντικά περισσότεροι, αποτελώντας το 26 % των συνολικών ονομάτων οντοτήτων (έναντι του 15 % στο Ελληνικό σώμα κειμένων). Επίσης η απόδοση στα ονόματα εταιριών παρέμεινε περίπου σταθερή ανάμεσα στις δύο γλώσσες, εμφανίζοντας μια πτώση περίπου 3 % στην Αγγλική γλώσσα.

| | Ακρίβεια | Ανάκληση | F-Measure | Σωστά | Επισημειωμένα | Απάντηση |
|---|---|---|---|---|---|---|
| Όλοι οι τύποι | 0,563 | 0,484 | 0,521 | 437 | 903 | 776 |
| **NE** | 0,558 | 0,458 | 0,503 | 347 | 758 | 622 |
| COUNTRY | 0,789 | 0,882 | 0,833 | 30 | 34 | 38 |
| EDU_TITLE | 0,188 | 0,100 | 0,130 | 3 | 30 | 16 |
| JOB_TITLE | 0,461 | 0,467 | 0,464 | 112 | 240 | 243 |
| LANGUAGE | 0,500 | 1,000 | 0,667 | 3 | 3 | 6 |
| MUNICIPALITY | 0,750 | 0,581 | 0,655 | 18 | 31 | 24 |
| ORGANIZATION | 0,734 | 0,348 | 0,473 | 69 | 198 | 94 |
| REGION | 0,737 | 0,424 | 0,538 | 14 | 33 | 19 |
| S/W | 0,538 | 0,519 | 0,528 | 98 | 189 | 182 |
| **TIMEX** | 0,627 | 0,885 | 0,734 | 69 | 78 | 110 |
| DATE | 0,500 | 0,829 | 0,624 | 29 | 35 | 58 |
| DURATION | 0,769 | 0,930 | 0,842 | 40 | 43 | 52 |
| **NUMEX** | 1,000 | 0,833 | 0,909 | 5 | 6 | 5 |
| MONEY | 1,000 | 0,833 | 0,909 | 5 | 6 | 5 |
| **TERM** | 0,410 | 0,262 | 0,320 | 16 | 61 | 39 |
| ORG_UNIT | 0,395 | 0,254 | 0,309 | 15 | 59 | 38 |
| SCHEDULE | 1,000 | 0,500 | 0,667 | 1 | 2 | 1 |

**Πίνακας 21: Αναλυτικά αποτελέσματα (ανά τύπο οντότητας) του συστήματος ML-HNERC για την Αγγλική γλώσσα.**

Στα πλαίσια του έργου CROSSMARC, κατασκευάστηκε ένα διαφορετικό σύστημα αναγνώρισης ονομάτων οντοτήτων, επίσης βασισμένο σε μηχανική μάθηση, από το πανεπιστήμιο του Εδιμβούργου, στο Ηνωμένο Βασίλειο. Το σύστημα αυτό, στο οποίο θα αναφερόμαστε με την ονομασία ML-ENERC, βασίζεται σε *ταξινομητή μεγιστοποίησης της εντροπίας (maximum entropy tagger)* [44], και συγκεκριμένα στον αλγόριθμο C&C [89]. Το σύστημα αναγνώρισης ονομάτων οντοτήτων ML-ENERC αποτιμήθηκε στο ίδιο σώμα κειμένων όπως και το σύστημα ML-HNERC, από το πανεπιστήμιο του Εδιμβούργου. Τα αναλυτικά αποτελέσματα της αξιολόγησης για το σύστημα ML-ENERC (για κάθε τύπο οντότητας) παρουσιάζονται στον ακόλουθο πίνακα (Πίνακας 22):

| | Ακρίβεια | Ανάκληση | F-Measure | Σωστά | Επισημειωμένα | Απάντηση |
|---|---|---|---|---|---|---|
| Όλοι οι τύποι | 0,648 | 0,430 | 0,517 | 388 | 903 | 599 |
| **NE** | 0,626 | 0,430 | 0,510 | 326 | 758 | 521 |
| COUNTRY | 0,833 | 0,882 | 0,857 | 30 | 34 | 36 |
| EDU_TITLE | 0,200 | 0,033 | 0,057 | 1 | 30 | 5 |
| JOB_TITLE | 0,588 | 0,446 | 0,507 | 107 | 240 | 182 |
| LANGUAGE | 0,400 | 0,667 | 0,500 | 2 | 3 | 5 |
| MUNICIPALITY | 0,727 | 0,516 | 0,604 | 16 | 31 | 22 |
| ORGANIZATION | 0,729 | 0,354 | 0,476 | 70 | 198 | 96 |
| REGION | 0,600 | 0,273 | 0,375 | 9 | 33 | 15 |
| S/W | 0,569 | 0,481 | 0,521 | 91 | 189 | 160 |
| **TIMEX** | 0,786 | 0,705 | 0,743 | 55 | 78 | 70 |
| DATE | 0,727 | 0,686 | 0,706 | 24 | 35 | 33 |
| DURATION | 0,838 | 0,721 | 0,775 | 31 | 43 | 37 |
| **NUMEX** | 1,000 | 0,333 | 0,500 | 2 | 6 | 2 |
| MONEY | 1,000 | 0,333 | 0,500 | 2 | 6 | 2 |
| **TERM** | 0,833 | 0,082 | 0,149 | 5 | 61 | 6 |
| ORG_UNIT | 0,833 | 0,085 | 0,154 | 5 | 59 | 6 |
| SCHEDULE | 0.000 | 0.000 | 0.000 | 0 | 2 | 0 |

**Πίνακας 22: Αναλυτικά αποτελέσματα (ανά τύπο οντότητας) του συστήματος ML-ENERC για την Αγγλική γλώσσα.**

Από την σύγκριση των πινάκων Πίνακας 21 και Πίνακας 22, προκύπτει ότι η απόδοση των δύο συστημάτων (ML-HNERC και ML-ENERC αντίστοιχα) είναι περίπου ισοδύναμη, με τα δύο συστήματα να εμφανίζουν μια συνολική απόδοση κοντά στο 52 % (F-Measure). Ωστόσο υπάρχει μια μικρή υπεροχή της προτεινόμενης προσέγγισης (ML-HNERC) σε σχέση με το σύστημα του πανεπιστημίου του Εδιμβούργου (ML-ENERC) στον απόλυτο αριθμό των σωστών αποτελεσμάτων (437 έναντι 388). Το σύστημα ML-HNERC επέστρεψε έναν σημαντικά μεγαλύτερο αριθμό ετικετών (776 έναντι 599), που σημαίνει ότι μπόρεσε να εντοπίσει έναν μεγαλύτερο αριθμό ονομάτων οντοτήτων, το οποία όμως δεν κατάφερε να τα αναγνωρίσει στην ολότητά τους (δηλ. να αναγνωρίσει επιτυχώς όλες τις λέξεις που τα απαρτίζουν), εμφανίζοντας μια πτώση στην ακρίβεια (από 64.8 % για το σύστημα ML-ENERC σε 56.3 % για το σύστημα ML-HNERC), η οποία αντισταθμίζει τα οφέλη που προκύπτουν στην ανάκληση. Παρόλα αυτά, θεωρούμε ότι το σύστημα ML-HNERC έχει πετύχει τους στόχους για τους οποίους κατασκευάστηκε. Πρόκειται για ένα σύστημα που έχει υποκαταστήσει πλήρως το υποσύστημα της γραμματικής με μια προσέγγιση που χρησιμοποιεί μηχανική μάθηση, η οποία μπορεί να προσαρμοστεί ευκολότερα σε νέες θεματικές περιοχές και γλώσσες. Η αποτίμηση εξέτασε το δυσκολότερο σενάριο, αυτό της προσαρμογής σε μια νέα γλώσσα (με αποτίμηση στην Αγγλική και Ελληνική γλώσσα, στην ίδια θεματική περιοχή), με την απόδοση του συστήματος να βρίσκεται στα ίδια επίπεδα με ένα σύστημα που κατασκευάστηκε για την Αγγλική γλώσσα, με ανάλογους περιορισμούς (δηλ. χωρίς την χρήση επιπρόσθετων πόρων, όπως λεξικά γνωστών ονομάτων οντοτήτων).

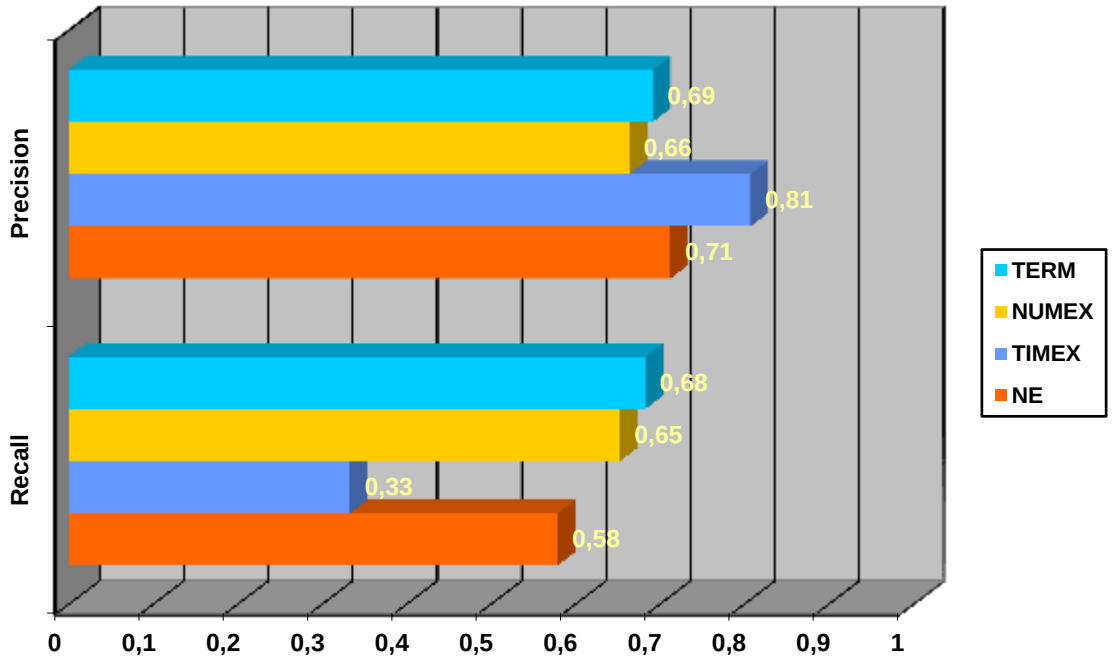
### Πειραματική αξιολόγηση για νέα θεματική περιοχή στην Ελληνική γλώσσα

Το σύστημα ML-HNERC αξιολογήθηκε και σε μια διαφορετική θεματική περιοχή για τα Ελληνικά, έχοντας σαν στόχο την αποτίμηση της προσαρμογής σε μια νέα θεματική περιοχή. Το σώμα κειμένων που χρησιμοποιήθηκε προέρχεται πάλι από το έργο CROSSMARC, και η θεματική του περιοχή είναι περιγραφές προϊόντων (φορητοί υπολογιστές – laptops) από ιστοσελίδες ηλεκτρονικών καταστημάτων. Το σώμα κειμένων που χρησιμοποιήθηκε για την αποτίμηση περιέχει 84 ιστοσελίδες, οι οποίες περιγράφουν 119 φορητούς υπολογιστές. Τα χαρακτηριστικά του σώματος κειμένων αποτίμησης παρουσιάζονται στον ακόλουθο πίνακα (Πίνακας 23):

| | |
|---|---|
| Ιστότοποι | 17 |
| Σελίδες | 84 |
| Προϊόντα (φορητοί υπολογιστές) | 119 |
| Ονόματα οντοτήτων (σύνολο) | 1985 |
| Αριθμητικές εκφράσεις (NUMEX) | 1373 |
| Χρονικές εκφράσεις (TIMEX) | 78 |
| Όροι (TERM) | 0 |

**Πίνακας 23: Τα χαρακτηριστικά του σωμάτων κειμένων αποτίμησης (θεματική περιοχή: περιγραφές φορητών υπολογιστών από ιστοσελίδες ηλεκτρονικών καταστημάτων).**

Τα αποτελέσματα της αξιολόγησης παρουσιάζονται στο ακόλουθο γράφημα (Εικόνα 17), ενώ τα αναλυτικά αποτελέσματα για κάθε τύπο οντότητας παρουσιάζονται στον πίνακα (Πίνακας 24):




**Εικόνα 17: Αποτελέσματα του συστήματος ML-HNERC για την Ελληνική γλώσσα (θεματική περιοχή: περιγραφές φορητών υπολογιστών από ιστοσελίδες ηλεκτρονικών καταστημάτων).**

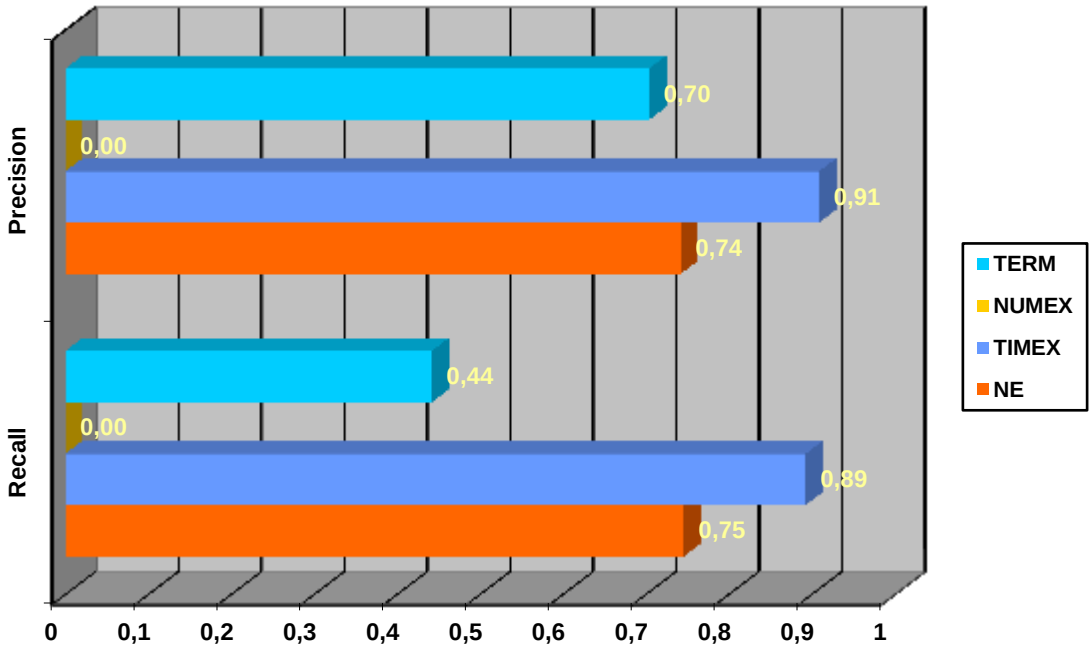
Όπως φαίνεται από τους πίνακες (Πίνακας 20) και (Πίνακας 24), η απόδοση του συστήματος ML-HNERC παραμένει περίπου η ίδια ανάμεσα στις δύο διαφορετικές περιοχές (αγγελίες προσφοράς εργασίας και περιγραφές φορητών υπολογιστών από ιστοσελίδες ηλεκτρονικών καταστημάτων). Η θεματική περιοχή των περιγραφών φορητών υπολογιστών κυριαρχείται από ονόματα οντοτήτων που σχετίζονται με αριθμητικές εκφράσεις, όπως οι χωρητικότητες σκληρών δίσκων, οι διαστάσεις των οθονών, οι συχνότητες των επεξεργαστών, οι τιμές πώλησης, το βάρος κάθε φορητού υπολογιστή κ.α., με την απόδοση του συστήματος ML-HNERC να κυμαίνεται γύρω στο 65 % (F-Measure) για αυτούς τους τύπους οντοτήτων, οι οποίοι δεν ήταν τόσο σημαντικοί για την προηγούμενη θεματική περιοχή των αγγελιών προσφοράς εργασίας. Όσον αφορά τα ονόματα οντοτήτων, οι τύποι που δυσκόλεψαν περισσότερο το σύστημα ML-HNERC είναι οι ονομασίες των μοντέλων φορητών υπολογιστών, και οι ονομασίες των εταιριών που κατασκευάζουν τους φορητούς υπολογιστές, το οποίο είναι αναμενόμενο καθώς αυτοί οι τύποι ονομάτων οντοτήτων αναμένεται να έχουν την μεγαλύτερη διακύμανση ανάμεσα στους υπόλοιπους τύπους, καθώς υπάρχουν περισσότεροι κατασκευαστές/μοντέλα σε σχέση με τις υπόλοιπες οντότητες, οι οποίες είναι περισσότερο τυποποιημένες.

| | Ακρίβεια | Ανάκληση | F-Measure | Σωστά | Επισημειωμένα | Απάντηση |
|---|---|---|---|---|---|---|
| Όλοι οι τύποι | 0.682 | 0.622 | 0.651 | 1265 | 2032 | 1855 |
| **NE** | 0.713 | 0.579 | 0.639 | 367 | 634 | 515 |
| MANUFACTUR. | 0.607 | 0.542 | 0.572 | 91 | 168 | 150 |
| MODEL | 0.645 | 0.506 | 0.567 | 78 | 154 | 121 |
| PROCESSOR | 0.786 | 0.544 | 0.643 | 92 | 169 | 117 |
| SOFT_OS | 0.835 | 0.741 | 0.785 | 106 | 143 | 127 |
| **TIMEX** | 0.808 | 0.333 | 0.472 | 21 | 63 | 26 |
| DATE | 1.000 | 0.286 | 0.444 | 2 | 7 | 2 |
| DURATION | 0.792 | 0.339 | 0.475 | 19 | 56 | 24 |
| **NUMEX** | 0.664 | 0.653 | 0.659 | 767 | 1174 | 1155 |
| CAPACITY | 0.613 | 0.603 | 0.608 | 214 | 355 | 349 |
| LENGTH | 0.824 | 0.709 | 0.762 | 117 | 165 | 142 |
| MONEY | 0.456 | 0.631 | 0.529 | 118 | 187 | 259 |
| PERCENT | 0.558 | 0.967 | 0.707 | 29 | 30 | 52 |
| RESOLUTION | 0.727 | 0.533 | 0.615 | 8 | 15 | 11 |
| SIMPLE | 0.000 | 0.000 | 0.000 | 0 | 14 | 0 |
| SPEED | 0.816 | 0.690 | 0.748 | 271 | 393 | 332 |
| WEIGHT | 1.000 | 0.667 | 0.800 | 10 | 15 | 10 |
| **TERM** | 0.692 | 0.683 | 0.688 | 110 | 161 | 159 |
| TERM | 0.692 | 0.683 | 0.688 | 110 | 161 | 159 |

**Πίνακας 24: Αναλυτικά αποτελέσματα (ανά τύπο οντότητας) του συστήματος ML-HNERC για την Ελληνική γλώσσα (θεματική περιοχή: περιγραφές φορητών υπολογιστών από ιστοσελίδες ηλεκτρονικών καταστημάτων).**

### Πειραματική αξιολόγηση με προσθήκη γνώσης για την θεματική περιοχή

Όλα τα αποτελέσματα που παρουσιάστηκαν μέχρι τώρα, αφορούν το σύστημα ML-HNERC όπως αυτό έχει εκπαιδευτεί στην εκάστοτε θεματική περιοχή/γλώσσα, χρησιμοποιώντας μόνο όση πληροφορία περιέχεται στο σώμα κειμένων που έχει χρησιμοποιηθεί για εκπαίδευση του συστήματος (το οποίο έχει το ίδιο μέγεθος και τα ίδια χαρακτηριστικά με το σώμα κειμένων αποτίμησης σε όλες τις περιπτώσεις που ήδη εξετάστηκαν). Για παράδειγμα, οι λίστες γνωστών ονομάτων δημιουργούνται μόνο από τις επισημειωμένες οντότητες που βρίσκονται εντός του σώματος κειμένου εκπαίδευσης, χρησιμοποιώντας την διαδικασία που περιγράφηκε στην παράγραφο 4.7.1 (Γλωσσική προ-επεξεργασία). Ωστόσο, θα ήταν ενδιαφέρον να εξεταστεί η μεταβολή στην απόδοση του συστήματος ML-HNERC, στην περίπτωση όπου υπάρχει επιπλέον γνώση για την θεματική περιοχή, με την μορφή υπαρχουσών λιστών γνωστών ονομάτων οντοτήτων. Σε αυτή την αξιολόγηση, εξετάζουμε την απόδοση του συστήματος όταν χρησιμοποιεί έναν *αναγνωριστή γνωστών ονομάτων οντοτήτων (gazetteer list lookup)*, εμπλουτίζοντας τις λίστες γνωστών ονομάτων που εξάγονται από το σώμα κειμένων εκπαίδευσης, με επιπρόσθετες λίστες γνωστών ονομάτων. Για τον εμπλουτισμό των λιστών γνωστών ονομάτων χρησιμοποιήθηκαν λίστες οι οποίες συλλέχθηκαν χειρωνακτικά από ιστοσελίδες της ίδιας θεματικής περιοχής. Τα αποτελέσματα για την θεματική περιοχή των αγγελιών προσφοράς εργασίας, με χρήση επιπρόσθετων λιστών ονομάτων από την ίδια θεματική περιοχή, παρουσιάζονται στο ακόλουθο γράφημα (Εικόνα 18), ενώ τα αναλυτικά αποτελέσματα για κάθε τύπο οντότητας παρουσιάζονται στον πίνακα (Πίνακας 25):




**Εικόνα 18: Αποτελέσματα του συστήματος ML-HNERC για την Ελληνική γλώσσα, χρησιμοποιώντας εμπλουτισμένες λίστες γνωστών ονομάτων οντοτήτων.**

Συγκρίνοντας τα αποτελέσματα του πίνακα (Πίνακας 20), με τα αποτελέσματα του πίνακα (Πίνακας 25), παρατηρούμε μια αισθητή βελτίωση της απόδοσης του συστήματος, από το 68 % (F-Measure) στο 75 % (F-Measure). Ο εμπλουτισμός των λιστών γνωστών ονομάτων έχει βελτιώσει κυρίως την ανάκληση του συστήματος ML-

HNERC, η οποία έχει αυξηθεί από 62.1% σε 74.6% ενώ η ακρίβεια του συστήματος έχει παραμείνει στα ίδια επίπεδα. Αυτή η αύξηση στην ανάκληση είναι αναμενόμενη, δεδομένου ότι το σύστημα έχει πλέον περισσότερες ενδείξεις για μέρη πιθανών ονομάτων οντοτήτων από τις λίστες γνωστών ονομάτων, γεγονός που οδηγεί στη σωστή αναγνώριση περισσότερων ονομάτων οντοτήτων.

| | Ακρίβεια | Ανάκληση | F-Measure | Σωστά | Επισημειωμένα | Απάντηση |
|---|---|---|---|---|---|---|
| Όλοι οι τύποι | 0.751 | 0.746 | 0.748 | 675 | 905 | 899 |
| **NE** | 0.742 | 0.752 | 0.747 | 606 | 806 | 817 |
| COUNTRY | 0.800 | 0.889 | 0.842 | 8 | 9 | 10 |
| EDU_TITLE | 0.727 | 0.675 | 0.700 | 56 | 83 | 77 |
| JOB_TITLE | 0.608 | 0.626 | 0.617 | 87 | 139 | 143 |
| LANGUAGE | 0.971 | 0.971 | 0.971 | 67 | 69 | 69 |
| MUNICIPALITY | 0.844 | 0.731 | 0.784 | 38 | 52 | 45 |
| ORGANIZATION | 0.745 | 0.812 | 0.777 | 82 | 101 | 110 |
| REGION | 0.200 | 0.143 | 0.167 | 1 | 7 | 5 |
| S/W | 0.746 | 0.772 | 0.759 | 267 | 346 | 358 |
| **TIMEX** | 0.909 | 0.893 | 0.901 | 50 | 56 | 55 |
| DATE | 0.889 | 0.800 | 0.842 | 24 | 30 | 27 |
| DURATION | 0.929 | 1.000 | 0.963 | 26 | 26 | 28 |
| **NUMEX** | 0.000 | 0.000 | 0.000 | 0 | 0 | 0 |
| MONEY | 0.000 | 0.000 | 0.000 | 0 | 0 | 0 |
| **TERM** | 0.704 | 0.442 | 0.543 | 19 | 43 | 27 |
| ORG_UNIT | 0.708 | 0.425 | 0.531 | 17 | 40 | 24 |
| SCHEDULE | 0.667 | 0.667 | 0.667 | 2 | 3 | 3 |

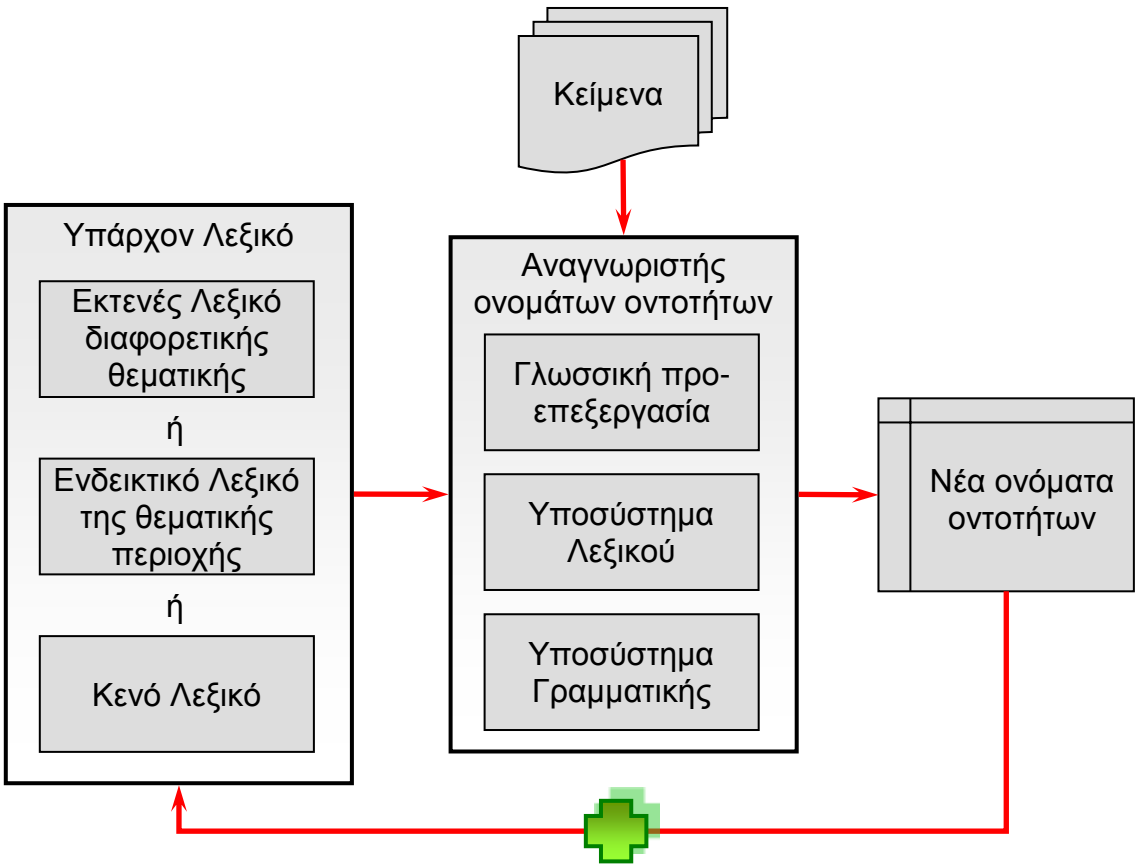
**Πίνακας 25: Αναλυτικά αποτελέσματα (ανά τύπο οντότητας) του συστήματος ML-HNERC για την Ελληνική γλώσσα, χρησιμοποιώντας εμπλουτισμένες λίστες γνωστών ονομάτων οντοτήτων.**

## 4.8 Προσαρμογή/εμπλουτισμός του υποσυστήματος του λεξικού

Το υποσύστημα του λεξικού είναι ένας γλωσσικός πόρος, η ύπαρξη του οποίου μπορεί να συνεισφέρει θετικά σε ένα σύστημα αναγνώρισης ονομάτων οντοτήτων, προτείνοντας πιθανά ονόματα οντοτήτων ή προσδιοριστών τους, τα οποία προέρχονται από λίστες γνωστών ονομάτων οντοτήτων. Η χειρωνακτική προσαρμογή τους σε νέες περιοχές και γλώσσες μοιράζεται τις ίδιες δυσκολίες με τα υπόλοιπα υποσυστήματα ενός συστήματος αναγνώρισης ονομάτων οντοτήτων (όπως η γραμματική), καθιστώντας την αυτόματη προσαρμογή του υποσυστήματος του λεξικού το ίδιο επιθυμητή. Στην ενότητα αυτή παρουσιάζεται μια προσέγγιση που σαν στόχο έχει τον εμπλουτισμό ενός λεξικού γνωστών ονομάτων οντοτήτων μέσω μηχανικής μάθησης. Η προσέγγιση βασίζεται στην ιδέα ότι ένας αναγνωριστής ονομάτων οντοτήτων μπορεί να λειτουργήσει έχοντας το υποσύστημα της γραμματικής σαν το βασικό υποσύστημα αναγνώρισης, στην περίπτωση που είτε το υποσύστημα λεξικού υπολειτουργεί (π.χ. λόγω του μικρού αριθμού γνωστών ονομάτων οντοτήτων που περιέχονται) ή στην περίπτωση που το υποσύστημα λεξικού απουσιάζει εντελώς.

### 4.8.1 Η προσέγγιση

Η προτεινόμενη προσέγγιση για την προσαρμογή/εμπλουτισμό του υποσυστήματος του λεξικού, βασίζεται στην χρήση ενός αναγνωριστή ονομάτων οντοτήτων για την εξαγωγή ονομάτων οντοτήτων, τα οποία μετέπειτα χρησιμοποιούνται για τον εμπλουτισμό των λιστών γνωστών ονομάτων ενός λεξικού. Το σύστημα αναγνώρισης ονομάτων οντοτήτων που θα χρησιμοποιηθεί μπορεί να περιλαμβάνει ένα *αρχικό ή ενδεικτικό (seed)* λεξικό, ένα λεξικό από μια άλλη θεματική περιοχή, ή να μην περιλαμβάνει καθόλου λεξικό, αφήνοντας το βάρος της αναγνώρισης ονομάτων οντοτήτων αποκλειστικά στο υποσύστημα της γραμματικής. Απαραίτητη προϋπόθεση για τον σωστό εμπλουτισμό ενός λεξικού με νέα ονόματα οντοτήτων, είναι το υποσύστημα της γραμματικής του αναγνωριστή ονομάτων οντοτήτων που θα χρησιμοποιηθεί να είναι κατάλληλο για την θεματική περιοχή που μας ενδιαφέρει, ώστε να εμφανίζει υψηλή *ακρίβεια (precision)* στον εντοπισμό των ονομάτων οντοτήτων της θεματικής περιοχής. Στην περίπτωση που το υποσύστημα της γραμματικής χρησιμοποιεί μηχανική μάθηση, συνίσταται η εκπαίδευση σε δεδομένα της ίδιας θεματικής περιοχής, για την οποία θέλουμε να προσαρμόσουμε/εμπλουτίσουμε το λεξικό. Σχηματικά, η προτεινόμενη προσέγγιση για τον εμπλουτισμό του υποσυστήματος του λεξικού παρουσιάζεται στην ακόλουθη εικόνα (Εικόνα 19):




**Εικόνα 19: Η αρχιτεκτονική του συστήματος προσαρμογής/εμπλουτισμού του υποσυστήματος του λεξικού.**

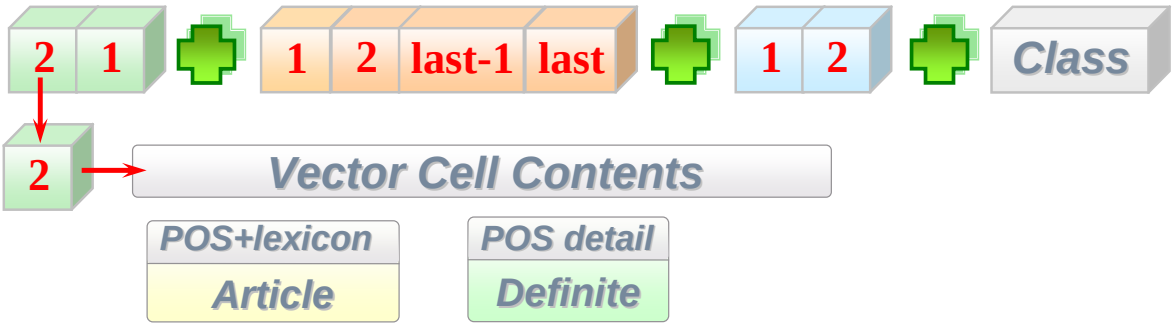
Η διαδικασία εμπλουτισμού περιλαμβάνει ένα σύστημα αναγνώρισης ονομάτων οντοτήτων, το οποίο είναι ικανό να αναγνωρίζει ονόματα οντοτήτων από την θεματική περιοχή που μας ενδιαφέρει με ικανοποιητική ακρίβεια, το λεξικό του οποίου εμπίπτει σε μια από τις ακόλουθες τρεις κατηγορίες:

- Εκτενές λεξικό για μια διαφορετική θεματική περιοχή: σε αυτήν την περίπτωση το λεξικό θα εμπλουτιστεί με ονόματα οντοτήτων από την θεματική περιοχή που μας ενδιαφέρει.
- Ενδεικτικό λεξικό: το λεξικό περιέχει έναν μικρό αριθμό από ονόματα γνωστών οντοτήτων από την θεματική περιοχή που μας ενδιαφέρει. Σε αυτήν την περίπτωση το λεξικό θα εμπλουτιστεί με ονόματα οντοτήτων από την θεματική περιοχή που μας ενδιαφέρει.
- Κενό λεξικό: σε αυτήν την περίπτωση θα δημιουργεί ένα νέο λεξικό με ονόματα γνωστών οντοτήτων από την θεματική περιοχή που μας ενδιαφέρει.

Έχοντας ένα σύστημα αναγνώρισης ονομάτων οντοτήτων για την θεματική περιοχή που μας ενδιαφέρει, και το οποίο χρησιμοποιεί το όποιο υπάρχον λεξικό, εφαρμόζουμε τον αναγνωριστή ονομάτων οντοτήτων σε κείμενα της θεματικής περιοχής, και συγκεντρώνουμε όλα τα αναγνωρισμένα ονόματα οντοτήτων, τα οποία μπορούν να εμπλουτίσουν το υπάρχον λεξικό, στην περίπτωση που κάποιο δεν υπάρχει ήδη στο λεξικό.

### 4.8.2 Πειραματική αξιολόγηση και αποτελέσματα

Για την πειραματική αξιολόγηση της προσαρμογής/εμπλουτισμού του υποσυστήματος του λεξικού, επιλέχθηκε ο αλγόριθμος μηχανικής μάθησης C4.5 [25] για να αντικαταστήσει το υποσύστημα της γραμματικής, και συγκεκριμένα η προσέγγιση της ενότητας 4.7.3, όπου η αναγνώριση ονομάτων οντοτήτων γίνεται σε επίπεδο φράσης. Ο αλγόριθμος C4.5 είναι ένας αλγόριθμος μηχανικής μάθησης που κατασκευάζει δέντρα αποφάσεων από τα δεδομένα εκπαίδευσης, τα οποία πρέπει να είναι κωδικοποιημένα σαν διανύσματα σταθερού μήκους. Η αναπαράσταση που επιλέξαμε είναι μια ελαφρά παραλλαγή της αναπαράστασης της ενότητας 4.7.3, με μια μικρή μείωση στον αριθμό των λέξεων της ονοματικής φράσης (από έξι σε τέσσερις) με ταυτόχρονη μείωση και του αριθμού λέξεων του περιβάλλοντος από τρεις λέξεις πριν και μετά το όνομα της οντότητας, σε δύο:




**Εικόνα 20: Η αναπαράσταση ενός διανύσματος του συστήματος προσαρμογής/εμπλουτισμού του υποσυστήματος του λεξικού.**

Όπως φαίνεται και από την Εικόνα 20, το διάνυσμα περιλαμβάνει πληροφορία σχετικά με τέσσερις λέξεις (τις δύο πρώτες και τις δύο τελευταίες λέξεις) ενός ονόματος οντότητας, καθώς και πληροφορία σχετικά με τις δύο λέξεις πριν και μετά το όνομα οντότητας, και φυσικά την επιθυμητή ετικέτα κατηγορίας (κλάση) στην περίπτωση δεδομένων εκπαίδευσης.

| Μέρος του λόγου | Ετικέτες επιπρόσθετης μορφολογικής πληροφορίας |
|---|---|
| Noun | singular, plural |
| Pronoun | 1st_singular, 2nd_singular, 3rd_singular, 3rd_singular_masculine, 3rd_singular_feminine, 3rd_singular_neuter, 1st_plural, 2nd_plural, 3rd_plural |
| Adjective, Adverb | positive_form, comparative_form, superlative_form |
| Verb | present_participle, simple_present, simple_present_1st_singular, simple_present_2nd_singular, simple_present_3rd_singular, simple_present_1st_plural, simple_present_2nd_plural, simple_present_3rd_plural, simple_present_plural, simple_present_negative_form, simple_present_passive_form, simple_present_passive_form_negative_form, present_perfect, present_perfect_negative_form, present_perfect_passive_form, present_perfect_passive_form_negative_form, present_progressive, present_progressive_negative_form, present_perfect_progressive, present_perfect_progressive_negative_form, present_conditional, present_conditional_negative_form, past_participle, simple_past, simple_past_1st_singular, simple_past_2nd_singular, simple_past_3rd_singular, simple_past_plural, simple_past_negative_form, simple_past_passive_form, simple_past_passive_form_negative_form, past_continuous, past_continuous_negative_form, past_perfect, past_perfect_passive_form, past_perfect_negative_form, past_perfect_passive_form_negative_form, base_form, future, future_passive_form, future_negative_form, future_passive_form_negative_form, perfect_conditional, perfect_conditional_negative_form |
| Punctuation | open, close |
| Article | definite, indefinite_singular |
| All Other Categories | Invariable |

**Πίνακας 26: Οι ετικέτες επιπρόσθετης μορφολογικής πληροφορίας για κάθε μέρος του λόγου.**

Η πληροφορία που περιέχεται για κάθε λέξη (είτε του ονόματος οντότητας είτε του περιβάλλοντός της) αποτελείται από δύο χαρακτηριστικά, τα οποία περιέχουν συμβολικές τιμές. Το πρώτο χαρακτηριστικό αντιπροσωπεύει το μέρος του λόγου της λέξης (π.χ. επίθετο, ουσιαστικό, ρήμα, αντωνυμία, κλπ.), επεκταμένο με τις ετικέτες από το υπάρχον λεξικό (π.χ. όνομα προσώπου, τοποθεσίας, ή οργανισμού), αν αυτό υπάρχει. Στην περίπτωση όπου μια η περισσότερες ετικέτες από το λεξικό είναι διαθέσιμες, οι ετικέτες αυτές συγχωνεύονται με τις ετικέτες των μερών του λόγου, δημιουργώντας μια νέα ετικέτα.

Το δεύτερο χαρακτηριστικό αντιπροσωπεύει επιπρόσθετες μορφολογικές πληροφορίες αν αυτές υπάρχουν, οι οποίες εξαρτώνται από το μέρος του λόγου της λέξης. Οι επιπρόσθετες μορφολογικές πληροφορίες περιλαμβάνουν τον αριθμό για τα ουσιαστικά και τα επίθετα, το χρόνο, το πρόσωπο και διάθεση για τα ρήματα, το πρόσωπο για τις

αντωνυμίες κλπ. (Το πλήρες σύνολο των ετικετών για κάθε μέρος του λόγου, εμφανίζεται στον πίνακα (Πίνακας 26)). Σημειώνεται ότι η ίδια η λέξη (είτε το λήμμα ή η ρίζα της) δεν συμπεριλαμβάνεται στο διάνυσμα χαρακτηριστικών.

Η πειραματική αξιολόγηση αφορά την Αγγλική γλώσσα, χρησιμοποιώντας σαν σώμα κειμένων μέρος του σώματος κειμένων "Wall Street Journal – WSJ" [90]. Το σώμα κειμένων WSJ περιέχει έγγραφα που καλύπτουν ένα ευρύ σύνολο θεματικών περιοχών. Το υποσύνολο του σώματος κειμένων που χρησιμοποιήθηκε για την πειραματική αξιολόγηση περιελάμβανε οχτώ τύπους ονομάτων οντοτήτων, από τους οποίους επιλέξαμε τρεις κατηγορίες ονομάτων οντοτήτων για την πειραματική αξιολόγηση (ονόματα προσώπων, τοποθεσιών, οργανισμών). Το σώμα κειμένων προ-επεξεργάστηκε με τη βοήθεια του συστήματος εξαγωγής πληροφορίας VIE [84]. Το σύστημα VIE αποτελεί ένα συμβατικό σύστημα εξαγωγής πληροφορίας, το οποίο αποτελείται από ένα λεξικό γνωστών ονομάτων οντοτήτων, και μια χειρωνακτικά κατασκευασμένη γραμματική για την αναγνώριση ονομάτων οντοτήτων. Με την βοήθεια του συστήματος VIE αναγνωρίστηκαν και επισημειώθηκαν τα ονόματα οντοτήτων στο σώμα κειμένων, από τις τρεις κατηγορίες που σχετίζονται με την αξιολόγηση (ονόματα προσώπων, τοποθεσιών, οργανισμών). Για κάθε ένα όνομα οντότητας δημιουργήθηκε ένα διάνυσμα από το κείμενο στο οποίο αυτό βρέθηκε. Ένα παράδειγμα τέτοιου διανύσματος φαίνεται στον πίνακα (Πίνακας 28), ενώ συνολικά δημιουργήθηκαν διανύσματα για 8341 μοναδικά ονόματα οντοτήτων που εντοπίστηκαν στο σώμα κειμένων από το σύστημα εξαγωγής πληροφορίας VIE. Τα διανύσματα που σχετίζονται με τα μισά περίπου ονόματα οντοτήτων χρησιμοποιήθηκαν σαν δεδομένα εκπαίδευσης, ενώ τα υπόλοιπα σαν δεδομένα αποτίμησης. Επίσης, τα ονόματα οντοτήτων των δεδομένων εκπαίδευσης χρησιμοποιήθηκαν σαν αρχικό λεξικό.

| **Δεδομένα εκπαίδευσης** | | **Δεδομένα αξιολόγησης** | |
|---|---|---|---|
| Πρόσωπα | 2397 | Πρόσωπα | 2390 |
| Οργανισμοί | 1460 | Οργανισμοί | 1462 |
| Τοποθεσίες | 316 | Τοποθεσίες | 316 |
| Σύνολο | 4173 | Σύνολο | 4168 |

**Πίνακας 27: Κατανομή των ονομάτων οντοτήτων στις τρεις σημασιολογικές κατηγορίες, για τα δεδομένα εκπαίδευσης και αξιολόγησης.**

Για την διαδικασία της αποτίμησης, όλα τα στιγμιότυπα των ονομάτων οντοτήτων που περιέχονται στο αρχικό λεξικό (το οποίο περιέχει τα ονόματα οντοτήτων από τα δεδομένα εκπαίδευσης του πίνακα (Πίνακας 27)) εντοπίστηκαν στο κείμενο. Για κάθε στιγμιότυπο, παρήχθησαν μια σειρά διανυσμάτων χαρακτηριστικών, όπως τα παραδείγματα που εμφανίζονται στον παραπάνω πίνακα (Πίνακας 28). Στην περίπτωση της απουσίας χαρακτηριστικών (που συνήθως οφείλεται στην απουσία μιας λέξης στη συγκεκριμένη θέση) χρησιμοποιείται ο χαρακτήρας "?", ο οποίος ερμηνεύεται από τον αλγόριθμο C4.5 σαν έλλειψη πληροφορίας. Στην περίπτωση ασάφειας, είτε στο σημασιολογικά εμπλουτισμένο μέρος του λόγου, ή στην επιπρόσθετη μορφολογικές πληροφορία, δημιουργούνται περισσότερα από ένα διανύσματα χαρακτηριστικών, προκειμένου να καλυφθούν όλοι οι πιθανοί συνδυασμοί των διφορούμενων ετικετών για κάθε χαρακτηριστικό.

Πρόταση: Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.
Όνομα οντότητας (από το σύστημα VIE): Pierre Vinken
Κατηγορία ονόματος οντότητας: person (όνομα προσώπου)
Χαρακτηριστικά που αποδόθηκαν από το στάδιο της γλωσσικής προ-επεξεργασίας

| Λέξη | Μέρος του λόγου + ετικέτα λεξικού | Ετικέτες επιπρόσθετης μορφολογικής πληροφορίας |
|---|---|---|
| Pierre | proper_noun | singular |
| | proper_noun_city | invariable |
| Vinken | ? | ? |
| , | punct_comma | invariable |
| 61 | adjective_cardinal_number | invariable |
| | common_noun | invariable |

Παραγόμενα διανύσματα χαρακτηριστικών:
Σειρά λέξεων: δεύτερη λέξη στα αριστερά (του ονόματος οντότητας), πρώτη λέξη αριστερά, πρώτη λέξη (του ονόματος οντότητας), δεύτερη λέξη, προ-τελευταία λέξη, τελευταία λέξη, πρώτη λέξη δεξιά, δεύτερη λέξη δεξιά.
[Δύο χαρακτηριστικά για κάθε μία από τις οκτώ λέξεις, συν την κατηγορία του ονόματος οντότητας.]

1) ?,?,?,?,proper_noun_city,invariable,?,?,proper_noun_city,invariable,?,?,punct_comma,invariable, adjective_cardinal_number,invariable,person.
2) ?,?,?,?,proper_noun_city,invariable,?,?,proper_noun_city,invariable,?,?,punct_comma,invariable, common_noun,invariable,person.
3) ?,?,?,?,proper_noun,singular,?,?,proper_noun,singular,?,?,punct_comma,invariable, adjective_cardinal_number,invariable,person.
4) ?,?,?,?,proper_noun,singular,?,?,proper_noun,singular,?,?,punct_comma,invariable,common_noun, invariable,person.

**Πίνακας 28: Παράδειγμα παραγόμενων διανυσμάτων από ένα στιγμιότυπο ονόματος οντότητας σε ένα κείμενο.**

Στην συνέχεια, ο αλγόριθμος μηχανικής μάθησης εκπαιδεύθηκε χρησιμοποιώντας όλα τα διανύσματα χαρακτηριστικών που δημιουργήθηκαν από τα ονόματα οντοτήτων του αρχικού λεξικού. Στην συνέχεια, το δέντρο αποφάσεων που δημιουργήθηκε χρησιμοποιήθηκε για την κατηγοριοποίηση των διανυσμάτων που αφορούν ονόματα οντοτήτων που περιλαμβάνονται στο σύνολο των δεδομένων αξιολόγησης. Η κατηγοριοποίηση κάθε διανύσματος χαρακτηριστικών συνοδεύεται από έναν βαθμό εμπιστοσύνης, με τιμή από το διάστημα $[0 \cdots 1]$, με το επίπεδο εμπιστοσύνης στο αποτέλεσμα της κατηγοριοποίησης από τον αλγόριθμο C4.5 να αυξάνει καθώς αυτός ο βαθμός εμπιστοσύνης αυξάνεται. Έχοντας σαν στόχο τον εμπλουτισμό ενός λεξικού, είναι επιθυμητή μια προτίμηση προς την υψηλή ακρίβεια όσον αφορά την απόδοση του συστήματος προσαρμογής/εμπλουτισμού του υποσυστήματος του λεξικού. Συνεπώς, επιλέχθηκε ένα *κατώφλι (threshold)*, το οποίο προσδιορίστηκε εμπειρικά για το συγκεκριμένο πείραμα στην τιμή 0.7, κάτω από το οποίο η κατηγοριοποίηση μετατρεπόταν στην ετικέτα «άγνωστο». Το αποτέλεσμα της κατηγοριοποίησης και του φιλτραρίσματος μέσω του επιλεγμένου κατωφλιού, ήταν τα διανύσματα αξιολόγησης κατηγοριοποιημένα σε τέσσερεις σημασιολογικές κατηγορίες (όνομα προσώπου, οργανισμού, τοποθεσίας, ή «άγνωστο» όνομα). Σε περιπτώσεις όπου διανύσματα που προέκυψαν από το ίδιο όνομα οντότητας (είτε λόγω αμφισημίας στα χαρακτηριστικά, είτε λόγω διαφορετικής σημασιολογικής κατηγορίας σε συγκεκριμένα έγγραφα) έλαβαν διαφορετικές κατηγοριοποιήσεις, επιλέχθηκε η συχνότερη κατηγοριοποίηση.

| Κατηγορία Ονομάτων | Αριθμός ονομάτων προς αναγνώριση | Σωστά Αναγνωρισθέντα ονόματα | Κατηγοριοποιημένα σαν «άγνωστο» | Ακρίβεια | Ανάκληση | F-measure |
|---|---|---|---|---|---|---|
| Πρόσωπο | 2390 | 1767 | 521 | 98.33 % | 73.93 % | 84.40 % |
| Οργανισμός | 1462 | 1005 | 384 | 91.36 % | 68.74 % | 78.45 % |
| Τοποθεσία | 316 | 268 | 30 | 79.76 % | 84.81 % | 82.21 % |
| Όλες οι κατηγορίες | 4168 | 3040 | 935 | 94.03 % | 72.94 % | 82.15 % |

**Πίνακας 29: Αποτελέσματα (ανά τύπο οντότητας) του συστήματος προσαρμογής/εμπλουτισμού του υποσυστήματος του λεξικού για την Αγγλική γλώσσα**

Τα αποτελέσματα της αξιολόγησης παρουσιάζονται στον πίνακα (Πίνακας 29). Όπως φαίνεται από τον πίνακα, 3233 από τα 4168 ονόματα οντοτήτων του συνόλου αξιολόγησης κατηγοριοποιήθηκαν, ενώ 935 ονόματα οντοτήτων έμειναν αταξινόμητα. Από αυτά που κατηγοριοποιήθηκαν, 3040 κατηγοριοποιήθηκαν σωστά. Η χρήση ενός υψηλού κατωφλιού (0.7) οδήγησε σε υψηλή ακρίβεια (94.03%), με την ανάκληση να βρίσκεται επίσης σε αποδεκτό επίπεδο (72.94%). Αξίζει να σημειωθεί ότι αρκετά από τα 935 ονόματα οντοτήτων που έμειναν αταξινόμητα από την προσέγγιση λόγω του κατωφλιού, τα 477 εμφανίζονται μόνο μία φορά στο σώμα κειμένων που χρησιμοποιήθηκε για την αξιολόγηση, καθιστώντας τον σωστό εντοπισμό τους εξαιρετικά δύσκολο.

## 4.9 Ενημέρωση συστήματος αναγνώρισης ονομάτων οντοτήτων

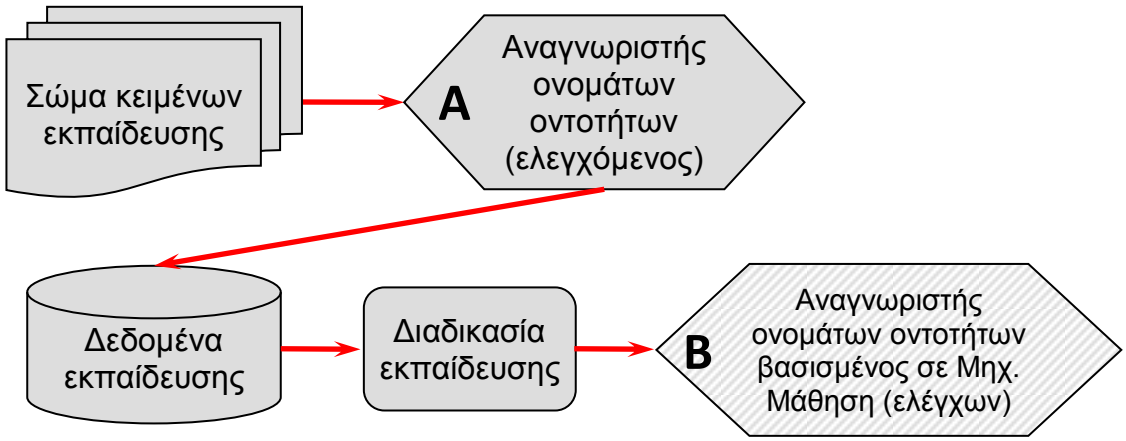
Το τρίτο και τελευταίο σενάριο χρήσης της μηχανικής μάθησης, αφορά την ανάγκη ενημέρωσης ενός συστήματος αναγνώρισης ονομάτων οντοτήτων. Αρκετές θεματικές περιοχές μεταβάλλονται ελαφρά με την πάροδο του χρόνου, όχι τόσο από την μεταβολή του τρόπου γραφής (π.χ. από την μεταβολή των δημοσιογράφων που συντάσσουν ειδησεογραφικά άρθρα) αλλά κυρίως από την εισαγωγή νέων ονομάτων οντοτήτων και την μεταβολή της σημασίας των υπαρχόντων ονομάτων οντοτήτων. Έχει παρατηρηθεί ότι η απόδοση των συστημάτων αναγνώρισης ονομάτων οντοτήτων υποβιβάζεται με την πάροδο του χρόνου ( [91], [92]), οπότε μια προσέγγιση που να μπορεί να σηματοδοτεί πότε ένα σύστημα αναγνώρισης ονομάτων οντοτήτων πρέπει να ενημερωθεί είναι επιθυμητή. Μια τέτοια προσέγγιση προτείνεται στην παρούσα ενότητα, βασισμένη στην ιδέα της απόκλισης μεταξύ δύο συστημάτων αναγνώρισης ονομάτων οντοτήτων με την πάροδο του χρόνου, τα οποία είχαν ενημερωθεί την ίδια χρονική στιγμή στο παρελθόν για να επεξεργάζονται κείμενα της ίδιας θεματικής περιοχής.

### 4.9.1 Η προσέγγιση

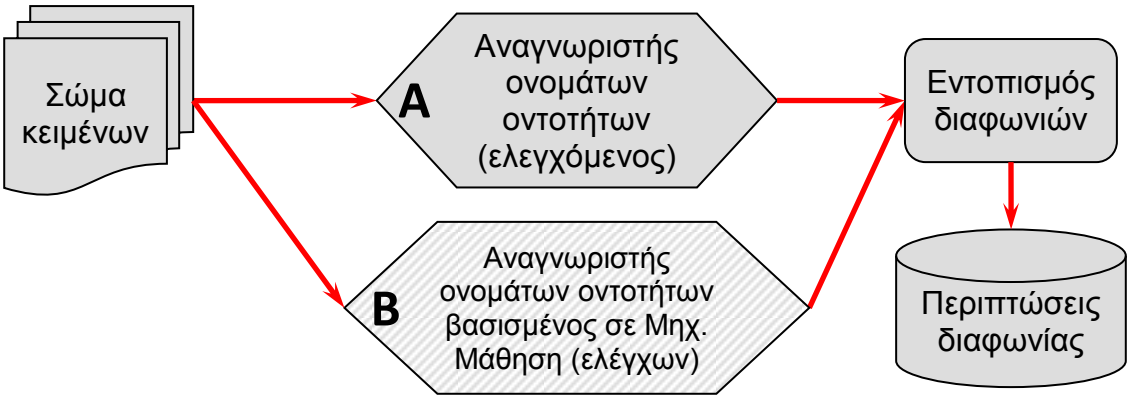
Η προτεινόμενη προσέγγιση για την ανάγκη ενημέρωσης ενός συστήματος αναγνώρισης ονομάτων οντοτήτων βασίζεται σε μια εναλλακτική και καινοτόμο χρήση της μηχανικής μάθησης για την κατασκευή αναγνωριστών ονομάτων οντοτήτων. Αντί για την χρήση μηχανικής μάθησης για την κατασκευή ενός αναγνωριστή ονομάτων οντοτήτων που θα χρησιμοποιηθεί αυτόνομα, χρησιμοποιείται για την κατασκευή ενός αναγνωριστή ονομάτων οντοτήτων που χρησιμοποιείται για να ελέγξει την απόδοση ενός υπάρχοντος αναγνωριστή ονομάτων οντοτήτων (είτε συμβατικού, ή βασισμένου σε μηχανική μάθηση), μέσω τις απόκλισης στην απόδοση μεταξύ των δύο αναγνωριστών.

Η προτεινόμενη προσέγγιση περιλαμβάνει δύο στάδια: το στάδιο εκπαίδευσης, κατά τη διάρκεια του οποίου ένας αναγνωριστής ονομάτων οντοτήτων βασισμένος σε μηχανική μάθηση, τουλάχιστον για το υποσύστημα της γραμματικής, προσαρμόζεται στα αποτελέσματα που παράγει ο προς έλεγχο αναγνωριστής, και το στάδιο ελέγχου, όπου

οι δύο αναγνωριστές ονομάτων οντοτήτων εφαρμόζονται στα ίδια έγγραφα και συγκρίνεται η έξοδός τους. Το διάγραμμα ροής της προτεινόμενης προσέγγισης εμφανίζεται στα ακόλουθα διαγράμματα, με το στάδιο εκπαίδευσης να εμφανίζεται στην Εικόνα 21, και το στάδιο ελέγχου να εμφανίζεται στην Εικόνα 22:




**Εικόνα 21: Η αρχιτεκτονική του συστήματος σηματοδότησης ανάγκης ενημέρωσης, για το στάδιο της εκπαίδευσης.**




**Εικόνα 22: Η αρχιτεκτονική του συστήματος σηματοδότησης ανάγκης ενημέρωσης, για το στάδιο του ελέγχου.**

### Σηματοδότηση ανάγκης ενημέρωσης: στάδιο εκπαίδευσης

Το στάδιο εκπαίδευσης έχει σαν στόχο την εξαγωγή όσο περισσότερης γνώσης είναι εφικτό από το σύστημα αναγνώρισης ονομάτων οντοτήτων (σύστημα Α), την απόδοση του οποίου επιθυμούμε να παρακολουθήσουμε (ελέγξουμε). Αυτό επιτυγχάνεται μέσω της επισημείωσης ενός ευμεγέθους σώματος κειμένων (σώμα κειμένων εκπαίδευσης) από το σύστημα αναγνώρισης ονομάτων οντοτήτων. Από το επισημειωμένο σώμα κειμένων εκπαίδευσης δημιουργούνται τα δεδομένα εκπαίδευσης του συστήματος αναγνώρισης ονομάτων οντοτήτων που βασίζεται σε μηχανική μάθηση (σύστημα Β), το οποίο θα ελέγχει το βασικό σύστημα αναγνώρισης ονομάτων οντοτήτων. Έχοντας τα δεδομένα εκπαίδευσης, το σύστημα αναγνώρισης ονομάτων οντοτήτων που βασίζεται σε μηχανική μάθηση (Β) εκπαιδεύεται, ώστε να προσαρμοστεί στην θεματική περιοχή του αρχικού συστήματος (Α). Αξίζει να σημειωθεί ότι το αρχικό σύστημα (Α) μπορεί να είναι ένα συμβατικό σύστημα αναγνώρισης ονομάτων οντοτήτων (π.χ. που δεν χρησιμοποιεί μηχανική μάθηση). Στην περίπτωση που το αρχικό σύστημα βασίζεται σε

μηχανική μάθηση, θα πρέπει να είναι αρκετά διαφορετικό από το σύστημα ελέγχου (σύστημα Β). Η προτεινόμενη προσέγγιση βασίζεται στο ότι τα δύο συστήματα είναι αρκούντως διαφορετικά (είτε στην επιλογή των εμπλεκόμενων αλγορίθμων μηχανικής μάθησης, ή στην επιλογή των αναπαραστάσεων), ώστε να βασίζονται σε διαφορετικά μοντέλα για την εργασία της αναγνώρισης ονομάτων οντοτήτων. Η χρήση διαφορετικών μοντέλων από τα δύο συστήματα καθιστά πιο πιθανή την απόκλιση των δύο συστημάτων μεταξύ τους στην περίπτωση που η θεματική περιοχή μεταβληθεί λόγω της παρόδου του χρόνου.

**Σηματοδότηση ανάγκης ενημέρωσης: στάδιο ελέγχου**

Το στάδιο ελέγχου έχει σαν στόχο την σύγκριση των αποτελεσμάτων των δύο συστημάτων αναγνώρισης ονομάτων οντοτήτων σε ένα νέο σώμα κειμένων, και τον εντοπισμό των διαφωνιών μεταξύ των δύο συστημάτων. Παρά το γεγονός ότι το σύστημα ελέγχου (σύστημα Β) έχει προσαρμοστεί στα αποτελέσματα του αρχικού συστήματος (Α), η διαφορετική φύση των δύο συστημάτων (π.χ. τα διαφορετικά μοντέλα αναγνώρισης και κατηγοριοποίησης ονομάτων οντοτήτων του υποσυστήματος της γραμματικής) θα οδηγήσουν σε διαφοροποίηση των αποτελεσμάτων, αν τα δύο συστήματα εφαρμοστούν σε ένα νέο σώμα κειμένων. Τα βήματα αυτού του σταδίου είναι τα ακόλουθα (Εικόνα 22):

1. Το αρχικό σύστημα αναγνώρισης ονομάτων οντοτήτων (σύστημα Α) εφαρμόζεται σε ένα νέο σώμα κειμένων. Αξίζει να σημειωθεί ότι τα έγγραφα αυτού του νέου σώματος κειμένων πρέπει να διαφέρουν με κάποιο χαρακτηριστικό τρόπο από τα έγγραφα του σώματος εκπαίδευσης. Την προτεινόμενη προσέγγιση την ενδιαφέρουν έγγραφα που διαφέρουν χρονολογικά, π.χ. τα έγγραφα να είναι πιο πρόσφατα χρονικά από εκείνα που απαρτίζουν το σώμα κειμένων εκπαίδευσης.
2. Εφαρμογή του συστήματος αναγνώρισης ονομάτων οντοτήτων ελέγχου (σύστημα Β) στο νέο σώμα κειμένων.
3. Σύγκριση των αποτελεσμάτων των δύο συστημάτων αναγνώρισης ονομάτων οντοτήτων (Α και Β) μεταξύ τους, με σκοπό των εντοπισμό των διαφωνιών.

Το αποτέλεσμα είναι ένα σύνολο ονομάτων οντοτήτων όπου τα δύο συστήματα διαφωνούν. Οι διαφωνίες έχουν διττό ρόλο: μπορούν να σηματοδοτήσουν την ανάγκη ενημέρωσης του αρχικού συστήματος (σύστημα Α), αν το ποσοστό τους είναι σημαντικό στο σύνολο των αναγνωρισμένων ονομάτων οντοτήτων, αλλά ταυτόχρονα μπορούν να καθοδηγήσουν και την διαδικασία της αναβάθμισης, καθώς οι περιπτώσεις διαφωνίας μπορούν να χρησιμοποιηθούν για τον συντονισμό του συστήματος.

### 4.9.2 Πειραματική αξιολόγηση και αποτελέσματα

Για την πειραματική αξιολόγηση της προσαρμογής/εμπλουτισμού του υποσυστήματος του λεξικού, επιλέχθηκε ο αλγόριθμος μηχανικής μάθησης C4.5 [25] για να αντικαταστήσει το υποσύστημα της γραμματικής, και συγκεκριμένα η προσέγγιση της ενότητας 4.7.3, όπου η αναγνώριση ονομάτων οντοτήτων γίνεται σε επίπεδο φράσης. Ο αλγόριθμος C4.5 είναι ένας αλγόριθμος μηχανικής μάθησης που κατασκευάζει δέντρα αποφάσεων από τα δεδομένα εκπαίδευσης, τα οποία πρέπει να είναι κωδικοποιημένα σαν διανύσματα σταθερού μήκους. Σαν αναπαράσταση χρησιμοποιήθηκε το συμβολικό διάνυσμα χαρακτηριστικών που περιγράφεται στην ενότητα 4.6.4. Για την πειραματική αξιολόγηση χρησιμοποιήθηκαν δύο σώματα κειμένων, από την ίδια θεματική περιοχή

(οικονομικές ειδήσεις), που παραχωρήθηκαν από την εταιρία Καπα TEL[3]. Το πρώτο σώμα κειμένων (σώμα κειμένων εκπαίδευσης) αποτελείται από 5.000 άρθρα ειδήσεων από τα έτη 1996 και 1997, και περιέχουν 10.010 *στιγμιότυπα (instances)* ονομάτων οντοτήτων (1.885 ονόματα προσώπων, 1.781 ονόματα τοποθεσιών, 6.344 ονόματα οργανισμών). Το δεύτερο σώμα κειμένων που σχετίζεται με το στάδιο ελέγχου, προέρχεται επίσης από την εταιρία Καπα TEL, αποτελείται από 5.779 ειδήσεις από τα έτη 1999 και 2000, και περιέχει 11.786 στιγμιότυπα ονομάτων οντοτήτων (1.137 ονόματα προσώπων, 810 ονόματα τοποθεσιών, 9.839 ονόματα οργανισμών). Σαν σύστημα αναγνώρισης ονομάτων οντοτήτων το οποίο ελέγχεται από την προσέγγιση (σύστημα Α), χρησιμοποιείται το σύστημα MITOS [16], [83] (το οποίο αποτελεί βελτίωση του συστήματος που παρουσιάζεται στο ΠΑΡΑΡΤΗΜΑ ΙΙ με μεγαλύτερο λεξικό και περισσότερους γραμματικούς κανόνες), το οποίο έχει αξιολογηθεί στην ίδια θεματική περιοχή στην ενότητα 4.6.6 (Πίνακας 14).

Ένας καλός τρόπος για να δοθεί μια επισκόπηση των περιπτώσεων της διαφωνίας των δύο συστημάτων είναι μέσω ενός *πίνακα συνάφειας (contingency table)*, όπως φαίνεται στον ακόλουθο πίνακα (Πίνακας 30). Οι σειρές του πίνακα αυτού αντιστοιχούν στα αποτελέσματα του χειρωνακτικά κατασκευασμένου συστήματος MITOS (σύστημα Α), ενώ οι στήλες αφορούν το σύστημα ελέγχου (σύστημα Β), που βασίζεται στον αλγόριθμο μηχανικής μάθησης C4.5.

| | Ονόματα οργανισμών | Ονόματα προσώπων | Ονόματα τοποθεσιών |
|---|---|---|---|
| Ονόματα οργανισμών | 9,906 | 250 | 32 |
| Ονόματα προσώπων | 230 | 649 | 14 |
| Ονόματα τοποθεσιών | 24 | 6 | 675 |

**Πίνακας 30: Αποτελέσματα αξιολόγησης του συστήματος ενημέρωσης αναγνωριστή ονομάτων οντοτήτων.**

Όπως φαίνεται και από τον πίνακα, στο 95% των περιπτώσεων τα δύο συστήματα βρίσκονται σε συμφωνία. Αυτό σημαίνει, ότι προκειμένου να ενημερωθεί το σύστημα Α, πρέπει να εξεταστεί η συμπεριφορά του μόνο στο 5% των περιπτώσεων, όπου τα δύο συστήματα διαφωνούν. Η εξέταση αυτών των περιπτώσεων μπορεί να οδηγήσει στον εντοπισμό προβλημάτων του συστήματος Α, μερικά παραδείγματα των οποίων παρουσιάζουμε στις ακόλουθες παραγράφους.

### Αναγνώριση ονομάτων οντοτήτων

Η εξέταση των περιπτώσεων διαφωνιών αποκάλυψε μερικά ενδιαφέροντα προβλήματα σχετικά με τον εντοπισμό των ονομάτων οντοτήτων του συστήματος Α (σύστημα αναγνώρισης ονομάτων οντοτήτων MITOS). Τα προβλήματα αυτά αφορούν κυρίως την μερική αναγνώριση κάποιων ονομάτων οντοτήτων, που οδήγησε σε ανακριβή αποτελέσματα στις περιπτώσεις αυτές. Παραδείγματος χάριν, στο στάδιο της *αρχικής οριοθέτησης* των ονομάτων οντοτήτων *(delimitation)*, οι κανονικές γραμματικής αδυνατούν να προσδιορίσουν ονόματα οντοτήτων που περιέχουν αριθμούς στα ονόματά τους, όπως τον οργανισμό «Αθήνα 2004», ο οποίος αντιπροσωπεύει την

[3] Καπα TEL: http://www.kapatel.gr/

οργανωτική επιτροπή των Ολυμπιακών Αγώνων του 2004. Επιπλέον, μερικοί από τους γραμματικούς κανόνες δεν έχουν λάβει υπόψη μερικές κλιτικές καταλήξεις, προκαλώντας λάθη στον εντοπισμό των λέξεων που περιέχονται στο όνομα μιας οντότητας. Παραδείγματος χάριν, στη φράση «ο υφ. Πολιτισμού Γ. Φλωρίδης» οι γραμματικοί κανόνες απέτυχαν να χωρίσουν το όνομα του προσώπου από τον τίτλο του, αναγνωρίζοντας σαν όνομα προσώπου την φράση «Πολιτισμού Γ. Φλωρίδης», μη εξετάζοντας την πτώση όλως των λέξεων που συμμετέχουν στο όνομα της οντότητας. Τέλος, προστέθηκαν νέοι κανόνες για κάποιες λέξεις που λειτουργούν σαν προσδιοριστές εταιριών, όπως η λέξη «γραμμών». Ονόματα οντοτήτων που περιέχουν αυτή την λέξη κατηγοριοποιούνταν συνήθως σαν ονόματα οργανισμών (δεδομένου ότι η λέξη «γραμμών» είναι ένα συχνό συστατικό αεροπορικών ή ακτοπλοϊκών εταιριών), αλλά με την πάροδο του χρόνου άρχισε να εμφανίζεται και σε οντότητες άλλων τύπων, όπως «γραμμών ISDN» (το οποίο αναγνωρίστηκε σαν όνομα οργανισμού από το σύστημα Α).

### Κατηγοριοποίηση ονομάτων οντοτήτων

Εκτός από τα προβλήματα της φάσης αναγνώρισης, η εξέταση των περιπτώσεων διαφωνίας αποκάλυψε προβλήματα και στην φάση κατηγοριοποίησης των ονομάτων οντοτήτων. Αρκετοί από τους γραμματικούς κανόνες ταξινόμησης βρέθηκαν να είναι γενικοί, οδηγώντας στις λανθασμένες κατηγοριοποιήσεις. Παραδείγματος χάριν, σύμφωνα με έναν από τους κανόνες μια ακολουθία δύο λέξεων, αρχίζοντας από τα κεφαλαία γράμματα, αποτελεί ένα όνομα προσώπου εάν προηγείται ένα οριστικό άρθρο, και οι καταλήξεις αυτών των δύο λέξεων ανήκουν σε ένα συγκεκριμένο σύνολο που δείχνουν συνήθως τα ονόματα προσώπων. Αυτός ο κανόνας προκάλεσε την κατηγοριοποίηση διαφόρων φράσεων που δεν είναι ονόματα οντότητας σαν ονόματα προσώπων, όπως για παράδειγμα την φράση «του Ολυμπιακού Χωριού». Ένα άλλο παράδειγμα ενός υπερβολικά γενικού κανόνα είναι η κατηγοριοποίηση σαν όνομα οργανισμού μιας ακολουθία συντμήσεων ή ουσιαστικών που ξεκινούν από κεφαλαίο γράμμα, εάν αυτή η ακολουθία προηγείται από ένα κόμμα και πριν από αυτό προηγείται ένα άλλο όνομα οργανισμού. Αυτός ο κανόνας προκάλεσε την κατηγοριοποίηση κάποιων ονομάτων προσώπων ως ονόματα οργανισμών, όπως στην περίπτωση της φράσης «ο διοικητής της Εθνικής Τράπεζας, Θ.Καρατζάς».

## 4.10 Συνεισφορά

Σε αυτό το κεφάλαιο εξετάστηκε το πρόβλημα της αναγνώρισης ονομάτων οντοτήτων από κειμενικά δεδομένα. Στο πλαίσιο αυτής της διατριβής εξετάστηκαν διάφορες προσεγγίσεις: έχοντας σαν αναφορά της δυσκολίες ανάπτυξης και προσαρμογής ενός κλασσικού συστήματος αναγνώρισης ονομάτων οντοτήτων (ΠΑΡΑΡΤΗΜΑ ΙΙ), παρουσιάστηκαν προσεγγίσεις που βασίζονται σε μηχανική μάθηση, εξετάζοντας διαφορετικούς αλγόριθμους μηχανικής μάθησης, συνθέσεις διαφορετικών αλγορίθμων μηχανικής μάθησης, αλλά και διαφορετικές προσεγγίσεις όσον αφορά την αναπαράσταση των δεδομένων. Το χειρωνακτικά κατασκευασμένο σύστημα που δημιουργήθηκε στο πλαίσιο της παρούσας διατριβής, εκτός του ότι αποτέλεσε ένα από τα πρώτα συστήματα αναγνώρισης ονομάτων οντοτήτων για την Ελληνική γλώσσα, επιβεβαίωσε τα προβλήματα χειρωνακτικής κατασκευής συστημάτων: η απόδοση ενός χειρωνακτικού συστήματος εξαρτάται από την επένδυση εργασίας που θα γίνει σε αυτό, και η οποία αφορά μια συγκεκριμένη θεματική περιοχή. Ταυτόχρονα, η ανάπτυξη ενός αναγνωριστή ονομάτων οντοτήτων παρουσιάζεται δυσκολότερη για τη Ελληνική σε σχέση με την Αγγλική γλώσσα, αφού γλωσσικές δομές που βοηθούν στον εντοπισμό ονομάτων δεν χρησιμοποιούνται τόσο συχνά στα Ελληνικά.

Το επόμενο βήμα ήταν η αντικατάσταση του υποσυστήματος της γραμματικής (το οποίο αποτελεί και το δυσκολότερο στην χειρωνακτική κατασκευή υποσύστημα) με κάποιον αλγόριθμο μηχανικής μάθησης. Δοκιμάστηκαν δύο αρκετά διαφορετικοί αλγόριθμοι, ένας συμβολικός και ένας αριθμητικός, σε όσο το δυνατόν παραπλήσιες αναπαραστάσεις δεδομένων. Έχοντας επιλέξει τα δέντρα αποφάσεων και τα νευρωνικά δίκτυα ως αντιπροσώπους των συμβολικών και αριθμητικών οικογενειών αλγορίθμων, έγινε συγκριτική αξιόλογη των δύο προσεγγίσεων, τόσο στην Ελληνική όσο και στην Αγγλική γλώσσα, εμφανίζοντας αυξημένη προσαρμοστικότητα όχι μόνο σε διαφορετικές περιοχές, αλλά ακόμα και σε διαφορετικές γλώσσες.

Το κύριο πλεονέκτημα των συμβολικών μεθόδων έναντι των αριθμητικών, είναι η δυνατότητά τους να παραγάγουν αποτελέσματα που είναι ευκολότερα να κατανοηθούν από ανθρώπους, και να μετατραπούν σε περισσότερο «παραδοσιακές» αναπαραστάσεις, όπως είναι οι γραμματικές. Το γεγονός ότι οι γλωσσικές πληροφορίες είναι πρώτιστα συμβολικές, είναι ένας περαιτέρω λόγος για την χρήση συμβολικών αλγορίθμων. Από την άλλη, η δυνατότητα τα αποτελέσματα να γίνουν κατανοητά ή να μετασχηματιστούν σε άλλες αναπαραστάσεις δεν είναι πάντα σημαντική για πρακτικές εφαρμογές, όπου η απόδοση είναι μεγαλύτερης σπουδαιότητας. Με τη σύγκριση μιας συμβολικής και μιας αριθμητικής προσέγγισης, θελήσαμε να εξετάσουμε εάν η απώλεια στη δυνατότητα κατανόησης για τα νευρωνικά δίκτυα ισορροπείται από ένα αντίστοιχο κέρδος στην απόδοση ταξινόμησης. Τα αποτελέσματά μας δεν υποστηρίζουν αυτήν την υπόθεση, δεδομένου ότι και οι δύο μέθοδοι πέτυχαν παρόμοια απόδοση τα σε όλα πειράματά μας. Συγκεκριμένα, χρησιμοποιώντας την ίδια αριθμητική αναπαράσταση για την κωδικοποίηση των ονοματικών φράσεων, η διάκριση μεταξύ των δύο μεθόδων με βάση την απόδοσή τους είναι αρκετά πολύ δύσκολη.

Ιδιαίτερη έκπληξη αποτέλεσε το γεγονός ότι η απόδοση της συμβολικής μεθόδου βελτιώνεται με τη χρήση της αριθμητικής αναπαράστασης. Αυτό ήταν απροσδόκητο, όχι μόνο επειδή τα δέντρα απόφασης είναι εξ' αρχής σχεδιασμένα για να χρησιμοποιηθούν με συμβολικά χαρακτηριστικά, αλλά και επειδή η συγκεκριμένη αριθμητική αναπαράσταση που έχουμε χρησιμοποιήσει αγνοεί τη σειρά των λέξεων. Τα αποτελέσματά μας δείχνουν ότι αυτή η μείωση πληροφορίας απλοποιεί το στόχο εκμάθησης, βελτιώνοντας κατά συνέπεια την απόδοση της ταξινόμησης. Αυτό φαίνεται να ισχύει, τουλάχιστον για τα γλωσσικά χαρακτηριστικά (ετικέτες μέρους του λόγου και λεξικού) που έχουμε περιλάβει στην αναπαράσταση.

Δεδομένου ότι η εργασία της αναγνώρισης ονομάτων οντοτήτων μπορεί να διαιρεθεί σε δύο υπο-εργασίες, δηλαδή την αναγνώριση ονομάτων οντοτήτων και την μετέπειτα ταξινόμησή τους σε προκαθορισμένες κατηγορίες, εξετάσαμε τη συμπεριφορά και των δύο αλγορίθμων σε κάθε υπο-εργασία. Το κύριο συμπέρασμα αυτών των πειραμάτων είναι ότι και οι δύο μέθοδοι εκμάθησης επιτυγχάνουν υψηλή απόδοση στην συνολική εργασία του NERC, και η οποία είναι συγκρίσιμη με την απόδοση κατά την διαίρεση του προβλήματος σε δύο χωριστά υπο-προβλήματα.

Στο πλαίσιο της αντικατάστασης του υποσυστήματος της γραμματικής ενός αναγνωριστή ονομάτων οντοτήτων, εξετάστηκε και μια δεύτερη προσέγγιση (προσέγγιση Β), η οποία συνδυάζει διάφορους κατηγοριοποιητές (είτε με διαφορετικούς αλγορίθμους μηχανικής μάθησης, ή με διαφορετικές αναπαραστάσεις δεδομένων), οι οποίοι λειτουργούν και σε επίπεδο λέξεων και σε επίπεδο φράσεων. Η δεύτερη αυτή προσέγγιση αξιολογήθηκε τόσο σε διάφορες γλώσσες (Αγγλικά και Ελληνικά), όσο και σε διαφορετικές θεματικές περιοχές για την Ελληνική γλώσσα (περιγραφές προϊόντων φορητών υπολογιστών σε διαδικτυακά καταστήματα, αγγελίες προσφοράς εργασίας). Η αξιολόγηση έδειξε ότι η προτεινόμενη προσέγγιση εμφανίζει μια μικρή υπεροχή σε σχέση με άλλες προσεγγίσεις, όπως το σύστημα του Πανεπιστημίου του Εδιμβούργου

που αξιολογήθηκε στα ίδια κείμενα για την Αγγλική γλώσσα. Η απόδοση της προτεινόμενης προσέγγισης παρέμεινε σταθερή ανάμεσα στις θεματικές περιοχές που εξετάστηκαν, ενώ εμφανίζεται σημαντική αύξηση στην απόδοση όταν προστεθεί γνώση για την θεματική περιοχή, κυρίως στην ανάκληση του συστήματος αναγνώρισης ονομάτων οντοτήτων, μέσω του υποσυστήματος του λεξικού.

Στην συνέχεια, εξετάστηκε η αυτόματη προσαρμογή/εμπλουτισμός του υποσυστήματος του λεξικού, χρησιμοποιώντας ένα σύστημα αναγνώρισης ονομάτων οντοτήτων βασισμένο σε μηχανική μάθηση και ταυτόχρονα προσανατολισμένο να εμφανίζει μια προτίμηση προς την ακρίβεια έναντι της ανάκλησης. Η προτεινόμενη προσέγγιση που εξετάστηκε κατάφερε να αυξήσει ένα *αρχικό (seed)* λεξικό κατά 36 %, πετυχαίνοντας ακρίβεια 94.03 % και καταφέρνοντας να εντοπίσει το 72.94 % των ονομάτων οντοτήτων που περιέχονταν στο σώμα κειμένων αξιολόγησης και ταυτόχρονα δεν περιέχονταν στο λεξικό.

Τέλος, εξετάσαμε την ενημέρωση ενός συστήματος αναγνώρισης ονομάτων οντοτήτων, μέσω της κατασκευής ενός παράλληλου συστήματος, το οποίο επιτηρεί το αρχικό σύστημα μέσω του εντοπισμού των διαφορών τους. Η προσέγγιση αυτή χρησιμοποιεί την μηχανική μάθηση με έναν καινοτόμο τρόπο (σαν επιτηρητή ενός άλλου συστήματος), και το αποτέλεσμα της προσέγγισης μπορεί να χρησιμοποιηθεί για να προσδιοριστεί τόσο η χρονική στιγμή που απαιτείται η ενημέρωση του συστήματος (π.χ. βάζοντας ένα άνω όριο στις διαφορές μεταξύ των δύο συστημάτων, του συστήματος επιτήρησης και του επιτηρούμενου συστήματος), αλλά και να καθοδηγήσει την ενημέρωση του επιτηρούμενου συστήματος, εστιάζοντας στην βελτίωση του συστήματος στις περιπτώσεις που εμφανίζονται διαφωνίες.

# 5. Επαγωγική Εξαγωγή Γραμματικών: Ο Αλγόριθμος egGRIDS+

Το κεφάλαιο αυτό παρουσιάζει έναν νέο αλγόριθμο μηχανικής μάθησης, τον αλγόριθμο egGRIDS+, ο οποίος εντάσσεται στην κατηγορία των αλγορίθμων *επαγωγικής εξαγωγής γραμματικών (grammatical inference)*. Ο αλγόριθμος egGRIDS+ βασίζεται σε προγενέστερη έρευνα στον χώρο της επαγωγικής εξαγωγής γραμματικών, και κυρίως στον αλγόριθμο GRIDS [13], ο οποίος με την σειρά του βασίζεται σε προηγούμενη εργασία του Wolff [93], [94] και το σύστημα SNPR. Όπως και οι προκάτοχοί του, ο egGRIDS+ χρησιμοποιεί μια προτίμηση προς την *απλότητα (simplicity bias)*, ώστε να εξάγει επαγωγικά ανεξάρτητες από τα συμφραζόμενα γραμματικές, έχοντας σαν είσοδο μόνο θετικά παραδείγματα. Ιδιαίτερη προσοχή έχει δοθεί στην επεξεργαστική αποδοτικότητα και την *εφαρμοσιμότητα (scalability)* του αλγορίθμου σε μεγάλα σύνολα παραδειγμάτων, ικανοποιώντας τον στόχο που διέπει αυτή την διατριβή, για προσεγγίσεις που μπορούν να χρησιμοποιηθούν σε πρακτικά συστήματα επεξεργασίας φυσικής γλώσσας. Για παράδειγμα, η *συμπερασματική διαδικασία (inference process)* έχει βελτιστοποιηθεί μέσω της χρήσης των αποτελεσμάτων μιας θεωρητικής ανάλυσης σχετικά με τη δυναμική συμπεριφορά των τελεστών αναζήτησης και την πολυπλοκότητα που συνεπάγεται από την εφαρμογή τους κατά τη διαδικασία της αναζήτησης στον χώρο των πιθανών γραμματικών.

## 5.1 Ορισμός προβλήματος

Ο όρος *«επαγωγική εξαγωγή γραμματικών» (grammatical inference)* αναφέρεται σε μια κατηγορία αλγορίθμων μηχανικής μάθησης, οι οποίοι αποσκοπούν στην εξαγωγή γραμματικών από παραδείγματα προτάσεων. Με απλά λόγια, το πρόβλημα της επαγωγικής εξαγωγής γραμματικών μπορεί να συνοψιστεί ως εξής: έχοντας σαν είσοδο ένα σύνολο προτάσεων που υπακούουν σε μια *τυπική γλώσσα (formal language)*, δημιούργησε μια γραμματική για την γλώσσα αυτή.

Πιο συνοπτικά, έχοντας ένα αλφάβητο $\Sigma$, και ένα σύνολο προτάσεων $S$, το οποίο απαρτίζεται από σύμβολα του $\Sigma$, και προέρχεται από μια άγνωστη γλώσσα $L \subseteq \Sigma^*$, στόχος είναι να προσδιοριστεί το $L$: για παράδειγμα να κατασκευαστεί μια γραμματική η οποία να αποδέχεται κάθε πρόταση της $L$, και να απορρίπτει κάθε πρόταση που δεν περιέχεται στην γλώσσα $L$. Επίσης, μερικές φορές είναι διαθέσιμο και ένα ακόμα σύνολο προτάσεων $S'$, το οποίο περιλαμβάνει προτάσεις που δεν ανήκουν στην $L$ (αρνητικά παραδείγματα).

Η *επιτυχία (tractability)* του εγχειρήματος εξαρτάται από διάφορες παραμέτρους, όπως το μέγεθος και την ποιότητα του $S$, την διαθεσιμότητα του $S'$, τον τύπο της γλώσσας στην οποία υπάγεται η $L$, αλλά και η αυστηρότητα στην αναγνώριση της $L$.

## 5.2 Βιβλιογραφική επισκόπηση

Το πρόβλημα της *επαγωγικής εξαγωγής γραμματικών* έχει μια πλούσια ιστορία στην διεθνή βιβλιογραφία, και συνεχίζει να προσελκύει σημαντικό ερευνητικό ενδιαφέρον. Αν και το γενικό πρόβλημα της εξαγωγής γραμματικών από παραδείγματα είναι αρκετά δύσκολο, υπάρχει σημαντικό ερευνητικό ενδιαφέρων για μεθόδους που εμφανίζουν καλά εμπειρικά αποτελέσματα, οι οποίες επικεντρώνονται είτε σε συγκεκριμένο τύπο γλωσσών (π.χ. *κανονικές γλώσσες – regular grammars*), ή χρησιμοποιούν ευριστικά.

Η πλειοψηφία των αλγορίθμων επαγωγικής εξαγωγής γραμματικών που έχουν παρουσιαστεί στη βιβλιογραφία μοιράζονται μια κοινή μεθοδολογία. Έχοντας σαν βάση ένα αρχικό σύνολο θετικών παραδειγμάτων εκπαίδευσης, μια υπερβολικά συγκεκριμένη γραμματική κατασκευάζεται, η οποία είναι σε θέση να αναγνωρίσει μόνο αυτά τα παραδείγματα. Κατόπιν, ένα σύνολο τελεστών χρησιμοποιείται για να γενικεύσει αυτήν

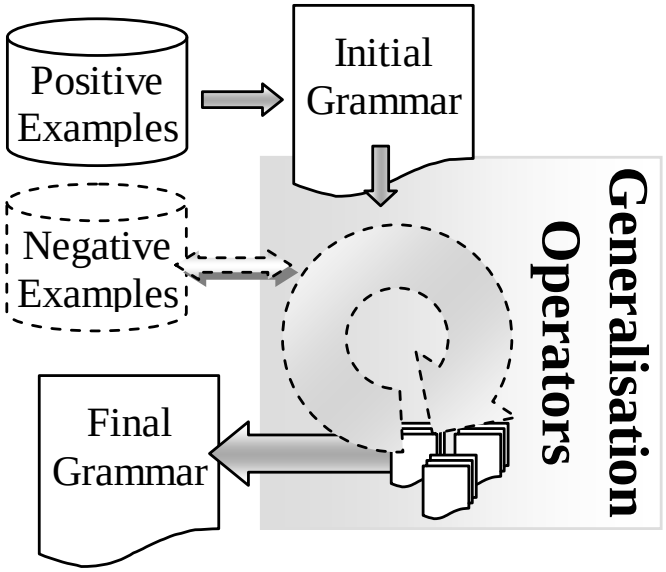
την αρχική γραμματική, συνήθως συνεπικουρούμενο από ένα σύνολο *αρνητικών παραδειγμάτων*, δηλ. προτάσεις που δεν πρέπει να αναγνωρίζονται από τη γραμματική (Εικόνα 23). Η ύπαρξη αρνητικών παραδειγμάτων είναι μια απαίτηση των περισσότερων αλγορίθμων, λόγω της ανάγκης να περιοριστεί η έκταση της γενίκευσης, δεδομένου ότι μια υπερβολικά γενική γραμματική δεν θα αντικρουστεί ποτέ από ένα νέο θετικό παράδειγμα: αν στην διαδικασία της εκπαίδευσης προκύψει μια γραμματική, της οποίας η γλώσσα είναι μεγαλύτερη από την άγνωστη γλώσσα στόχο, συχνά αυτή είναι μια αμετάκλητη ενέργεια, δεδομένου ότι κανένα θετικό παράδειγμα δεν μπορεί να βοηθήσει στην ανίχνευση αυτού του λάθους. Υπερβολικά γενικές γραμματικές μπορούν να ανιχνευθούν μόνο εάν αρνητικά παραδείγματα είναι διαθέσιμα, αφού μια γενική γραμματική είναι πιθανό να περιλαμβάνει και μερικά αρνητικά παραδείγματα. Συνεπώς, ένας αλγόριθμος μηχανικής μάθησης που χρησιμοποιεί αρνητικά παραδείγματα πρέπει κατά κύριο λόγο να αποφύγει το φαινόμενο της *απομνημόνευσης (overspecialisation/overfitting)*, αφού η *υπεραπλούστευση (overgeneralisation)* μπορεί να ελεγχθεί μέσω των αρνητικών παραδειγμάτων. Αντίθετα, ένας αλγόριθμος εκμάθησης που μαθαίνει μόνο από θετικά παραδείγματα, πρέπει να αποτρέψει και το φαινόμενο της απομνημόνευσης αλλά και της υπεραπλούστευσης.

Η δυνατότητα εκμάθησης των διάφορων τύπων γλωσσών (language classes), είτε στην ιεραρχία Chomsky είτε όχι, έχει μελετηθεί εκτενώς στην βιβλιογραφία. Στα δύο σχετικά *μοντέλα αναφοράς (reference models)* που χρησιμοποιούνται στη *θεωρία εκμάθησης (learning theory)*, δηλαδή στον *προσδιορισμό στο όριο (identification in the limit)* του Gold [95] και στο μοντέλο “probably approximately correct” (PAC) του Valiant [96], τα αποτελέσματα δεν είναι ενθαρρυντικά, ακόμη και για μια από τις απλούστερες κατηγορίες επίσημων γλωσσών, δηλ. την κατηγορία των κανονικών γλωσσών. Σύμφωνα με τον Gold [95], εάν μια κατηγορία γλωσσών περιέχει όλες τις πεπερασμένες γλώσσες και τουλάχιστον μια άπειρη γλώσσα, τότε *δεν μπορεί να προσδιοριστεί στο όριο, μόνο από θετικά παραδείγματα*. Κατά συνέπεια, οι κανονικές γραμματικές δεν είναι δυνατό να εξαχθούν μόνο από θετικά παραδείγματα στο μοντέλο του πρότυπου του Gold. Επιπρόσθετα, η Angluin [97] αποδεικνύει ότι οι ανεξάρτητες από συμφραζόμενα γραμματικές (context-free grammars) δεν μπορούν να εξαχθούν μόνο από θετικά παραδείγματα σε πολυωνυμικό χρόνο, ακόμα και στην περίπτωση που μπορούν να τεθούν *ερωτήματα ιδιότητας μέλους (membership queries)* κατά την διαδικασία της εξαγωγής. Η κατάσταση είναι ακόμα χειρότερη σύμφωνα με το μοντέλο εκμάθησης PAC, καθώς έχει αποδειχθεί ότι ούτε οι απλές κατηγορίες γλωσσών δεν μπορούν να εξαχθούν [98], [99]. Εντούτοις, πρέπει να σημειωθεί ότι αυτά τα θεωρητικά αποτελέσματα είναι *σενάρια χειρότερης περίπτωσης (worst case scenarios)*: στην πράξη μπορούν να επιτευχθούν ικανοποιητικές προσεγγίσεις της γλώσσας στόχου, ακόμα και αν η πραγματική γλώσσα είναι θεωρητικά αδύνατο να βρεθεί.

Ωστόσο, δεν είναι πάντα εύκολη η απόκτηση αρνητικών παραδειγμάτων σε πραγματικά προβλήματα. Για παράδειγμα, στις περισσότερες εφαρμογές της επεξεργασίας φυσικής γλώσσας, μεγάλα σύνολα θετικών παραδειγμάτων είναι συνήθως διαθέσιμα, αλλά συνήθως κανένα αρνητικό παράδειγμα δεν είναι διαθέσιμο. Καθώς η απουσία αρνητικών στοιχείων προκύπτει συχνά στην πράξη, δύο λύσεις έχουν προταθεί για αυτό το πρόβλημα:

- Μερικές περιορισμένες κατηγορίες επίσημων γλωσσών έχει αποδειχθεί ότι μπορούν να εξαχθούν μόνο από θετικά παραδείγματα, όπως οι *αντιστρέψιμες γλώσσες (reversible languages)* [100], οι γλώσσες k-testable [101], γλώσσες code regular και code linear [102], οι *«καθαρά» ανεξάρτητες από τα συμφραζόμενα γλώσσες (pure context-free languages)* [103], [104], και τα *αυστηρώς ντετερμινιστικά αυτόματα (strictly deterministic automata)* [105].

- Έχουν προταθεί διάφορα *ευριστικά (heuristics)*, τα οποία στοχεύουν την αποφυγή της *υπεραπλούστευσης (overgeneralisation)*, χωρίς τη χρήση αρνητικών παραδειγμάτων [106], [13].




**Εικόνα 23: Η Αρχιτεκτονική ενός τυπικού αλγορίθμου επαγωγικής εξαγωγής γραμματικών.**

Ο αλγόριθμος egGRIDS+ που παρουσιάζεται σε αυτό το κεφάλαιο, ανήκει στην τελευταία κατηγορία και χρησιμοποιεί το κριτήριο της απλότητας ως ευριστικό για να κατευθύνει την συμπερασματική διαδικασία (inference process), βασιζόμενο αποκλειστικά σε θετική πληροφόρηση. Η προσέγγιση που ακολουθεί ο αλγόριθμος egGRIDS+ είναι παρόμοιος με αυτόν που ακολουθείται από πολλούς αλγορίθμους επαγωγικής εξαγωγής γραμματικών, όπως απεικονίζεται στην Εικόνα 23. Ο αλγόριθμος egGRIDS+ είναι βασισμένος στον αλγόριθμο GRIDS (*"GRammar Induction Driven by Simplicity"*) [13], οπότε επίσης ενσωματώνει μια *προτίμηση προς απλές γραμματικές* κατά την αναζήτηση στον χώρο των πιθανών γραμματικών. Θεωρώντας τις γραμματικές σαν κώδικα, το ευριστικό που χρησιμοποιεί ο egGRIDS+, βασισμένο στο *«ελάχιστο μήκος περιγραφής» (Minimum Description Length - MDL)*, επιδιώκει να συμπιέσει η ίδια τη γραμματική, καθώς επίσης και την κωδικοποίηση των προτάσεων εκπαίδευσης από τη γραμματική.

Ταυτόχρονα, ο egGRIDS+ βελτιώνει τον GRIDS σε αρκετά σημεία. Το σημαντικότερο μειονέκτημα του αλγορίθμου GRIDS συνδέεται με την αναζήτηση που εκτελείται από τους *τελεστές εκμάθησης (learning operators)* στον χώρο των πιθανών γραμματικών. Καθώς κάθε τελεστής απαριθμεί εξαντλητικά όλες τις θυγατρικές γραμματικές που παράγονται από την εφαρμογή του στην μητρική γραμματική, ο GRIDS δεν μπορεί να εφαρμοστεί σε σύνολα εκπαίδευσης που περιέχουν έναν μεγάλο αριθμό παραδειγμάτων ή που χρησιμοποιούν ένα μεγάλο λεξιλόγιο. Προκειμένου να αντιμετωπιστεί αυτή η ανεπάρκεια, η δυναμική συμπεριφορά των τελεστών εκμάθησης εξετάστηκε, και τα αποτελέσματα αυτής της θεωρητικής ανάλυσης οδήγησαν σε βελτιστοποιήσεις που επιτρέπουν την επεξεργασία μεγάλων συνόλων δεδομένων. Μια άλλη σημαντική βελτίωση αφορά την εισαγωγή νέων τελεστών εκμάθησης που επιταχύνουν την εκμάθηση συγκεκριμένων γλωσσικών φαινομένων, καθώς και την προσθήκη νέων στρατηγικών αναζήτησης που προσπαθούν να επιταχύνουν την αναζήτηση με ευριστικές τεχνικές, όπως η αναζήτηση μέσω στατιστικής πληροφορίας ή αναζήτηση μέσω γενετικών αλγορίθμων. Επιπλέον, ο egGRIDS+ εφαρμόζει ένα λίγο

διαφορετικό κριτήριο «απλότητας», επιτρέποντας ακριβέστερα αποτελέσματα υπό συνθήκη.

Ο egGRIDS+ μοιράζεται μερικά από τα κεντρικά χαρακτηριστικά του με προηγούμενες εργασίες στην επαγωγική εξαγωγή γραμματικών. Ο egGRIDS+ προέρχονται από τον GRIDS, ο οποίος είναι βασισμένος στο SNPR [93], [94]. Το σύστημα SNPR χρησιμοποιεί παρόμοιους τελεστές εκμάθησης με τον GRIDS, με την εξαίρεση ότι ενδεχομένως να μπορούν να χρησιμοποιηθούν περισσότερα από δύο σύμβολα από ένα τελεστή. Το SNPR έχει επίσης προτίμηση προς «απλές» γραμματικές, δεδομένου ότι χρησιμοποιεί το ευριστικό του «ελάχιστου μήκους περιγραφής» (Minimum Description Length – MDL) για την αποτίμηση και επιλογή των πιο εύλογων γραμματικών. Επιπλέον, ο SNPR, όπως και ο egGRIDS+, χρησιμοποιεί ευριστικά για την αποφυγή της παραγωγής όλων των πιθανών γραμματικών σε κάθε επανάληψη. Η επιλογή όμως των ευριστικών αυτών έχει γίνει διαισθητικά, γεγονός που μπορεί να μην επιτρέψει στο SNPR να συγκλίνει στη βέλτιστη γραμματική, σύμφωνα με το ευριστικό του $MDL$.

Αν και η πλειοψηφία των εργασιών στην επαγωγική εξαγωγή γραμματικών εστιάζει στις κανονικές γραμματικές, υπάρχει ένας μικρός αριθμός αλγορίθμων που εξάγουν ανεξάρτητες από τα συμφραζόμενα γραμματικές. Οι Stolcke και Omohundro [107], [108], έχουν παρουσιάσει μια προσέγγιση που συμπεραίνει πιθανοτικές, ανεξάρτητες από τα συμφραζόμενα, γραμματικές. Χρησιμοποιώντας ένα πλαίσιο με βάση το νόμο του Bayes, το σύστημά τους χρησιμοποιεί παρόμοιους τελεστές με τον egGRIDS+ και προσπαθεί να βρει μια γραμματική με τη *μέγιστη μεταγενέστερη πιθανότητα (maximal posterior probability)*, δοθέντος ενός συνόλου εκπαίδευσης. Αυτό το κριτήριο είναι ουσιαστικά ισοδύναμο με το ευριστικό του $MDL$ που χρησιμοποιείται από τον egGRIDS+. Στην εργασία [109] παρουσιάζεται ένας αλγόριθμος για την εκμάθηση ανεξάρτητων από τα συμφραζόμενα γραμματικών από θετικά και αρνητικά δομημένα παραδείγματα. Ένα δομημένο παράδειγμα είναι ένα παράδειγμα με μερικές παρενθέσεις που παρεμβάλλονται για να δείξουν τη μορφή του δέντρου ανάλυσης/παραγωγής του παραδείγματος με βάση κάποια γραμματική. Ο αλγόριθμος εκμάθησης CKY [110] υιοθετεί μια γενετική αναζήτηση για τη σύγκλιση σε μια τελική γραμματική. Ένας διάδοχος αυτού του αλγορίθμου περιγράφεται στην εργασία [111]. Μια πιο πρόσφατη έκδοση αυτού του αλγορίθμου [112] λειτουργεί σε μερικώς δομημένα παραδείγματα, αντί σε πλήρως δομημένα παραδείγματα, ενώ χρησιμοποιεί μια συνοπτική αναπαράσταση που οδηγεί σε έναν πιο προσαρμόσιμο και εφαρμόσιμο αλγόριθμο, έναντι του προκατόχου του. Ο αλγόριθμος έχει εφαρμοστεί επιτυχώς στις διάφορες απλές γλώσσες καθώς επίσης και στη μοντελοποίηση ακολουθιών DNA. Ο Synapse [113] είναι ένας ακόμα αλγόριθμος βασισμένος στον CKY, ο οποίος μαθαίνει *επαυξητικά (incrementally)* από θετικά και αρνητικά παραδείγματα, ενώ τα προκαταρκτικά αποτελέσματα δείχνουν ότι είναι σε θέση να συμπεράνει σε λογικό χρόνο τόσο *διφορούμενες (ambiguous)* όσο και *αποσαφηνισμένες (unambiguous)* ανεξάρτητες από τα συμφραζόμενα γραμματικές για απλές γλώσσες.

Διάφορες προσπάθειες έχουν γίνει επίσης για την εφαρμογή αλγορίθμων επαγωγικής εξαγωγής γραμματικών σε προβλήματα μάθησης από την περιοχή της επεξεργασίας φυσικής γλώσσας. Στην εργασία [114], μια στοχαστική παραλλαγή του αλγορίθμου του Sakakibara αξιολογείται σε μέρος του Penn Treebank [115], με στόχο την εκμάθηση του συντακτικού της γλώσσας από θετικά δομημένα παραδείγματα. Ο Freitag [116] εφαρμόζει τρεις αλγορίθμους επαγωγικής εξαγωγής γραμματικών για εξαγωγή πληροφορίας από κείμενα που περιγράφουν ανακοινώσεις σεμιναρίων. Στην ίδια εργασία, ένας ειδικός αλγόριθμος προτείνεται που μαθαίνει ειδικούς «μετατροπείς αλφάβητου», που επανακωδικοποιούν το κείμενο σε μια αφαιρετική αναπαράσταση,

περισσότερο κατάλληλη για την εφαρμογή τεχνικών επαγωγικής εξαγωγής γραμματικών. Στην εργασία [117] παρουσιάζεται ένας αλγόριθμος για το συμπερασμό ανεξάρτητων από τα συμφραζόμενα γραμματικών που προσπαθούν να εντοπίσουν προκαθορισμένα πεδία σε δομημένα έγγραφα. Ο αλγόριθμος λειτουργεί δημιουργώντας ένα αυτόματο δέντρου προθέματος (prefix tree automaton) από τις εμφανίσεις όλων των πεδίων στο κείμενο. Αυτά τα αυτόματα γενικεύονται με τη βοήθεια ευριστικών, μετασχηματισμένων σε κανονικές εκφράσεις, τα οποία συνδυαζόμενα σχηματίζουν τις τελικές (ανεξάρτητες από τα συμφραζόμενα) γραμματικές. Στην αναφορά [118], ο αλγόριθμος ALERGIA, μαζί με το νέο αλγόριθμο WIL, εφαρμόζονται στο πρόβλημα της εκμάθησης wrappers για την εξαγωγή πληροφορίας από ιστοσελίδες. Ο αλγόριθμος ABL [119] δέχεται ως είσοδο ένα επίπεδο (flat) σώμα κειμένων (μη δομημένων) προτάσεων και επιστρέφει ένα σώμα επισημειωμένων –και δομημένων– προτάσεων. Στο [120], οι αλγόριθμοι ABL και EMILE αξιολογούνται στο σώμα παραδειγμάτων ATIS. Ο Clark [121] παρουσιάζει ένα νέο αλγόριθμο, ο οποίος χρησιμοποιεί content distribution clustering για την επαγωγή στοχαστικών ανεξάρτητων από τα συμφραζόμενα γραμματικών από επισημειωμένα κείμενα. Αυτός ο αλγόριθμος αξιολογείται στο πρόβλημα της εκμάθησης *γραμματικών δομημένων φράσεων (phrase structured grammars)* από το British National Corpus, καθώς και από το σώμα παραδειγμάτων ATIS.

Η συντριπτική πλειοψηφία των αλγορίθμων που εφαρμόζονται σε προβλήματα μάθησης από την περιοχή της επεξεργασίας φυσικής γλώσσας, συμπεραίνει γραμματικές μόνο από θετικά παραδείγματα, δεδομένου ότι η επεξεργασία φυσικής γλώσσας είναι πεδίο εφαρμογής όπου αρνητικά παραδείγματα είναι σπάνια διαθέσιμα. Μια άλλη ενδιαφέρουσα παρατήρηση σχετικά με την εφαρμογή των αλγορίθμων επαγωγικής εξαγωγής γραμματικών στην επεξεργασία φυσικής γλώσσας, είναι ότι πολλοί από αυτούς τους αλγορίθμους επιδεικνύουν στην πράξη έναν πολύ μεγάλο χρόνο σύγκλισης. Εκτός από την απουσία αρνητικών παραδειγμάτων, η εκμάθηση στην περιοχή της φυσικής γλώσσας βάζει επιπρόσθετους περιορισμούς στους αλγορίθμους, καθώς απαιτεί εκμάθηση από γλώσσες με μεγάλο αλφάβητο, οι οποίες είναι επίσης σχετικά σύνθετες, καθιστώντας αναγκαία την εξέταση ενός σχετικού μεγάλου όγκου παραδειγμάτων. Η ικανότητα ενός αλγορίθμου επαγωγικής εξαγωγής γραμματικών να συγκλίνει σε λογικό χρόνο όταν εκπαιδεύεται με μεγάλα σύνολα παραδειγμάτων δεν θα πρέπει σε καμία περίπτωση να υποτιμηθεί.

## 5.3 Ο αλγόριθμος egGRIDS+: επαγωγική εξαγωγή γραμματικών από θετικά παραδείγματα

Όπως έχουμε ήδη σημειώσει, η επαγωγική εξαγωγή γραμματικών από θετικά παραδείγματα είναι δυσκολότερη από την εξαγωγή από σύνολα δεδομένων όπου αρνητική πληροφορία είναι παρούσα. Προκειμένου να αντισταθμιστεί η έλλειψη αρνητικών παραδειγμάτων, ένας αλγόριθμος εκμάθησης μπορεί είτε να μάθει περιορισμένες κατηγορίες επίσημων γλωσσών (για τις οποίες να έχει αποδειχθεί ότι μπορούν να εξαχθούν από θετικά παραδείγματα), ή να χρησιμοποιήσουν ευριστικά προς αποφυγή της υπεραπλούστευσης.

Όπως ο GRIDS, έτσι και ο egGRIDS+ ακολουθεί το τελευταίο παράδειγμα με την ενσωμάτωση έναν ευριστικού που τιμωρεί τις υπερβολικά γενικές γραμματικές. Αυτή η επιλογή προτιμήθηκε εν μέρει από το μακροπρόθεσμο στόχο κατασκευής πρακτικών συστημάτων με τον egGRIDS+, που να μπορούν να χρησιμοποιηθούν σε πρακτικά συστήματα επεξεργασίας φυσικής γλώσσας. Για τον ίδιο λόγο, ήταν επιθυμητή η εκμάθηση γραμματικών μιας γλωσσικής κατηγορίας που να μπορεί να αντιπροσωπεύσει μια μεγάλη ποικιλία των γλωσσικών φαινομένων. Αυτή η απαίτηση οδήγησε στον αποκλεισμό των περιορισμένων γλωσσών, και στην επιλογή των

*γραμματικών ανεξάρτητων από συμφραζόμενα (context-free grammars)* σαν την κατηγορία στόχο, οι οποίες έχουν αρκετή εκφραστικότητα για να αναπαραστήσουν μια πληθώρα γλωσσικών φαινομένων. Αν και η ιδανικότερη κατηγορία γραμματικών θα ήταν οι *γραμματικές συμφραζομένων (context-sensitive grammars)*, κάποιες ιδιότητές τους όπως η *αναποφασιστικότητα (undecidability)* και η πολυπλοκότητα ανάλυσης με αυτές τις καθιστούν υπολογιστικά ακριβές για χρήση σε πρακτικά προβλήματα. Ταυτόχρονα, μόνο λίγα γλωσσικά φαινόμενα (σε κάποιες γλώσσες) δεν μπορούν να μοντελοποιηθούν με γραμματικές ανεξάρτητες από τα συμφραζόμενα, όπως το «cross-serial dependency», τα οποία χρειάζονται γραμματικές με συμφραζόμενα. Αυτά τα φαινόμενα εκτός από σπάνια, στην πράξη εμφανίζονται περιορισμένα (π.χ. δεν περιλαμβάνουν αναδρομή πολλών επιπέδων), επιτρέποντας την μοντελοποίηση τους με λιγότερο εκφραστικές γραμματικές. Για αυτούς τους λόγους, για τον αλγόριθμο egGRIDS+ επιλέξαμε τις γραμματικές ανεξάρτητες από συμφραζόμενα έναντι αυτών με συμφραζόμενα.

Ο αλγόριθμος GRIDS εξάγει γραμματικές ανεξάρτητες από τα συμφραζόμενα από θετικά παραδείγματα. Υλοποιεί μια ευριστική αναζήτηση προς «απλές» γραμματικές, χρησιμοποιώντας δύο τελεστές εκμάθησης, οι οποίοι παράγουν νέες γραμματικές. Η διαδικασία αναζήτησης οργανώνεται σαν *αναζήτηση δέσμης (beam search)*. Βασισμένος στον GRIDS, ο egGRIDS+ μοιράζεται τέσσερα χαρακτηριστικά γνωρίσματα με τον προκάτοχό του:

- Αναπαράσταση γνώσης (ανεξάρτητες από τα συμφραζόμενα γραμματικές).
- Προτίμηση προς απλές γραμματικές (όπως αποτιμούνται από το ελάχιστο μήκος περιγραφής).
- Δύο βασικούς τελεστές εκμάθησης.
- Αναζήτηση μέσω αναζήτησης δέσμης.

Ωστόσο, υπάρχουν και σημαντικές διαφορές μεταξύ των δύο αλγορίθμων, καθώς ο egGRIDS+:

- Βελτιστοποιεί τη διαδικασία αναζήτησης, χρησιμοποιώντας τα αποτελέσματα θεωρητικής ανάλυσης της δυναμικής συμπεριφοράς των τελεστών εκμάθησης
- Ενσωματώνει έναν νέο τελεστή εκμάθησης, που μπορεί να οδηγήσει σε «απλούστερες»
- Βελτιώνει το μέτρο απλότητας, το οποίο βασίζεται στο ελάχιστο μήκος περιγραφής (MDL)

### 5.3.1 Αναπαράσταση γνώσης στον egGRIDS+

Κάθε γραμματική στον egGRIDS+ αποτελείται από ένα σύνολο από ανεξάρτητους από τα συμφραζόμενα παραγωγικούς (ή επανεγγραφής) κανόνες (*context-free production (or rewrite) rules*), οι οποίοι εκφράζονται χρησιμοποιώντας τερματικός ένα σύνολο από *μη τερματικά σύμβολα (non-terminal symbols)* (π.χ. φράσεις στις φυσικές γλώσσες) και ένα σύνολο *τερματικών συμβόλων (terminal symbols)* (π.χ. λέξεις). Το σύνολο των μη τερματικών συμβόλων περιέχει ένα πρόσθετο σύμβολο έναρξης (*"S"* το οποίο συμβολίζει την πρόταση). Κάθε κανόνας παραγωγής (production rule) αποτελείται από μια κεφαλή (head) και ένα σώμα (body), στην ακόλουθη μορφή:

$$\underbrace{X}_{Head} \rightarrow \underbrace{Y \ \ldots \ Z}_{Body}$$

όπου τα X, Y,…, Z αντιπροσωπεύουν μοναδικά σύμβολα. Η κεφαλή (ή αλλιώς η αριστερή πλευρά του κανόνα) αποτελείται μόνο από ένα μη τερματικό σύμβολο ενώ το

σώμα (ή διαφορετικά η δεξιά πλευρά του κανόνα) μπορεί να περιέχει ένα ή περισσότερα σύμβολα, είτε τερματικά είτε μη τερματικά. Η σημασία ενός κανόνα παραγωγής είναι ότι κάποιος μπορεί να αντικαταστήσει κάθε *εμφάνιση (occurrence)* της κεφαλής του κανόνα με το σώμα του (και το αντίστροφο) κατά την αναγνώριση ή παραγωγή μιας πρότασης. Όπως ο αλγόριθμος GRIDS αλλά και προγενέστερη εργασία [122], θα θεωρήσουμε τους ακόλουθους περιορισμούς για τη μορφή που μπορούν να λάβουν οι κανόνες παραγωγής:

1. Κανένας κανόνας παραγωγής δεν μπορεί να έχει κενή κεφαλή.
2. Οι κανόνες της μορφής "X → Y" επιτρέπονται μόνο εάν το σύμβολο "Y" είναι ένα τερματικό σύμβολο, δηλ., ένα μη τερματικό σύμβολο δεν μπορεί να ισοδυναμεί με ένα διαφορετικό μη τερματικό σύμβολο.
3. Κάθε μη τερματικό σύμβολο εμφανίζεται στην αναγνώριση ή την παραγωγή κάποιας πρότασης (θετικό παράδειγμα).

Επιπλέον θα προσθέσουμε τον ακόλουθο περιορισμό:

4. Τελικά σύμβολα επιτρέπονται μόνο σε κανόνες της μορφής *"$X \rightarrow Y$"*, και όχι σε κανόνες που τα σώματα των οποίων περιέχουν περισσότερα από ένα σύμβολα.

Αυτές οι τέσσερις απαιτήσεις δεν έχουν επιπτώσεις στην δύναμη αναπαράστασης μιας γραμματικής, καθώς οποιαδήποτε ανεξάρτητη από τα συμφραζόμενα γραμματική *(CFG)* μπορεί να μετατραπεί σε μια γραμματική που συμμορφώνεται με αυτούς τους περιορισμούς [110]. Ο λόγος για τον τέταρτο περιορισμό είναι η παραγωγή γραμματικών με χαμηλότερο μήκος περιγραφής. Ένα παράδειγμα μιας απλής γραμματικής, ανεξάρτητης από τα συμφραζόμενα, παρουσιάζεται στον πίνακα 1.

| | |
|---|---|
| S → NP VERB1<br>S → NP VERB2<br>NP → ART NOUN | ART → the<br>NOUN → dog<br>VERB1 → ran<br>VERB2 → barked |
| The dog ran<br>The dog barked | |

**Πίνακας 31: Παράδειγμα μιας απλής CFG, μαζί με το πλήρες σύνολο προτάσεων που μπορούν να παραχθούν ή να αναλυθούν από αυτήν την γραμματική, όπως παρουσιάζεται στην εργασία [13].**

Τυπικότερα, μια CFG είναι μια τετράδα $G = (V_{NT}, V_T, P, "S")$ όπου $V_{NT}$ είναι ένα σύνολο *μη τερματικών* συμβόλων, $V_T$ ένα σύνολο από *τερματικά σύμβολα*, $P$ ένα σύνολο *κανόνων παραγωγής*, ικανό να παράγει έγκυρες προτάσεις της γλώσσας, και $"S" \in V_{NT}$ είναι ένα ειδικό σύμβολο, αποκαλούμενο *σύμβολο έναρξης*. Το σύνολο των ανεξάρτητων από τα συμφραζόμενα κανόνων παραγωγής $P$ κατασκευάζεται χρησιμοποιώντας το σύμβολο έναρξης, τα σύμβολα από το σύνολο μη-τερματικών συμβόλων $V_{NT}$, και σύμβολα από το σύνολο $V_T$ των τερματικών συμβόλων. Κάθε κανόνας παραγωγής είναι της μορφής $\alpha \rightarrow \beta$, όπου $\alpha \in V_{NT}$, $|\alpha| = 1$ και $\beta \in (V_{NT} \cup V_T)^*$.

### 5.3.2 Αναζήτηση στον egGRIDS+: προτίμηση προς «απλές» γραμματικές

Καθώς ο egGRIDS+ δεν χρησιμοποιεί αρνητικά παραδείγματα, απαιτείται κάποιο κριτήριο για να κατευθύνει την αναζήτηση μέσα στον χώρο των ανεξάρτητων από τα συμφραζόμενα γραμματικών, και να αποφύγει τις υπερβολικά γενικές γραμματικές. Όπως έχουμε ήδη αναφέρει, το κριτήριο αυτό παρέχει μια προτίμηση προς *απλές* γραμματικές. Θεωρώντας τόσο την γραμματική όσο και τα παραδείγματα σαν κώδικα, μια γραμματική $A$ θεωρείται απλούστερη από μια γραμματική $B$ εάν το άθροισμα

(a) του αριθμού των *δυφίων (δυαδικών ψηφίων – bits)* που απαιτούνται για την κωδικοποίηση της γραμματικής $A$, και
(b) του αριθμού των δυφίων που απαιτούνται για την κωδικοποίηση των παραδειγμάτων εκπαίδευσης ως *παραγωγές (derivations)* της γραμματικής $A$

είναι χαμηλότερο από το αντίστοιχο άθροισμα για την γραμματική $B$.

Δεδομένου ότι η *συμπερασματική διαδικασία (inference process)* πρέπει να προστατευτεί τόσο από πολύ συγκεκριμένες γραμματικές (που απομνημονεύουν απλά τα παραδείγματα) όσο και από υπεραπλουστευμένες γραμματικές (που αποδέχονται κάθε πρόταση), υπάρχουν δύο είδη τετριμμένων γραμματικών που θέλουμε να αποφύγουμε. Το πρώτο είδος είναι μια γραμματική που έχει έναν κανόνα παραγωγής για κάθε παράδειγμα, και δεν μπορεί να χειριστεί νέα, *απαρατήρητα (unseen) παραδείγματα*. Το άλλο είδος τετριμμένης γραμματικής που θέλουμε να αποφύγουμε είναι μια γραμματική που αποδέχεται οποιοδήποτε παράδειγμα σαν ορθό. Ακολουθώντας τον αλγόριθμο GRIDS, υιοθετούμε την προσέγγιση του *ελάχιστου μήκους περιγραφής (minimum description length – MDL)*, που κατευθύνει τη διαδικασία αναζήτησης προς γραμματικές που είναι συμπαγείς, δηλ., αυτές που απαιτούν λίγα δυφία για να κωδικοποιηθούν, ενώ συγχρόνως κωδικοποιούν το σύνολο των παραδειγμάτων με έναν συμπαγή τρόπο, δηλ. να απαιτούνται λίγα δυφία για να κωδικοποιήσουν τα παραδείγματα χρησιμοποιώντας τη γραμματική.

Η κεντρική ιδέα πίσω από το *MDL* [123] μπορεί να αναλυθεί σε τέσσερα βασικά βήματα:

1. Την κατάρτιση ενός μοντέλου, βασισμένο στο σύνολο των παραδειγμάτων.
2. Την χρήση αυτού του μοντέλου για την κωδικοποίηση (συμπίεση) του συνόλου παραδειγμάτων, και την απόδοση ενός μήκους στην συμπτυγμένη μορφή, χρησιμοποιώντας συνήθως έννοιες από τη *θεωρία πληροφοριών (information theory)*.
3. Τον υπολογισμό του μήκους του μοντέλου, χρησιμοποιώντας πάλι έννοιες από τη θεωρία πληροφοριών.
4. Την αναζήτηση του καλύτερου δυνατού μοντέλου, που αντιστοιχεί στην ελαχιστοποίηση του μήκους της γραμματικής και των συμπιεσμένων με την γραμματική παραδειγμάτων.

Με απλά λόγια, το MDL στοχεύει σε μια ελάχιστη συμπαγή αναπαράσταση και του *μοντέλου και των δεδομένων*, ταυτόχρονα. Αξίζει να σημειωθεί ότι το MDL δεν παρέχει τα μέσα για την παραγωγή μοντέλων, παρά μόνο για την αποτίμηση αυτών. Στην εξαγωγή γραμματικών, το MDL προσφέρει απλά έναν μηχανισμό για σύγκριση γραμματικών και επιλογής εκείνης που είναι περισσότερο «συμπαγής», όσον αφορά το μήκος της γραμματικής και της κωδικοποίησης των παραδειγμάτων από τη γραμματική.

Προκειμένου να χρησιμοποιηθεί το MDL, πρέπει να μετρήσουμε το μήκος κωδικοποίησης της γραμματικής και του συνόλου των παραδειγμάτων, όπως κωδικοποιείται από τη γραμματική. Υποθέτοντας μια ανεξάρτητη από τα συμφραζόμενα γραμματική $G$ και ένα σύνολο παραδειγμάτων (προτάσεις) $T$, τα οποία μπορούν να αναγνωριστούν (αναλυθούν) από τη γραμματική $G$, το συνολικό μήκος περιγραφής μιας γραμματικής (εφεξής $ML$ – *μήκος περιγραφής του μοντέλου – model description length*) είναι το άθροισμα δύο ανεξάρτητων μηκών:

- Το μήκος περιγραφής γραμματικής (grammar description length – GDL), δηλ. τα δυφία που απαιτούνται για να κωδικοποιήσουν τους κανόνες της γραμματικής και να τους διαβιβάσουν σε έναν παραλήπτη που έχει ελάχιστη γνώση αναπαράστασης γραμματικών, και

- το μήκος περιγραφής παραγωγών (derivations description length – DDL), δηλ. τα δυφία που απαιτούνται για να κωδικοποιήσουν και να διαβιβάσουν όλα τα παραδείγματα του συνόλου $T$, υπό τον όρο ότι ο παραλήπτης ξέρει ήδη τη γραμματική $G$.

$$ML = GDL + DDL$$

Το πρώτο συστατικό του ευριστικού $ML$ κατευθύνει την αναζήτηση μακριά από το είδος της τετριμμένης γραμματικής που έχει έναν χωριστό κανόνα για κάθε πρόταση εκπαίδευσης, δεδομένου ότι αυτή η γραμματική θα έχει ένα μεγάλο GDL. Εντούτοις, το ίδιο συστατικό οδηγεί στο άλλο είδος τετριμμένης γραμματικής, μια γραμματική που αποδέχεται όλες τις προτάσεις. Προκειμένου να αποφευχθεί αυτό, το δεύτερο συστατικό υπολογίζει τη *δυναμικότητα παραγωγής (derivation power)* της γραμματικής (ή εναλλακτικά τη γλώσσα της γραμματικής) και βοηθά να αποφευχθεί η υπεραπλούστευση με την τιμωρία των γενικών γραμματικών. Ο προφανής τρόπος να μετρηθεί η δύναμη παραγωγής μιας γραμματικής είναι να μετρηθούν όλες της οι παραγωγές, δηλ. οι προτάσεις που μπορούν να παραχθούν. Ωστόσο, αυτό δεν είναι πάντα εφικτό ή επιθυμητό, αφού συχνά οι γραμματικές παράγουν μια άπειρη γλώσσα μέσω αναδρομής. Ακόμα κι αν κάποιος μπορούσε να μετρήσει όλες τις παραγωγές μιας γραμματικής, στις περισσότερες περιπτώσεις το αποτέλεσμα θα ήταν ένας πολύ μεγάλος αριθμός έναντι του μήκους της γραμματικής, που θα καθιστούσε το ευριστικό ακατάλληλο για χρήση. Αντίθετα, η μέτρηση της δύναμης παραγωγής σύμφωνα το πώς τα παραδείγματα εκπαίδευσης παράγονται από τη γραμματική είναι υπολογιστικά εφικτό και δίνει τιμές πλησιέστερες στο GDL, καθιστώντας το περισσότερο κατάλληλο: όσο υψηλότερη η δύναμη παραγωγής της γλώσσας, τόσο υψηλότερο θα είναι και το DDL. Η αρχική – υπερβολικά συγκεκριμένη – γραμματική είναι η καλύτερη δυνατή από την άποψη του DDL, δεδομένου ότι συνήθως υπάρχει μια-προς-μια αντιστοιχία μεταξύ των παραδειγμάτων και των κανόνων γραμματικής, δηλ. η δύναμη παραγωγής της είναι χαμηλή. Από την άλλη, η πιο γενική γραμματική έχει το χειρότερο DDL, καθώς θα απαιτούνται πολλοί κανόνες στην παραγωγή κάθε πρότασης, απαιτώντας επιπρόσθετα δυφία για να καταγραφούν όλοι οι χρησιμοποιούμενοι κανόνες μονοσήμαντα.

### *Μήκος περιγραφής γραμματικής (GDL)*

Προκειμένου να μετρηθούν τα δυφία που απαιτούνται για να διαβιβαστεί μια γραμματική $G$ σε έναν παραλήπτη, πρέπει να συμφωνηθεί ο τρόπος με τον οποίο η γραμματική θα κωδικοποιηθεί και θα διαβιβαστεί. Η προσέγγισή μας είναι βασισμένη στο χωρισμό του συνόλου των κανόνων σε τρία ανεξάρτητα υποσύνολα, που διαβιβάζονται διαδοχικά. Το πρώτο υποσύνολο περιέχει όλους τους κανόνες της γραμματικής $G$ των οποίων η κεφαλή είναι το σύμβολο έναρξης της γραμματικής (start symbol subset – $SB_1$). Το δεύτερο υποσύνολο περιέχει όλους τους κανόνες της μορφής *"*$X \rightarrow Y$*"* (terminal subset – $SB_2$). Τέλος, το τρίτο υποσύνολο περιέχει όλους τους κανόνες της $G$ που δεν ανήκουν στα δύο πρώτα υποσύνολα (non-terminal subset – $SB_3$). Ο λόγος για αυτόν τον χωρισμό είναι το γεγονός ότι τα πρώτα δύο υποσύνολα παρουσιάζουν χαρακτηριστικά που μπορούν να χρησιμοποιηθούν για να μειώσουν τον αριθμό δυφίων που απαιτούνται για την κωδικοποίηση των αντίστοιχων κανόνων. Γενικά, ο αριθμός δυφίων που απαιτείται για την κωδικοποίηση ενός κανόνα είναι:

$$Bits_{Rule} = Bits_{Head} + Bits_{Body} + Bits_{End\ of\ rule}$$

Με άλλα λόγια, ο συνολικός αριθμός δυφίων που απαιτούνται για την κωδικοποίηση ενός κανόνα είναι το άθροισμα των δυφίων που απαιτούνται για την κωδικοποίηση της κεφαλής του κανόνα $(Bits_{Head})$ και του σώματός του $(Bits_{Body})$. Επιπλέον, όταν η γραμματική διαβιβάζεται, ένα πρόσθετο μοναδικό σύμβολο (π.χ. "$STOP$") πρέπει να επικολληθεί σε κάθε κανόνα προκειμένου να επισημανθεί το τέλος του κανόνα στον παραλήπτη (αφού οι κανόνες έχουν μεταβλητά μήκη). Αυτό το πρόσθετο σύμβολο αντιμετωπίζεται σαν μη τερματικό, απαιτώντας $Bits_{End}$ για να κωδικοποιηθεί.

Υποθέτοντας ότι η γραμματική $G$ έχει $A_{UNT}$τα μοναδικά μη τερματικά σύμβολα, εξαιρώντας το ειδικό σύμβολο "$STOP$", τα δυφία που απαιτούνται για την κωδικοποίηση μιας εμφάνισης (instance) ενός μη τερματικού συμβόλου είναι[4]:

$$Bits_{NT} = \log_2(A_{UNT} + 1)$$

όπου το πρόσθετο σύμβολο είναι το "$STOP$". Υποθέτοντας επίσης ότι η γραμματική έχει $T$ μοναδικά τερματικά σύμβολα (δηλ. λέξεις), ο αριθμός δυφίων που απαιτείται για την κωδικοποίηση μιας εμφάνισης ενός τερματικού συμβόλου είναι:

$$Bits_T = \log_2(T)$$

Οι κανόνες στο υποσύνολο συμβόλων έναρξης έχουν ως κεφαλή το σύμβολο έναρξης της γραμματικής. Αυτή η ιδιότητα μπορεί να χρησιμοποιηθεί για να μειώσει περεταίρω τα δυφία που χρειάζονται, αφού δεν υπάρχει ανάγκη να κωδικοποιηθεί η κεφαλή αυτών των κανόνων (που είναι κοινό και γνωστό στον παραλήπτη). Συνεπώς, προκειμένου να κωδικοποιηθεί ένας κανόνας του υποσυνόλου συμβόλων έναρξης, μπορεί να χρησιμοποιηθεί η ακόλουθη εξίσωση:

$$Bits_{Rule} = \left( \sum_{\substack{\forall NT \\ in\ rule\ body}} \log(A_{UNT} + 1) \right) + \log(A_{UNT} + 1)$$

Αν και η μείωση φαίνεται να είναι μικρής σημασίας, καθώς αφαιρούμε απλά ένα μικρό αριθμό μη τερματικών συμβόλων (σε σχέση με τον συνολικό αριθμό συμβόλων που πρέπει να κωδικοποιηθούν), αυτή έχει μια σημαντική παρενέργεια: ο συνολικός αριθμός των μοναδικών μη τερματικών $(A_{UNT})$ μειώνεται κατά ένα σύμβολο.

Όσον αφορά το υποσύνολο τερματικών συμβόλων, το οποίο περιέχει κανόνες της μορφής *"$X \rightarrow Y$"* (το *"$Y$"* είναι τερματικό σύμβολο), όλοι οι κανόνες του υποσυνόλου έχουν καθορισμένο μήκος, οπότε το σύμβολο "$STOP$" δεν απαιτείται για να επισημάνει το τέλος των κανόνων. Σε αυτήν την περίπτωση, τα δυφία που απαιτούνται για την κωδικοποίηση ενός κανόνα είναι:

[4] Όλοι οι λογάριθμοι στο παρόν έγγραφο έχουν σαν βάση το 2. Για λόγους αναγνωσιμότητας, δεν θα περιλαμβάνουμε τη βάση του λογαρίθμου στις εξισώσεις.

$$Bits_{Rule} = \log(A_{UNT} + 1) + \log(T)$$

Το συνολικό μήκος περιγραφής γραμματικής (GDL) είναι το άθροισμα των δυφίων που απαιτούνται για να κωδικοποιήσουν καθένα από τα τρία υποσύνολα της γραμματικής $G$, συν δύο πρόσθετα σύμβολα "$STOP$" που απαιτούνται για να χωριστούν τα τρία υποσύνολα:

$$\begin{aligned} GDL = & \sum_{\substack{\forall\ rule\ in \\ Start\ Symbol \\ Subset}} \left( \sum_{\substack{\forall\ NT \\ in\ rule\ body}} Bits_{NT} + Bits_{STOP} \right) + \Bigg| \begin{array}{l} Start\ Symbol\ Subset: \\ Body\ NTs + STOP \\ (No\ bits\ needed\ for\ Head) \end{array} \\ & Bits_{STOP} + \Big| Subset\ Separator \\ & \sum_{\substack{\forall\ rule\ in \\ Terminal \\ Subset}} (Bits_{NT} + Bits_{T}) + \Bigg| \begin{array}{l} Terminal\ Subset: \\ Head + Body \\ \left( \begin{array}{l} No\ STOP\ symbol \\ required \end{array} \right) \end{array} \\ & Bits_{STOP} + \Big| Subset\ Separator \\ & \sum_{\substack{\forall\ rule\ in \\ Non\text{-}terminal \\ Subset}} \left( Bits_{NT} + \sum_{\substack{\forall\ NT \\ in\ rule\ body}} Bits_{NT} + Bits_{STOP} \right) \Bigg| \begin{array}{l} Non-Terminal\ Subset: \\ Head + Body\ NTs + \\ STOP \end{array} \end{aligned} \tag{5.1}$$

όπου $Bits_{STOP} = Bits_{NT}$.

Σημειώνεται ότι ο τρόπος που έχουμε επιλέξει να κωδικοποιήσουμε τη γραμματική κάνει τέσσερις υποθέσεις για τη γνώση που ο έχει ο παραλήπτης σχετικά με την γραμματική $G$:

- Ο παραλήπτης γνωρίζει το σύνολο των τερματικών συμβόλων.
- Ο παραλήπτης γνωρίζει ότι οι κανόνες της γραμματικής χωρίζονται σε τρία υποσύνολα, και γνωρίζει και την σειρά που αυτά διαβιβάζονται.
- Ο παραλήπτης γνωρίζει πώς ένας κανόνας διαβιβάζεται για κάθε υποσύνολο.
- Ο παραλήπτης γνωρίζει ότι δύο συνεχόμενα "$STOP$" σηματοδοτούν την αρχή ενός νέου υποσυνόλου κανόνων.

Ένα παράδειγμα υπολογισμού του GDL μιας απλής γραμματικής $G$ παρουσιάζεται στον ακόλουθο πίνακα (Πίνακας 32).

| | |
|---|---|
| $\boldsymbol{A_{UNT}}$ = 5 (with out *"S" and "STOP"*) $\boldsymbol{A_{UT}}$ = 4<br>$\boldsymbol{Bits_{NT}}$ = log(5+1) = 2.58 | |
| ***Start Symbol Subset***<br>S → NP VERB1<br>S → NP VER<br>S O | $2\cdot 2.58+1\cdot 2.58+$<br>$2\cdot 2.58+1\cdot 2.58+$<br>$1\cdot 2.58+$ |
| ***Terminal Subset***<br>ART → the<br>NOUN → dog<br>VERB1 → ran<br>VERB2 → barked<br>S OP | $1\cdot 2.58+1\cdot \log(4)+$<br>$1\cdot 2.58+1\cdot \log(4)+$<br>$1\cdot 2.58+1\cdot \log(4)+$<br>$1\cdot 2.58+1\cdot \log(4)+$<br>$1\cdot 2.58+$ |
| ***Non-Terminal Subset***<br>NP → ART NOUN | $3\cdot 2.58+1\cdot 2.58$ |
| ***Total GDL:*** | $16\cdot 2.58+4\cdot 2=49.28$ Bits |

**Πίνακας 32: Υπολογίζοντας το μήκος περιγραφής (GDL) μιας γραμματικής $G$.**

### *Υποστηρίζοντας τερματικά σύμβολα που ανήκουν σε πολλαπλές κατηγορίες*

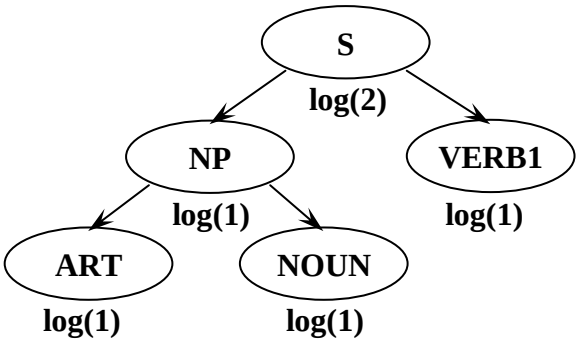
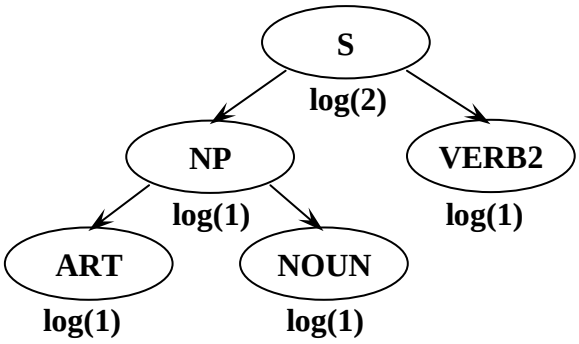
Στον αλγόριθμο GRIDS, τα τερματικά σύμβολα θεωρούνται ότι ταξινομούνται σε κατηγορίες, π.χ. μέρη του λόγου, οι οποίες ήταν γνωστές τόσο από τον αποστολέα όσο και από τον παραλήπτη. Με βάση αυτήν την υπόθεση, ο αποστολέας πρέπει να προσδιορίσει τη λέξη μέσα σε κάθε κατηγορία, απαιτώντας $log(Number\ of\ Unique\ Terminals)$ δυφία για την κωδικοποίηση κάθε τερματικού συμβόλου. Όμως, η κωδικοποίηση αυτή υποθέτει ότι ο παραλήπτης είναι σε θέση να προσδιορίσει μονοσήμαντα την κατηγορία στην οποία ανήκει ένα εισερχόμενο τερματικό σύμβολο. Αυτό μπορεί να είναι εφικτό όταν η κεφαλή ενός κανόνα δηλώνει την κατηγορία. Ωστόσο, οι κατηγορίες αντιπροσωπεύονται από μη τερματικά σύμβολα που μπορούν να τροποποιηθούν κατά την αναζήτηση. Σε αυτή την περίπτωση, ο παραλήπτης έχει ανεπαρκή πληροφόρηση για να εκτελέσει την αποκωδικοποίηση, δηλ. όταν ένας κανόνας τερματικού συμβόλου φθάσει έχοντας ως κεφαλή ένα νέο μη τερματικό σύμβολο, ο παραλήπτης δεν μπορεί να προσδιορίσει την αρχική κατηγορία του τερματικού συμβόλου. Η μέτρηση του $ML$ μιας γραμματικής σε τέτοιες περιπτώσεις είναι προβληματική, δεδομένου ότι δεν αντιπροσωπεύει μια σωστή κωδικοποίηση της γραμματικής. Προκειμένου να διορθωθεί αυτή η ανεπάρκεια, πρέπει να δοθούν οι πληροφορίες που λείπουν: εκτός από τα δυφία που απαιτούνται για την κωδικοποίηση των τερματικών συμβόλων, ο αποστολέας πρέπει επίσης να διευκρινίσει την κατηγορία στην οποία ανήκει το τερματικό σύμβολο.

Σαν απάντηση σε αυτό το πρόβλημα, ο egGRIDS+ εισάγει μια νέα προσέγγιση στην κωδικοποίηση των τερματικών συμβόλων της γραμματικής και στον υπολογισμό της συμμετοχής τους στο $ML$. Ενοποιώντας τον τρόπο που αντιμετωπίζονται τα τερματικά σύμβολα με τα μη τερματικά, ο egGRIDS+ υποθέτει ότι τα τερματικά σύμβολα ανήκουν στο ίδιο σύνολο με τα μη τερματικά, χρησιμοποιώντας $log(Number\ of\ Unique\ Terminals)$ δυφία για την κωδικοποίηση κάθε τερματικού συμβόλου. Μια ενδιαφέρουσα παρενέργεια του γεγονότος ότι τα τερματικά σύμβολα δεν

ομαδοποιούνται στον egGRIDS+, είναι η δυνατότητα χειρισμού τερματικών συμβόλων που είναι ταξινομημένα σε περισσότερες από μια κατηγορίες, δυνατότητα που είναι χρήσιμη για την μοντελοποίηση λέξεων ανήκουν σε πολλαπλές συντακτικές κατηγορίες.

### *Μήκος περιγραφής παραγωγών γραμματικής (DDL)*

Το μήκος περιγραφής παραγωγών μετρά τον αριθμό δυφίων που απαιτείται για την κωδικοποίηση και να διαβίβαση ενός συνόλου προτάσεων, όπως έχει αναγνωριστεί (αναλυθεί) από μια γραμματική $G$ (ή ισοδύναμα έχει παραχθεί από αυτή την γραμματική), υπό τον όρο ότι ο παραλήπτης ξέρει ήδη την γραμματική $G$. Το σύνολο των προτάσεων είναι στην περίπτωση του egGRIDS+ το σύνολο των δεδομένων εκπαίδευσης. Η κωδικοποίηση του τρόπου που ένα παράδειγμα παράγεται/αναλύεται από μια γραμματική $G$ απαιτεί τον σαφή προσδιορισμό όλων των κανόνων που συμμετέχουν στην παραγωγή/ανάλυση του παραδείγματος, δηλ. διευκρινίζοντας το πλήρες δέντρο παραγωγής/ανάλυσης του παραδείγματος. Στα πλαίσια ενός παραδείγματος υπολογισμού του DDL μιας γραμματικής, ας θεωρήσουμε το πλήρες σύνολο των παραγωγών της γραμματικής $G$ που παρουσιάζεται στο πίνακα (Πίνακας 31).

| **The dog ran** | **The dog barked** |
|---|---|
|  |  |
| $\Delta_{DDL} = \log(2) + 4 \cdot \log(1)$ | $\Delta_{DDL} = \log(2) + 4 \cdot \log(1)$ |

**Πίνακας 33: Υπολογίζοντας την συνεισφορά στο μήκος *DDL* δύο προτάσεων.**

| | |
|---|---|
| S → NP VERB1 | 1 |
| S → NP VERB2 | 1 |
| NP → ART NOUN | 2 |
| ART → *the* | 2 |
| NOUN → *dog* | 2 |
| VERB1 → *ran* | 1 |
| VERB2 → *barked* | 1 |

**Πίνακας 34: Η γραμματική $G$ και οι αντίστοιχες συχνότητες κανόνων.**

Προκειμένου να κωδικοποιηθεί η πρόταση *“The dog ran”*, πρέπει να διευκρινιστεί στον παραλήπτη ότι ο πρώτος από τους δύο κανόνες έναρξης ("$S$") πρέπει να χρησιμοποιηθεί. Προκειμένου να κωδικοποιηθούν αυτές οι πληροφορίες χρειάζονται

$\log(Number\ of\ Start\ Rules) = \log(2)$ δυφία. Έπειτα, ο κανόνας πρέπει να αναλυθεί στα συστατικά του, σε αυτήν την περίπτωση "$NP$" και "$VERB1$", και πρέπει να διευκρινιστεί σαφώς να ποιος κανόνας "$NP$" και "$VERB1$" θα χρησιμοποιηθεί. Για την κωδικοποίηση αυτής της πληροφορίας χρειάζονται $\log(1) + \log(1)$ δυφία, καθώς η γραμματική έχει μόνο έναν κανόνα που αρχίζει είτε με "$NP$" ή "$VERB1$". Δεδομένου ότι ο κανόνας "$VERB1$" είναι ένας τερματικός κανόνας, καμία περαιτέρω ανάλυση δεν απαιτείται. Το σύμβολο "$NP$" από την άλλη, μπορεί να αναλυθεί περαιτέρω σε "$ART$" και "$NOUN$", απαιτώντας επιπλέον $\log(1) + \log(1)$ δυφία για να αρθεί αυτή η αμφισημία. Με δεδομένο ότι ούτε το "$ART$", ούτε το "$NOUN$", αναλύονται περαιτέρω, η κωδικοποίηση ολοκληρώνεται. Συνεπώς, η συμβολή αυτής της πρότασης $\Delta_{DDL}$ στο $DDL$ είναι $\log(2) + 4 \cdot \log(1)$ δυφία. Η κωδικοποίηση της δεύτερης πρότασης απαιτεί ακριβώς τον ίδιο αριθμό δυφίων (Πίνακας 33). Σαν αποτέλεσμα, το $DDL$ της γραμματικής $G$, λαμβάνοντας υπόψη το συγκεκριμένο σύνολο προτάσεων, είναι $2 \cdot \log(2) + 8 \cdot \log(1) = 2$ δυφία.

Ένας εύκολος και πρακτικός τρόπος να υπολογιστεί το $DDL$ μιας γραμματικής $G$ είναι να αντιστοιχιστεί μια *συχνότητα* σε κάθε κανόνα της γραμματικής, η οποία είναι ο αριθμός προτάσεων από το σύνολο εκπαίδευσης στις οποίες ο κανόνας συμμετέχει στην παραγωγή/ ανάλυση. Η γραμματική $G$ και οι σχετικές συχνότητες κανόνων παρουσιάζονται στον πίνακα: Πίνακας 34.

Προκειμένου να υπολογιστεί το $DDL$ μιας γραμματικής, εξετάζεται και υπολογίζεται το $DDL$ κάθε κανόνα. Εάν ο κανόνας ανήκει στο υποσύνολο συμβόλων έναρξης, πρέπει σαφώς να προσδιοριστεί ο συγκεκριμένος κανόνας από τους υπόλοιπους κανόνες του υποσυνόλου. Ο αριθμός δυφίων που απαιτούνται για την κωδικοποίηση αυτής της «αποσαφήνισης» δίνεται από το λογάριθμο του αριθμού των κανόνων που μοιράζονται την ίδια κεφαλή. Κατόπιν, για κάθε μη τερματικό σύμβολο $X$ στο σώμα του κανόνα, προστίθεται στο $DDL$ ο αριθμός δυφίων που απαιτείται για την ταυτοποίηση του κανόνα που έχει το σύμβολο $X$ σαν κεφαλή, ο οποίος εξαρτάται από τον συνολικό αριθμό κανόνων που έχουν το $X$ σαν κεφαλή. Τέλος, ο συνολικός αριθμός δυφίων που απαιτείται για να την κωδικοποίηση του κανόνα πρέπει να πολλαπλασιαστεί με τη συχνότητα του κανόνα, υπολογίζοντας τη χρήση του κανόνα στην κωδικοποίηση των παραδειγμάτων.

Για τα άλλα δύο υποσύνολα κανόνα, η κεφαλή κάθε κανόνα δεν πρέπει να κωδικοποιηθεί, καθώς ο απαραίτητος αριθμός δυφίων έχει ήδη υπολογιστεί κατά την εξέταση των κανόνων που ανήκουν στο υποσύνολο των κανόνων συμβόλου έναρξης. Μόνο το μη τερματικά σύμβολα στα σώματα των κανόνων πρέπει να εξεταστούν. Συνεπώς, δεν υπάρχει καμία ανάγκη να εξεταστούν οι κανόνες που ανήκουν στο υποσύνολο τερματικών κανόνων, δεδομένου ότι τα σώματα αυτών περιέχουν μόνο ένα τερματικό σύμβολο που δεν μπορεί να αναλυθεί περαιτέρω. Γενικά, το $DDL$ μιας γραμματικής $G$ μπορεί να υπολογιστεί ως:

$$DDL = \sum_{\substack{\forall\, rule\, in \\ Start\ Symbol\ Subset}} \left( \log\left(H_{Start\ Symbol}\right) + \sum_{\substack{\forall\, X\, in \\ rule\ body}} \log\left(H_X\right) \right) \cdot F_{rule} + \sum_{\substack{\forall\, rule\, in \\ Non\text{-}Terminal\ Subset}} \left( \sum_{\substack{\forall\, X\, in \\ rule\ body}} \log\left(H_X\right) \cdot F_{rule} \right) \tag{5.2}$$

όπου:

- $H_X = \begin{cases} Number\ of\ times\ X\ appears\ as\ Head\ of\ a\ rule \\ 1 \qquad\quad if\ X\ does\ not\ appear\ as\ Head\ of\ a\ rule \end{cases}$
- $F_{rule}$: η συχνότητα του κανόνα.

### 5.3.3 Υπολογιστική πολυπλοκότητα της μέτρησης του μήκους του μοντέλου

Δεδομένου ότι το $MDL$ καθοδηγεί τη διαδικασία αναζήτησης του egGRIDS+, ο υπολογισμός του $ML$ είναι μια θεμελιώδης δράση που πρέπει να επαναληφθεί αρκετές φορές. Συνεπώς, είναι χρήσιμο να γίνει γνωστή η πολυπλοκότητα που συνδέεται με αυτόν τον υπολογισμό. Η διαδικασία υπολογισμού του μήκους του μοντέλου μιας γραμματικής, όπως περιγράφεται από τις εξισώσεις (5.1) και (5.2), μπορεί να πραγματοποιηθεί από τον απλό αλγόριθμο που παρουσιάζεται στην Εικόνα 24.

```
for each Rule in Grammar {
    Count the frequency H_X of the rule head X
}
for each Rule in Grammar {
    for each Symbol in Rule {
        Update statistics:
        Occurrence frequency of each Non Terminal
        Occurrence frequency of each Terminal
        Frequency of Non-Terminals starting rules
        ...
    }
}
Use the various statistics to measure the model length
```

**Εικόνα 24: Ψευδοκώδικας για τον υπολογισμό του μήκους μοντέλου $ML$ μιας γραμματικής.**

Ο παραπάνω αλγόριθμος περιλαμβάνει δύο επαναλήψεις: η πρώτη επανάληψη χρειάζεται αποκλειστικά για τον υπολογισμό του $DDL$, δεδομένου ότι αφορά τη συχνότητα κάθε κεφαλής κανόνα (συνάρτηση $HX$ στην εξίσωση 5.2). Κατά τη διάρκεια του δεύτερου βρόχου, όλα τα σύμβολα μέσα σε κάθε κανόνα εξετάζονται προκειμένου να συλλεχθούν τα δεδομένα για την αποτίμηση των εξισώσεων (5.1) και (5.2) (όπως ο αριθμός των μοναδικών μη τερματικών και τερματικών συμβόλων, ο συνολικώς αριθμός εμφανίσεως όλων των τερματικών/μη τερματικών συμβόλων, καθώς και πόσες φορές ένα μη τερματικό σύμβολο εμφανίζεται σαν κεφαλή ενός κανόνα). Το μήκος του μοντέλου μπορεί να υπολογιστεί άμεσα από αυτά τα δεδομένα.

Συνεπώς, εάν με $R$ συμβολίζεται ο αριθμός των κανόνων στη γραμματική και με $S$ το μέσο μήκος ενός κανόνα, τότε η πολυπλοκότητα της ανωτέρω διαδικασίας είναι $O(R) + O(R) \cdot O(S)$. Ο αριθμός των κανόνων στη γραμματική είναι της ίδιας τάξης μεγέθους με τον αριθμό των παραδειγμάτων εκπαίδευσης $(N)$, τουλάχιστον για την αρχική γραμματική η οποία γενικά έχει το μεγαλύτερο αριθμό κανόνων μεταξύ όλων των διαδοχικών γραμματικών, αφού το $MDL$ κατευθύνει την αναζήτηση σε πιο συμπαγείς γραμματικές από την αρχική. Το μέσο μήκος ενός κανόνα είναι χαρακτηριστικό των παραδειγμάτων εκπαίδευσης, και πολύ συχνά είναι ένας μικρός αριθμός, ασήμαντος συγκριτικά με τον αριθμό των παραδειγμάτων εκπαίδευσης. Σαν αποτέλεσμα, η πολυπλοκότητα της μέτρησης του μήκους μοντέλου $(C_{ML})$ μιας γραμματικής, όσον αφορά τον αριθμό παραδειγμάτων εκπαίδευσης $N$ είναι περίπου γραμμική:

$$C_{ML} = O(N) + O(N) \cdot O(S) = O\big(N \cdot (1 + S)\big) \approx O(N) \qquad (5.3)$$

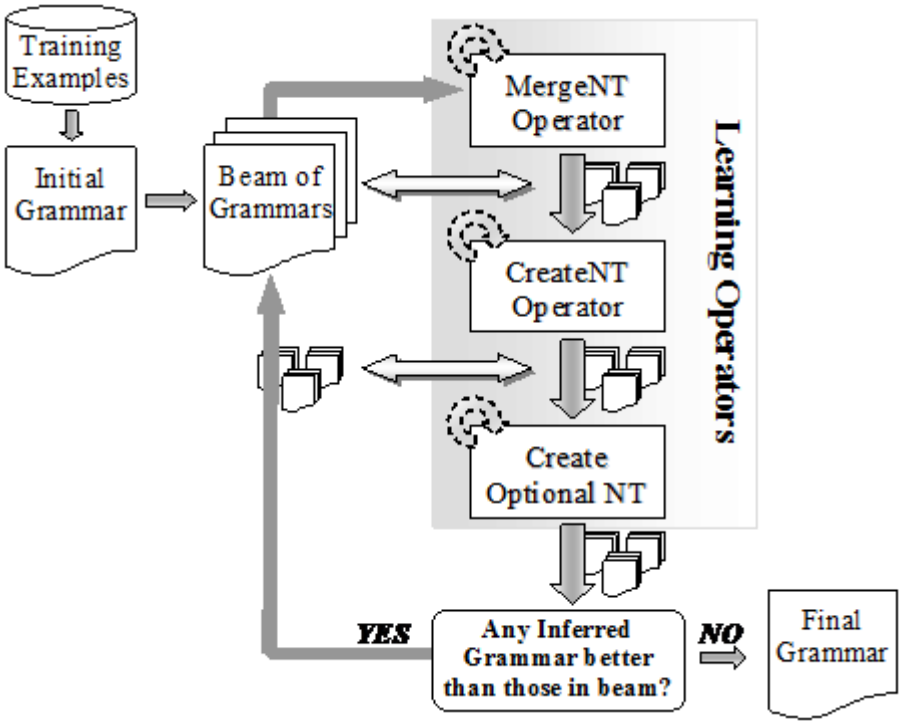
### 5.3.4 Η Αρχιτεκτονική του egGRIDS+ και οι τελεστές εκμάθησης

Όπως έχει ήδη αναφερθεί, ο egGRIDS+ εξάγει γραμματικές από θετικά παραδείγματα εκπαίδευσης, χωρίς απαίτηση οποιωνδήποτε αρνητικών παραδειγμάτων. Συνδυαζόμενο με το γεγονός ότι εξάγει γραμματικές ανεξάρτητες από συμφραζόμενα αντί για κανονικές γραμματικές, ο egGRIDS+ εμφανίζεται ικανός για τη χρήση σε περιοχές όπως η επεξεργασία φυσικής γλώσσας, όπου τα αρνητικά παραδείγματα είναι σπάνια διαθέσιμα και πιο εκφραστικές αναπαραστάσεις από τις κανονικές γραμματικές είναι επιθυμητές. Το σχήμα 2 συνοψίζει την αρχιτεκτονική egGRIDS+.

Όπως και αρκετοί άλλοι αλγόριθμοι επαγωγικής εξαγωγής γραμματικών, έτσι και ο egGRIDS+ χρησιμοποιεί τις προτάσεις εκπαίδευσης προκειμένου να κατασκευάσει μια αρχική, «επίπεδη» γραμματική. Αυτή η αρχική γραμματική κατασκευάζεται με απλή μετατροπή καθενός από τα παραδείγματα εκπαίδευσης σε ένα γραμματικό κανόνα, όπως στον φαίνεται στον πίνακα: Πίνακας 35.

Επομένως, ο αριθμός κανόνων στην αρχική γραμματική (παραλείποντας αυτούς που αντικαθιστούν τα μη-τερματικά σύμβολα με τα τερματικά) είναι ίσος με τον αριθμό των παραδειγμάτων εκπαίδευσης. Αυτή η αρχική γραμματική είναι υπερβολικά συγκεκριμένη, δεδομένου ότι μπορεί να αναγνωρίσει μόνο τις προτάσεις που περιλαμβάνονται στο σύνολο παραδείγματος εκπαίδευσης. Τα δέντρα ανάλυσης που δημιουργούνται από την αρχική γραμματική έχουν βάθος μόνο ενός επιπέδου, χαρακτηρίζοντας τη γραμματική ως επίπεδη.

Η διαδικασία εκπαίδευσης του egGRIDS+ οργανώνεται ως αναζήτηση δέσμης (beam search). Αρχικά, η δέσμη περιέχει μόνο την αρχική γραμματική. Έχοντας μια αρχική υπόθεση (την αρχική γραμματική) στην δέσμη, ο egGRIDS+ χρησιμοποιεί τρεις τελεστές μάθησης προκειμένου να ερευνήσει τον χώρο των ανεξάρτητων από τα συμφραζόμενα γραμματικών. Ο τελεστής “Create NT” – CreateNT δημιουργεί ένα νέο μη τερματικό σύμβολο $X$, το οποίο ορίζεται ως μια ακολουθία δύο υπαρχόντων μη τερματικών συμβόλων. Το $X$ ορίζεται ως ένας νέος κανόνας παραγωγής που αποσυνθέτει το $X$ στα δύο σύμβολά του. Ο τελεστής “Merge NT” – MergeNT συγχωνεύει δύο μη τερματικά σύμβολα σε ένα ενιαίο σύμβολο "$Y$", αντικαθιστώντας όλες τις εμφανίσεις των δύο αρχικών συμβόλων στις κεφαλές και τα σώματα των κανόνων από το "$Y$". Τέλος, ο τελεστής “Create Optional NT” – CreateOptionalNT δημιουργεί ένα αντίγραφο από κανόνα που έχει δημιουργηθεί από τον τελεστή CreateNT, και επισυνάπτει ένα μη τερματικό σύμβολο στον κανόνα, καθιστώντας το σύμβολο αυτό προαιρετικό. Οι τρεις τελεστές δημιουργούν γραμματικές που έχουν είτε την ίδια είτε μεγαλύτερη εκφραστικότητα από τη μητρική γραμματική. Καθώς οι τελεστές δεν αφαιρούν ποτέ κανόνες από μια γραμματική, οι προκύπτουσες γραμματικές έχουν τουλάχιστον την ίδια κάλυψη με την μητρική γραμματική, δηλ. μπορούν να αναγνωρίσουν τουλάχιστον το ίδιο σύνολο προτάσεων.




**Εικόνα 25: Η αρχιτεκτονική του egGRIDS+.**

| | | |
|---|---|---|
| the dog ran | ⇒ | S → THE DOG RAN |
| the dog barked | | S → THE DOG BARKED |
| | | THE → the |
| | | DOG → dog |
| | | RAN → ran |
| | | BARKED → barked |

**Πίνακας 35: Μετατροπή προτάσεων εκπαίδευσης σε μια αρχική γραμματική.**

Κατά τη διαδικασία της εκπαίδευσης, ο egGRIDS+ εναλλάσσεται μεταξύ τριών καταστάσεων, όπου κάθε κατάσταση χαρακτηρίζεται από την επαναλαμβανόμενη εφαρμογή του ίδιου τελεστή. Στην πρώτη κατάσταση (κατάσταση «συγχώνευσης» – *"merge"* mode), ο αλγόριθμος εξετάζει όλους τους τρόπους συγχώνευσης μη τερματικών συμβόλων, με την επανειλημμένη εφαρμογή του τελεστή MergeNT. Αυτή η διαδικασία επαναλαμβάνεται για κάθε γραμματική στην δέσμη, οδηγώντας σε αντίστοιχες διάδοχες γραμματικές. Αφότου έχουν εξεταστεί όλες οι γραμματικές στην δέσμη, οι προκύπτουσες γραμματικές αξιολογούνται. Εάν οποιαδήποτε από αυτές τις γραμματικές αξιολογηθεί εξίσου καλά ή καλύτερα από μια από τις γραμματικές στην δέσμη, η διάδοχη γραμματική αντικαθιστά τη γραμματική από την δέσμη που έχει την χαμηλότερη αξιολόγηση. Εάν τουλάχιστον μια από τις διάδοχες γραμματικές κατορθώσει να εισαχθεί στην δέσμη, ο αλγόριθμος παραμένει στην ίδια κατάσταση.

Εντούτοις, εάν καμία από τις διάδοχες γραμματικές δεν εισαχθεί στην δέσμη, ο egGRIDS+ μεταβαίνει σε διαφορετική κατάσταση. Στην δεύτερη κατάσταση (κατάσταση «δημιουργίας» – *"create"* mode), ο egGRIDS+ εξετάζει όλους τους τρόπους δημιουργίας νέων συμβόλων, ομαδοποιώντας ζεύγη συμβόλων που εμφανίζονται διαδοχικά στην γραμματική, με επανειλημμένη εφαρμογή του τελεστή CreateNT, λειτουργώντας με τον ίδιο τρόπο όπως περιγράφεται ανωτέρω για τον τελεστή

«συγχώνευσης». Πάλι οι καλύτερες διάδοχες γραμματικές επιλέγονται και τοποθετούνται στην δέσμη, αλλάζοντας την τρέχουσα υπόθεση, η οποία θα επεκταθεί περαιτέρω από το τελεστή CreateNT. Εάν καμία από τις διάδοχες γραμματικές δεν εισαχθεί στην δέσμη, ο egGRIDS+ μεταβαίνει στην τελευταία κατάσταση, της «δημιουργίας προαιρετικών» συμβόλων.

Ο τελευταίος των τριών τελεστών, ο τελεστής CreateOptionalNT, εξετάζει όλους τους πιθανούς τρόπους αντιγραφής ενός κανόνα και προσθήκης ενός πρόσθετου συμβόλου στο τέλος του (καθιστώντας αυτό το σύμβολο προαιρετικό). Αυτός ο τελεστής χρησιμοποιείται επανειλημμένα με τον ίδιο τρόπο ακριβώς όπως οι προηγούμενοι δύο τελεστές.

Ο αλγόριθμος συνεχίζει εναλλασσόμενος ανάμεσα στις τρεις καταστάσεις, έως ότου είναι ανίκανος να παράξει μια διάδοχη γραμματική που να αποτιμηθεί σαν καλύτερη από αυτές που βρίσκονται ήδη στην δέσμη. Σε αυτή την περίπτωση, η διαδικασία της εκπαίδευσης τερματίζεται.

### 5.4 Οι τελεστές αναζήτησης του egGRIDS+

Στις ακόλουθες τρεις ενότητες περιγράφουμε λεπτομερώς τους τελεστές αναζήτησης που χρησιμοποιούνται από τον αλγόριθμο egGRIDS+ προκειμένου να ερευνηθεί ο χώρος των πιθανών ανεξάρτητων από τα συμφραζόμενα γραμματικών. Ιδιαίτερη προσοχή δίνεται στον τρόπο που ο κάθε τελεστής επηρεάζει το μήκος περιγραφής του μοντέλου (γραμματικής).

#### 5.4.1 Ο τελεστής “Create NT”

Ο τελεστής CreateNT δημιουργεί ένα νέο μη τερματικό σύμβολο που ορίζεται ως μια ακολουθία δύο υπαρχόντων μη τερματικών συμβόλων. Μετονομάζοντας μια ακολουθία δύο συμβόλων "$X$" και "$Y$" σε ένα νέο μη τερματικό σύμβολο "$Z$", προκαλεί την εισαγωγή στη γραμματική ενός νέου κανόνα της μορφής *“*$Z \rightarrow X\ Y$*”*, ο οποίος αποσυνθέτει το σύμβολο "$Z$" στα συστατικά του. Επιπλέον, όλες οι εμφανίσεις της ακολουθίας *“*$X\ Y$*”* στη γραμματική αντικαθίστανται από το σύμβολο "$Z$". Ο Πίνακας 36 παρουσιάζει την επίδραση αυτού του τελεστή.

| Operator “Create NT”: Creating symbol AP1 | | |
|---|---|---|
| NP → ART ADJ NOUN<br>NP → ART ADJ ADJ NOUN | ⇒ | NP → ART AP1<br>NP → ART ADJ AP1<br>AP1 → ADJ NOUN |

**Πίνακας 36: Η επίδραση του τελεστή CreateNT, όπως παρουσιάζεται στην αναφορά [13].**

Στις γραμματικές φυσικής γλώσσας, τα σύμβολα που δημιουργούνται από αυτόν τον τελεστή θα αντιπροσωπεύσουν συνήθως συγκεκριμένες φράσεις ή δευτερεύουσες προτάσεις. Η εισαγωγή τέτοιων φράσεων είναι χρήσιμη όταν ορισμένοι συνδυασμοί λέξεων (ή υπο-φράσεις) τείνουν να εμφανίζονται μαζί σε προτάσεις. Η επίδραση αυτού του τελεστή είναι μια απλή *αλλαγή αναπαράστασης (representation change)*. Δεδομένου ότι ο τελεστής CreateNT αντικαθιστά απλά μια ακολουθία δύο συμβόλων με ένα νέο, δεν αυξάνει ή δεν μειώνει την κάλυψη μιας γραμματικής, δηλ. η διάδοχη γραμματική αναγνωρίζει ακριβώς το ίδιο σύνολο προτάσεων με την αρχική. Αν και αυτός ο τελεστής δεν αυξάνει την κάλυψη μιας γραμματικής, η αλλαγή αναπαράστασης που επιτυγχάνει είναι σημαντική, προετοιμάζοντας τη γραμματική για την εφαρμογή του δεύτερου τελεστή (του τελεστή «συγχώνευσης μη τερματικών συμβόλων» – “Merge NT”).

Η επίδραση της εφαρμογής του τελεστή CreateNT σε μια γραμματική μπορεί να συνοψιστεί ως εξής:

- Όλες οι εμφανίσεις της ακολουθίας "$X\ Y$" αντικαθίστανται από το μη τερματικός σύμβολο "$Z$".
- Ένας νέος κανόνας της μορφής "$Z \rightarrow X\ Y$" προστίθεται στην γραμματική.
- Η κάλυψη της γραμματικής δεν επηρεάζεται, δεδομένου ότι η εισαγωγή του νέου κανόνα δεν επιτρέπει στη γραμματική για να αναλύσει νέες, προηγουμένως μη αναλυόμενες, προτάσεις.
- Το $DDL$ της γραμματικής δεν επηρεάζεται. Κανένα πρόσθετο δυφίο δεν απαιτείται για να διευκρινιστεί ο νέος κανόνας, δεδομένου ότι είναι ο μόνος κανόνας που έχει ως κεφαλή το σύμβολο "$Z$".
- Το $GDL$ της γραμματικής αυξάνεται εν μέρει λόγω της εισαγωγής ενός νέου κανόνα και ενός νέου μη τερματικού συμβόλου. Η εισαγωγή ενός νέου μη τερματικού συμβόλου σημαίνει ότι περισσότερα δυφία απαιτούνται προκειμένου να αντιπροσωπευθεί κάθε μη τερματικό σύμβολο.
- Το $GDL$ της γραμματικής επίσης εν μέρει μειώνεται, δεδομένου ότι κάθε εμφάνιση της ακολουθίας "$X\ Y$" αντικαθίσταται από σύμβολο "$Z$".

Στις ακόλουθες ενότητες παρουσιάζουμε μερικά αποτελέσματα σχετικά με την πολυπλοκότητα και τη δυναμική συμπεριφορά του τελεστή "CreateNT".

### 5.4.2 Η πολυπλοκότητα της κατάστασης "Create NT"

Σε αυτήν την ενότητα υπολογίζουμε την πολυπλοκότητα του τελεστή CreateNT καθώς και την πολυπλοκότητα ενός πλήρους βήματος «δημιουργίας» (δηλ. μιας ολόκληρης κατάστασης λειτουργίας του αλγορίθμου egGRIDS+ όπου εφαρμόζεται μόνο ο τελεστής CreateNT). Αυτό το βήμα χαρακτηρίζεται από την συνεχή εφαρμογή του τελεστή CreateNT σε όλες τις πιθανές μη τερματικές ακολουθίες (*διγράμματα – bigrams*) και τον υπολογισμό του μήκους περιγραφής όλων των διάδοχων γραμματικών. Η διαδικασία ενός πλήρους βήματος «δημιουργίας» μπορεί να πραγματοποιηθεί με τον απλό αλγόριθμο που παρουσιάζεται στην Εικόνα 26.

Σαν πρώτο βήμα, όλες οι πιθανές ακολουθίες (διγράμματα) πρέπει να προσδιοριστούν. Ο προσδιορισμός απαιτεί μια επανάληψη σε όλα τα σύμβολα όλων των κανόνων. Η πολυπλοκότητα του εντοπισμού όλων των διγραμμάτων είναι $O(N \cdot S)$, υποθέτοντας ότι η διαδικασία αποθήκευσης των διγραμμάτων και τις συχνότητες εμφάνισής τους είναι σταθερής πολυπλοκότητας $\left(O(1)\right)$. Ο αριθμός των παραγόμενων διγραμμάτων (ο οποίος συμβολίζεται με $K$) μπορεί να είναι το πολύ $K = N \cdot S$.

```
for each Rule in Grammar {
  for i=0, i< Rule body symbol number -1, i=i+1 {
    Store_Bigram (symbol[i], symbol[i+1])
  }
}
for each stored Bigram {
  for each Rule in Grammar {
    for i=0, i < Rule body symbol number -1, i=i+1 {
      if Bigram equals "symbol[i] symbol[i+1]" {
        => Replace "symbol[i] symbol[i+1]" with Bigram
      }
    }
  }
Measure Grammar Model Length [O(N · (1 + S))]
}
```

**Εικόνα 26: Ψευδοκώδικας ενός βήματος (κατάστασης λειτουργίας του egGRIDS+) του τελεστή CreateNT.**

Σαν δεύτερο βήμα, ο τελεστής CreateNT πρέπει να εφαρμοστεί σε όλα τα διγράμματα. Η διαδικασία εφαρμογής του τελεστή περιλαμβάνει μια επανάληψη σε όλα τα σύμβολα στα σώματα όλων των κανόνων, έχοντας πολυπλοκότητα $O(N \cdot S)$, υποθέτοντας ότι η διαδικασία αντικατάστασης σύμβολων στα σώματα των κανόνων είναι σταθερής πολυπλοκότητας. Η πολυπλοκότητα της εξέτασης όλων των διγραμμάτων είναι $O\big(K \cdot N \cdot S + N \cdot (1 + S)\big)$. Κατά συνέπεια, η πολυπλοκότητα ολόκληρου του βήματος $C_{CreateNT}$ είναι τετραγωνική:

$$C_{CreateNT} = O(K \cdot N \cdot S + 2 \cdot N \cdot S + N) = O(N^2 \cdot S^2 + 2 \cdot N \cdot S + N) \approx O(N^2).$$

### 5.4.3 Η επίδραση του τελεστή “Create NT” στο μήκος περιγραφής γραμματικής

Υποθέτοντας μια ανεξάρτητη από τα συμφραζόμενα γραμματική $G$, με τα ακόλουθα χαρακτηριστικά:

- $A_{UNT}$: αριθμός μοναδικών μη τερματικών συμβόλων, αποκλείοντας το σύμβολο έναρξης "$S$" και το πρόσθετο σύμβολο "$STOP$".
- $A_{UT}$: αριθμός μοναδικών τερματικών συμβόλων.
- $A_{NT}$: αριθμός εμφανίσεων όλων των μη τερματικών συμβόλων, συμπεριλαμβανομένου του συμβόλου έναρξης της γραμματικής "$S$" αλλά αποκλείοντας το πρόσθετο σύμβολο "$STOP$".
- $A_T$: αριθμός εμφανίσεων όλων των τερματικών συμβόλων.
- $A_S$: αριθμός κανόνων του υποσύνολου των κανόνων συμβόλου έναρξης.
- $A_R$: αριθμός κανόνων του υποσύνολου κανόνων συμβόλου έναρξης και του υποσύνολου μη τερματικών κανόνων. Το $A_R$ μετρά έμμεσα τον αριθμό εμφανίσεων του συμβόλου "$STOP$" που χρησιμοποιείται για τον διαχωρισμό των γραμματικών κανόνων.
- $BF(X,Y)$: αριθμός εμφανίσεων του διγράμματος "$X\,Y$" που πρόκειται να αντικατασταθεί από το "$Z$" στη γραμματική $G$.
- $Bits_{NT} = \log(A_{UNT} + 1)$: αριθμός δυφίων που απαιτούνται για την κωδικοποίηση κάθε εμφάνισης ενός μη τερματικού συμβόλου.

- $Bits_T = \log(A_{UT})$: αριθμός δυφίων που απαιτούνται για την κωδικοποίηση κάθε εμφάνισης ενός τερματικού συμβόλου.
- $Bits_{NT}^{Fin}$: αριθμός δυφίων που απαιτούνται για την κωδικοποίηση κάθε εμφάνισης ενός μη τερματικού συμβόλου στην διάδοχη γραμματική, δηλ. μετά την εφαρμογή του τελεστή στην γραμματική $G$.

Το αρχικό μήκος περιγραφής γραμματικής $GDL_{In}$ της $G$ (πριν την εφαρμογή του τελεστή) μπορεί να υπολογιστεί ως το άθροισμα των δυφίων που απαιτούνται για να κωδικοποιηθούν όλα τα μη τερματικά σύμβολα $(A_{NT} \cdot Bits_{NT} - A_S \cdot Bits_{NT})$, συν τα δυφία που απαιτούνται για την κωδικοποίηση όλων των τερματικών σύμβολων $(A_T \cdot Bits_T)$, συν τα δυφία που απαιτούνται για να κωδικοποιήσουν όλες τις εμφανίσεις του πρόσθετου συμβόλου "$STOP$" $(A_R \cdot Bits_{NT} + 2 \cdot Bits_{NT})$:

$$GDL_{In} = (A_{NT} + A_R - A_S + 2) \cdot Bits_{NT} + A_T \cdot Bits_T$$

Το διάδοχο μήκος περιγραφής γραμματικής $GDL_{Fin}$, μετά την εφαρμογή του τελεστή, μπορεί να υπολογιστεί ως εξής:

$$GDL_{Fin} = \big(A_{NT} + A_R - A_S + 2 + 4 - BF(X,Y)\big) \cdot Bits_{NT}^{Fin} + A_T \cdot Bits_T$$

Δεδομένου ότι η εφαρμογή του τελεστή εισάγει έναν νέο κανόνα της μορφής "$Z \rightarrow X\,Y$", ο συνολικός αριθμός των μη-τερματικών συμβόλων αυξάνεται κατά τέσσερα, τα τρία σύμβολα του νέου κανόνα συν το σύμβολο "$STOP$". Αφ' ενός, κάθε εμφάνιση του "$X\,Y$" αντικαθίσταται από το "$Z$", οδηγώντας στην εξοικονόμηση $BF(X,Y) \cdot Bits_{NT}^{Fin}$ δυφίων. Δεδομένου ότι ο αριθμός των μοναδικών μη τερματικών συμβόλων αυξάνει κατά ένα, $Bits_{NT}^{Fin} = \log\big(A_{UNT}^{Fin} + 1\big) = \log(A_{UNT} + 1 + 1)$. Καθώς το $DDL$ δεν επηρεάζεται από αυτόν τον τελεστή, η συνολική συμβολή αυτού του τελεστή στο μήκος μοντέλου της γραμματικής $(\Delta ML)$ είναι ίση με την αλλαγή στο $GDL$ $(\Delta_{GDL})$:

$$\Delta ML = \big(A_{NT} + A_R - A_S + 2\big) \cdot \log\left(\frac{A_{UNT} + 2}{A_{UNT} + 1}\right) + \big(4 - BF\big(X,Y\big)\big) \cdot \log\big(A_{UNT} + 2\big) \quad (5.4)$$

### 5.4.4 Επιταχύνοντας τον τελεστή CreateNT

Μια σημαντική ιδιότητα του egGRIDS+ σχετίζεται με την υπολογιστική αποδοτικότητα (computational efficiency). Η διαδικασία παραγωγής υποψηφίων (διαδόχων) γραμματικών του αλγορίθμου GRIDS είναι μια υπολογιστικά σύνθετη διαδικασία, δεδομένου ότι είναι βασισμένη στην απαρίθμηση όλων των γραμματικών που μπορούν να παραχθούν από την εφαρμογή ενός τελεστή. Η εφαρμογή ενός τελεστή και η αποτίμηση της νέας γραμματικής απαιτούν μια σημαντική επεξεργασία, ειδικά για μεγάλες γραμματικές, με μερικές χιλιάδες κανόνες και μερικές χιλιάδες τερματικά σύμβολα. Συνεπώς, είναι σημαντικό η διαδικασία εφαρμογής ενός τελεστή και η αποτίμηση της διάδοχης γραμματικής να είναι όσο το δυνατόν υπολογιστικά αποδοτική, καθώς η σύγκληση σε μια τελική γραμματική μπορεί να απαιτήσει έναν μεγάλο αριθμό εφαρμογών αυτής της θεμελιώδους ενέργειας.

Τα αποτελέσματα της ανάλυσης της δυναμικής συμπεριφοράς ενός τελεστή μπορούν να παράσχουν πολύτιμη βοήθεια στη βελτιστοποίηση της εφαρμογής του, καθώς τα αποτελέσματα αυτά μπορούν να χρησιμοποιηθούν προκειμένου να *προβλεφθεί το*

*μήκος μοντέλου μιας διάδοχης γραμματικής, χωρίς εφαρμογή του τελεστή* για να παραχθεί η διάδοχη γραμματική. Επιπλέον, σε μερικές περιπτώσεις αυτά τα αποτελέσματα μπορούν να αποκαλύψουν πληροφορίες χρήσιμες στον περιορισμό του συνόλου συμβόλων, πάνω στο οποίο ο τελεστής πρέπει να εφαρμοστεί, μειώνοντας αποτελεσματικά τις εφαρμογές του τελεστή.

Όσον αφορά τον τελεστή CreateNT, προκειμένου να παραχθεί μια διάδοχη γραμματική με χαμηλότερο μήκος μοντέλου περιγραφής από την γονική, το $\Delta ML$ πρέπει να είναι αρνητικό. Η εφαρμογή αυτού του περιορισμού στην εξίσωση 5.4 οδηγεί στην ακόλουθη σχέση:

$$\left(A_{NT}+A_R-A_S+2\right)\cdot\log\left(\frac{A_{UNT}+2}{A_{UNT}+1}\right)+\left(4-BF\left(X,Y\right)\right)\cdot\log\left(A_{UNT}+2\right)<0\Rightarrow$$

$$BF(X,Y)>\frac{(A_{NT}+A_R-A_S+2)\cdot\log\left(\frac{A_{UNT}+2}{A_{UNT}+1}\right)}{\log(A_{UNT}+2)}+4 \qquad (5.5)$$

Τα αποτελέσματα που επιτυγχάνονται από αυτήν την ανάλυση επισημαίνουν μια ιδιότητα του τελεστή CreateNT, η οποία φαίνεται πολύ λογική και προφανής: η αντικατάσταση ενός διγράμματος από ένα νέο μη τερματικό σύμβολο, πρέπει να εφαρμόζεται μόνο όταν τα δύο σύμβολα του διγράμματος εμφανίζονται συχνά μαζί σε σειρά. Η μόνη παράμετρος της εξίσωσης (5.5) που είναι πραγματικά ανεξάρτητη είναι η συχνότητα του διγράμματος $BF(X,Y)$, καθώς όλες οι υπόλοιπες παράμετροι είναι χαρακτηριστικά της γονικής γραμματικής. Επιπλέον, όσο υψηλότερη είναι η συχνότητα του διγράμματος, τόσο μεγαλύτερη είναι η μείωση που μπορεί να επιτευχθεί στο μήκος περιγραφής μοντέλου της διάδοχης γραμματικής από την εφαρμογή του τελεστή. Συνεπώς, προκειμένου να δημιουργηθούν οι καλύτερες $N$ διάδοχες γραμματικές, αρκεί να εφαρμοστεί ο τελεστής CreateNT χρησιμοποιώντας τα $N$ διγράμματα με τις υψηλότερες συχνότητες εμφάνισης. Αυτή η βελτιστοποίηση είναι πολύ απλή και μειώνει δραστικά τον αριθμό συνδυασμών που πρέπει να εξεταστούν από το τελεστή CreateNT. Επιπλέον, εάν η συχνότητα ενός διγράμματος είναι μικρότερη από το κατώτατο όριο που προβλέπει η εξίσωση (5.5), αυτό το δίγραμμα δεν πρέπει να εξεταστεί καθόλου από τον CreateNT, δεδομένου ότι η διάδοχη γραμματική θα έχει το υψηλότερο μήκος μοντέλου από τη μητρική γραμματική. Η σημαντικότερη πτυχή αυτής της βελτιστοποίησης είναι ότι είναι συνολικά ισοδύναμη με την εξαντλητική απαρίθμηση (exhaustive enumeration), όσον αφορά την επίδρασή της στη συνολική διαδικασία αναζήτησης. Αυτή η βελτιστοποίηση χρησιμοποιήθηκε επίσης ως ευριστικό για την καθοδήγηση της αναζήτησης στο σύστημα SNPR [93], όπου εφαρμόστηκε διαισθητικά. Η θεωρητική ανάλυσή μας υποστηρίζει αυτήν την διαίσθηση και παρέχει την ακριβή σχέση μεταξύ της απαραίτητης συχνότητας του διγράμματος και των δομικών ιδιοτήτων της μητρικής γραμματικής.

Εκτός από τη σημαντική μείωση των διγραμμάτων που πρέπει να εξεταστούν, η θεωρητική ανάλυση προσφέρει τη δυνατότητα να υπολογιστεί το μήκος μοντέλου της διάδοχης γραμματικής άμεσα από την εξίσωση (5.5), χωρίς παραγωγή της διάδοχης γραμματικής. Στην παράγραφο 5.4.2 η πολυπλοκότητα ενός βήματος της κατάστασης του τελεστή CreateNT υπολογίστηκε ως $C_{CreateNT}=O(N^2\cdot S^2+2\cdot N\cdot S+N)$. Εάν η εξίσωση (5.4) χρησιμοποιηθεί για να προβλέψει το μήκος μοντέλου (αντί της εφαρμογής του τελεστή για την παραγωγή της γραμματικής και την μετέπειτα αποτίμηση της γραμματικής), η πολυπλοκότητα μπορεί να μειωθεί σε $C_{CreateNT}=O(2\cdot N\cdot S)\approx O(N)$.

### 5.4.5 Ο τελεστής “Merge NT”

Αυτός ο τελεστής συγχωνεύει δύο μη τερματικά σύμβολα από τη γραμματική σε ένα νέο μη τερματικό σύμβολο. Ο χειριστής MergeNT διαφέρει από το χειριστή CreateNT σε δύο σημαντικές πτυχές:

- Αυτός ο τελεστής δεν εισάγει νέους κανόνες στη γραμματική, αλλά τροποποιεί τους σχετικούς κανόνες ώστε να εφαρμοστούν οι απαραίτητες αλλαγές.
- Δεν υπάρχει καμία προϋπόθεση για την σχετική θέση των υποψηφίων συμβόλων προς συγχώνευση στη γραμματική, π.χ. τα υποψηφίων σύμβολα για να συγχωνευτούν δεν είναι απαραίτητο να εμφανίζονται σαν δίγραμμα.

Η συγχώνευση δύο μη τερματικών σύμβολων "$X$" και "$Y$" σε ένα κοινό μη τερματικό σύμβολο "$Z$", προκαλεί την αντικατάσταση όλων των εμφανίσεων και του "$X$" και του "$Y$" από το "$Z$", περιλαμβάνοντας και αντικαταστάσεις στις κεφαλές των κανόνων. Ο Πίνακας 37 παρουσιάζει την επίδραση αυτού του τελεστή.

| Operator “Merge NT”: Merging symbols AP1 and AP2 | | |
|---|---|---|
| NP → ART AP1<br>NP → ART AP2<br>AP1 → ADJ NOUN<br>AP2 → ADJ AP1 | ⇒ | NP → ART AP3<br>AP3→ ADJ NOUN<br>AP3→ ADJ AP3 |

**Πίνακας 37: Η επίδραση του τελεστή MergeNT, όπως παρουσιάζεται στην αναφορά [13].**

Στις γραμματικές φυσικής γλώσσας, τα σύμβολα που δημιουργούνται από αυτόν τον τελεστή θα αντιπροσωπεύσουν συνήθως συγκεκριμένες κατηγορίες λέξεων (π.χ. ουσιαστικά ή ρήματα) ή *φραστικές κατηγορίες (phrasal classes)* (π.χ. ονοματικές φράσεις). Η επίδραση αυτού του τελεστή *δεν* είναι μια απλή αλλαγή αναπαράστασης (representation change), όπως στην περίπτωση του προηγούμενου τελεστή (CreateNT). Η συγχώνευση δύο διαφορετικών συμβόλων αυξάνει πάντα την κάλυψη της γραμματικής, καθώς το σύνολο προτάσεων που αναγνωρίζονται από τη γραμματική αυξάνεται. Επιπλέον, η χρήση αυτού του τελεστή μπορεί να οδηγήσει διαφορετικούς κανόνες να γίνουν πανομοιότυποι, επιτρέποντας την απαλοιφή των διπλών (πλέον) κανόνων από τη γραμματική, όπως φαίνεται και στο παράδειγμα του παραπάνω πίνακα (Πίνακας 37). Μια άλλη σημαντική επίδραση αυτού του τελεστή είναι η εισαγωγή (άμεσης ή έμμεσης) *αναδρομής (recursion)* στη γραμματική, όπως επίσης φαίνεται στο παράδειγμα του πίνακα: Πίνακας 37.

Προκειμένου να μελετηθεί η επίδραση αυτού του τελεστή στο μήκος περιγραφής μοντέλου μιας γραμματικής, είναι πάλι χρήσιμο να αποσυντεθεί το μήκος περιγραφής μοντέλου στα δύο συστατικά του: το μήκος περιγραφής γραμματικής $(GDL)$ και το μήκος περιγραφής παραγωγών $(DDL)$. Γενικά, η επίδραση του τελεστή MergeNT σε μια γραμματική μπορεί να συνοψιστεί ως εξής:

- Όλες οι εμφανίσεις των μη τερματικών συμβόλων "$X$" και "$Y$" αντικαθίστανται από το μη τερματικό σύμβολο "$Z$".
- Κανόνες μπορούν να αφαιρεθούν, καθώς η συγχώνευση δύο συμβόλων μπορεί να οδηγήσει σε διπλότυπους κανόνες.
- Η κάλυψη της γραμματικής πάντα αυξάνει.
- Το $GDL$ της γραμματικής πάντα μειώνεται. Αυτή η μείωση έχει δύο αιτίες:

a) Δεδομένου ότι δύο μη τερματικά σύμβολα συγχωνεύονται σε ένα ενιαίο, ο αριθμός δυφίων που απαιτείται για την κωδικοποίηση των μη-τερματικών συμβόλων μειώνεται.
b) Δεδομένου ότι αυτός ο τελεστής μπορεί να αποβάλει κανόνες, ο συνολικός αριθμός συμβόλων στη γραμματική μπορεί να μειωθεί.

- Το $DDL$ της γραμματικής μπορεί είτε να αυξηθεί, είτε να μειωθεί, λόγω των ακόλουθων αιτιών:

a) Ο αριθμός των κανόνων που αναλύουν το νέο σύμβολο "$Z$" είναι μεγαλύτερος από αυτόν για καθένα από τα δύο συγχωνευμένα σύμβολα ("$X$", "$Y$"). Κατά συνέπεια, περισσότερα δυφία απαιτούνται για να κωδικοποιήσουν κάθε εμφάνιση του νέου συμβόλου έναντι των δύο αντικατασταθέντων.
b) Η εισαγωγή του νέου συμβόλου "$Z$" μπορεί να οδηγήσει κανόνες που γίνονται ίδιοι σε αποβολή από τη γραμματική. Διπλότυποι κανόνες μπορούν αν προκύψουν είτε επειδή οι κεφαλές τους έγιναν ίδιες (το σύμβολο "$Z$"), είτε επειδή τα σώματά τους έγιναν ίδια μετά από την αντικατάσταση των "$X$", "$Y$", με "$Z$". Ένα παράδειγμα της τελευταίας περίπτωσης συμβαίνει με τους δύο κανόνες "$NP$" στον Πίνακας 7. Η αποβολή των διπλότυπων κανόνων μπορεί να μειώσει τον αριθμό κανόνων που μοιράζονται το ίδιο σύμβολο κεφαλής (είτε το σύμβολο "$Z$" είτε οποιοδήποτε άλλο μη τερματικό σύμβολο), άρα και τον αριθμό δυφίων που απαιτείται για να κωδικοποιήσει τις εμφανίσεις αυτού του συμβόλου.

### 5.4.6 Η πολυπλοκότητα της κατάστασης “ Merge NT”

Σε αυτήν την ενότητα υπολογίζουμε την πολυπλοκότητα του τελεστή MergeNT καθώς και την πολυπλοκότητα ενός πλήρους βήματος «συγχώνευσης» (δηλ. μιας ολόκληρης κατάστασης λειτουργίας του αλγορίθμου egGRIDS+ όπου εφαρμόζεται μόνο ο τελεστής MergeNT). Αυτό το βήμα χαρακτηρίζεται από την συνεχή εφαρμογή του τελεστή MergeNT σε όλους τους πιθανούς συνδυασμούς όλων των πιθανών ζευγών μη τερματικών συμβόλων της γραμματικής, και τον υπολογισμό του μήκους περιγραφής όλων των διάδοχων γραμματικών. Η διαδικασία ενός πλήρους βήματος «συγχώνευσης» μπορεί να πραγματοποιηθεί με τον απλό αλγόριθμο που παρουσιάζεται στην Εικόνα 27.

Σαν πρώτο βήμα, όλα τα μη τερματικά σύμβολα στη γραμματική πρέπει να προσδιοριστούν, ενέργεια η οποία απαιτεί μια επανάληψη σε όλα τα σύμβολα σε όλα τα σώματα των κανόνων. Η πολυπλοκότητα αυτής της δράσης είναι $O(N \cdot S)$, υποθέτοντας ότι η διαδικασία αποθήκευσης αυτών των σύμβολων είναι σταθερής πολυπλοκότητας. Ο αριθμός των μοναδικών μη τερματικών συμβόλων μπορεί να είναι το πολύ $N$ (ο μέγιστος αριθμός κανόνων στη γραμματική).

```
    for each Rule in Grammar {
      for each Symbol in Rule {
        Store(Symbol)
      }
    }
    for each Symbol_1 {
      for each Symbol_2 after Symbol_1 {
        for each Rule in Grammar {
          for each Symbol in Rule {
            if Symbol equals to either Symbol_1 or Symbol_2
{
                Replace Symbol with Symbol_1
            }
          }
        }
        Measure Grammar Model Length O(N · (1 + S))
      }
    }
```

**Εικόνα 27: : Ψευδοκώδικας ενός βήματος (κατάστασης λειτουργίας του egGRIDS+) του τελεστή MergeNT.**

Σαν δεύτερο βήμα, μια τετραγωνική αναζήτηση εκτελείται ώστε να εφαρμοστεί ο τελεστής MergeNT σε όλους τους συνδυασμούς ανά δυάδες, όλων των μη τερματικών συμβόλων, ο αριθμός των οποίων είναι $N^2/2$. Η διαδικασία εφαρμογής του τελεστή περιλαμβάνει μια επανάληψη σε όλα τα σύμβολα, σε όλα τα σώματα κανόνων, έχοντας πολυπλοκότητα $O(N \cdot S)$, υποθέτοντας ότι η διαδικασία αντικατάστασης σύμβολων στα σώματα των κανόνων είναι σταθερής πολυπλοκότητας. Η πολυπλοκότητα της εξέτασης όλων των συνδυασμών συμβόλων είναι $O\left(N^2/2 \cdot \left(N \cdot S + N \cdot (1+S)\right)\right)$. Κατά συνέπεια, η πολυπλοκότητα ολόκληρου του βήματος $C_{MergeNT}$ είναι κυβική:

$$C_{MergeNT} = O\left(\frac{N^3 \cdot (2 \cdot S + 1)}{2} + N \cdot S\right) \approx O(N^3)$$

### 5.4.7 Η επίδραση του τελεστή “Merge NT” στο μήκος περιγραφής γραμματικής

Υποθέτοντας μια ανεξάρτητη από τα συμφραζόμενα γραμματική $G$, με τα ακόλουθα χαρακτηριστικά:

- $A_{UNT}$: αριθμός μοναδικών μη τερματικών συμβόλων, αποκλείοντας το σύμβολο έναρξης "$S$" και το πρόσθετο σύμβολο "$STOP$".
- $A_{UT}$: αριθμός μοναδικών τερματικών συμβόλων.
- $A_{NT}$: αριθμός εμφανίσεων όλων των μη τερματικών συμβόλων, συμπεριλαμβανομένου του συμβόλου έναρξης της γραμματικής "$S$" αλλά αποκλείοντας το πρόσθετο σύμβολο "$STOP$".
- $A_T$: αριθμός εμφανίσεων όλων των τερματικών συμβόλων.
- $A_S$: αριθμός κανόνων του υποσύνολου των κανόνων συμβόλου έναρξης.
- $A_R$: αριθμός κανόνων του υποσύνολου κανόνων συμβόλου έναρξης και του υποσύνολου μη τερματικών κανόνων. Το $A_R$ μετρά έμμεσα τον αριθμό εμφανίσεων του συμβόλου "$STOP$" που χρησιμοποιείται για τον διαχωρισμό των γραμματικών κανόνων.
- $Bits_{NT} = \log(A_{UNT} + 1)$: αριθμός δυφίων που απαιτούνται για την κωδικοποίηση κάθε εμφάνισης ενός μη τερματικού συμβόλου.
- $Bits_T = \log(A_{UT})$: αριθμός δυφίων που απαιτούνται για την κωδικοποίηση κάθε εμφάνισης ενός τερματικού συμβόλου.
- $Bits_{NT}^{Fin}$: αριθμός δυφίων που απαιτούνται για την κωδικοποίηση κάθε εμφάνισης ενός μη τερματικού συμβόλου στην διάδοχη γραμματική, δηλ. μετά την εφαρμογή του τελεστή στην γραμματική $G$.

Όπως και στην περίπτωση του τελεστή CreateNT, το αρχικό μήκος περιγραφής γραμματικής $GDL_{In}$ της γραμματικής $G$ (πριν την εφαρμογή του τελεστή), μπορεί να υπολογιστεί ως εξής:

$$GDL_{In} = (A_{NT} + A_R - A_S + 2) \cdot \log(A_{UNT} + 1) + A_T \cdot \log(A_{UT})$$

Όμοια, το διάδοχο μήκος περιγραφής γραμματικής $GDL_{Fin}$, μετά την εφαρμογή του τελεστή, μπορεί να υπολογιστεί ως εξής:

$$GDL_{Fin} = (A_{NT} + A_R - A_S + 2) \cdot \log(A_{UNT}) + A_T \cdot \log(A_{UT}) - \mathrm{E}$$

Ο πρώτος όρος της παραπάνω εξίσωσης αντιπροσωπεύει τα δυφία που απαιτούνται για να κωδικοποιηθούν όλα τα μη-τερματικά σύμβολα. Δεδομένου ότι η εφαρμογή του χειριστή έχει αντικαταστήσει κάθε εμφάνιση είτε του "$X$", είτε του "$Y$" με το "$Z$", ο συνολικός αριθμός των μοναδικών μη τερματικών συμβόλων έχει μειωθεί κατά ένα. Κατά συνέπεια, $Bits_{NT}^{Fin} = \log(A_{UNT})$. Ο όρος $E$ αντιπροσωπεύει την αναμενόμενη μείωση στο $GDL$ που μπορεί να προκύψει εάν η εφαρμογή αυτού του τελεστή οδηγήσει σε διπλότυπους κανόνες, οι οποίοι θα αφαιρεθούν. Για κάθε κανόνα που αποβάλλεται από τη γραμματική $G$, πρέπει να αφαιρέσουμε τα δυφία που απαιτούνται για την κωδικοποίησης της κεφαλής και του σώματος του κανόνα, καθώς και του συμβόλου "$STOP$" που τερματίζει τον κανόνα.

$$\mathrm{E} = \sum_{j \in \Omega_1} \left(L_j + 1\right) \cdot \log(A_{UNT}) + |\Omega_2| \cdot \left(\log(A_{UNT}) + \log(A_{UT})\right) + \sum_{j \in \Omega_3} \left(L_j + 2\right) \cdot \log(A_{UNT}) \quad (5.6a)$$

Όπου:

- $\Omega_1$: το σύνολο κανόνων από το υποσυνόλου συμβόλων έναρξης, που αποβάλλεται από την $G$.
- $\Omega_2$: το σύνολο κανόνων από το υποσύνολο τερματικών συμβόλων, που αποβάλλεται από την $G$.
- $\Omega_3$: το σύνολο κανόνων από το υποσύνολο μη-τερματικών συμβόλων, που αποβάλλεται από την $G$.
- $L_j$: ο αριθμός μη-τερματικών συμβόλων στο σώμα του κανόνα $j$.

Στην εξίσωση 5.6a, ο όρος $\left(L_j + 1\right) \cdot \log(A_{UNT})$ αντιπροσωπεύει τα δυφία που απαιτούνται για την κωδικοποίηση των μη τερματικών συμβόλων στο σώμα του κανόνα (μαζί με το σύμβολο "$STOP$"), ενός κανόνα από το υποσύνολο συμβόλου έναρξης. Ο όρος $\log(A_{UNT}) + \log(A_{UT})$ αντιπροσωπεύει τα δυφία που απαιτούνται για να κωδικοποιήσουν την κεφαλή και το τερματικό σύμβολο στο σώμα ενός κανόνα από το υποσύνολο τερματικών κανόνων. Ο τελευταίος όρος $\left(L_j + 2\right) \cdot \log(A_{UNT})$ αντιπροσωπεύει τα δυφία που απαιτούνται για να κωδικοποιήσουν την κεφαλή και τα μη τερματικά του σώματος (μαζί με το σύμβολο "$STOP$"), ενός κανόνα από το υποσύνολο μη τερματικών συμβόλων.

Η αλλαγή $\Delta_{GDL}$ στο μήκος περιγραφής μοντέλου $GDL$ λόγω αυτού του τελεστή είναι:

$$\Delta_{GDL} = (A_{NT} + A_R - A_S + 2) \cdot \log\left(\frac{A_{UNT}}{A_{UNT} + 1}\right) - \left( \sum_{j \in \Omega_1} \left(L_j + 1\right) \cdot \log(A_{UNT}) + |\Omega_2| \cdot \left(\log(A_{UNT}) + \log(A_{UT})\right) + \sum_{j \in \Omega_3} \left(L_j + 2\right) \cdot \log(A_{UNT}) \right) \quad (5.6b)$$

Ο πρώτος όρος στην εξίσωση 5.6b είναι αρνητικός και μπορεί να θεωρηθεί «σταθερός», καθώς εξαρτάται από τα χαρακτηριστικά της αρχικής γραμματικής, ανεξάρτητα από τα δύο συγκεκριμένα σύμβολα που συγχωνεύονται. Ο δεύτερος όρος της εξίσωσης 5.6b είναι επίσης πάντα αρνητικός (λόγω του αρνητικού πρόσημου). Κατά συνέπεια, η μείωση του $GDL$ είναι μεγαλύτερη όσο μεγαλύτερος είναι ο αριθμός των διπλότυπων κανόνων που αφαιρούνται από τη γραμματική και όσο το μήκος τους αυξάνει.

### 5.4.8 Η επίδραση του τελεστή “Merge NT” στο μήκος περιγραφής παραγωγών

Με δεδομένη γραμματική $G$ και ένα σύνολο προτάσεων που αναγνωρίζονται από την $G$, το μήκος περιγραφής παραγωγών $DDL$ μπορεί να υπολογιστεί ως εξής:

$$DDL = \sum_{\substack{\forall\, rule\ in \\ Start\ Symbol\ Subset}} \left( \log\left(H_{Start\ Symbol}\right) + \sum_{\substack{\forall\, X\ in \\ rule\ body}} \log(H_X) \right) \cdot F_j^S + \sum_{\substack{\forall\, rule\ in \\ Non-Terminal\ Subset}} \left( \sum_{\substack{\forall\, X\ in \\ rule\ body}} \log(H_X) \cdot F_j^S \right) \quad (5.6c)$$

Όπου:

- Το $X$ αντιπροσωπεύει κάθε μη τερματικό σύμβολο στην κεφαλή και το σώμα ενός κανόνα $j$.
- $H_X = \begin{cases} Number\ of\ times\ X\ appears\ as\ Head\ of\ a\ rule \\ 1 \qquad\quad if\ X\ does\ not\ appear\ as\ Head\ of\ a\ rule \end{cases}$
- Το $F_j^S$ συμβολίζει τον αριθμό των προτάσεων που περιλαμβάνουν τον κανόνα $j$ στην παραγωγή/ανάλυσή τους.

Δεδομένου ότι ενδιαφερόμαστε κυρίως για τη μελέτη της επίδρασης που αυτός ο τελεστής έχει στη γραμματική $G$, υπολογίζουμε κατ' ευθείαν την αλλαγή $\Delta_{DDL}$ στο $DDL$ λόγω αυτού του τελεστή. Προκειμένου να υπολογιστεί αυτή η αλλαγή, διαιρούμε την επίδραση αυτού του τελεστή σε δύο «φαινόμενα», τα οποία θα μελετηθούν ανεξάρτητα το ένα από το άλλο. Το πρώτο «φαινόμενο» είναι η αντικατάσταση όλων των κεφαλών των κανόνων που είναι "$X$" ή "$Y$" με "$Z$", χωρίς εξάλειψη οποιωνδήποτε διπλότυπων κανόνων που πιθανών να εμφανιστούν. Θα αναφερθούμε σε αυτό το γεγονός ως «συγχώνευση συνόλων από κανόνες». Το δεύτερο «φαινόμενο» είναι η αποβολή των διπλότυπων κανόνων, ώστε να παραμείνει μόνο ένας. Όπως έχει ήδη αναφερθεί, η εφαρμογή του τελεστή MergeNT μπορεί να οδηγήσει σε διπλούς κανόνες με δύο τρόπους. Οι κανόνες που μοιράζονται τα ίδια σώματα αλλά οι κεφαλές τους είναι είτε "$X$" είτε "$Y$", θα γίνουν ίδιοι όταν αντικατασταθούν οι κεφαλές τους από το νέο σύμβολο "$Z$". Χαρακτηριστικό παραδείγματα αποτελούν οι τερματικοί κανόνες, όπου ένα τερματικό σύμβολο μπορεί να ανήκει σε δύο κατηγορίες: ένας από τους δύο κανόνες πρέπει να αποβληθεί, εάν τα μη τερματικά σύμβολα που αντιπροσωπεύουν τις δύο κατηγορίες συγχωνευτούν. Η δεύτερη περίπτωση όπου εμφανίζονται διπλότυποι κανόνες είναι η αντικατάσταση είτε του "$X$" είτε του "$Y$", από το "$Z$" στα σώματα κανόνων. Είναι δυνατό και οι δύο αυτές περιπτώσεις να συμβούν ταυτόχρονα σε κάποιον κανόνα. Στις ακόλουθες παραγράφους εξετάζουμε την επίδραση κάθε ενός από τα δύο φαινόμενα στο $\Delta_{DDL}$.

### Συγχώνευση συνόλων από κανόνες

Όπως έχει ήδη αναφερθεί, ο τελεστής MergeNT δεν εισάγει νέους κανόνας στην γραμματική. Αντί αυτού, ο τελεστής τροποποιεί τις κεφαλές μερικών κανόνων για να υλοποιηθούν οι απαραίτητες αλλαγές. Αυτό παρουσιάζεται στον πίνακα: Πίνακας 38.

Στην αρχική γραμματική $G_{In}$ του πίνακα (Πίνακας 38), τα δυφία που απαιτούνται προκειμένου να προσδιοριστεί μονοσήμαντα κάθε κανόνας που έχει ως κεφαλή το σύμβολο "$X$" είναι $log(3)$, ενώ τα δυφία που απαιτούνται για τον προσδιορισμό ενός κανόνα που έχει ως κεφαλή το σύμβολο "$Y$" είναι $log(2)$. Στην διάδοχη γραμματική $G_{Fin}$, ο προσδιορισμός ενός κανόνα που έχει το σύμβολο "$Z$" σαν κεφαλή απαιτεί $log(5)$

δυφία. Στη γενική περίπτωση, τα δυφία που απαιτούνται για να προσδιορίσουν κάθε κανόνα που έχει το νέο σύμβολο "$Z$" ως κεφαλή είναι:

$$Bits_Z = \log(F_X + F_Y)$$

Όπου:

- Το $F_X$ συμβολίζει τον αριθμό των κανόνων που έχουν το σύμβολο "$X$" ως κεφαλή.
- Το $F_Y$ συμβολίζει τον αριθμό των κανόνων που έχουν το σύμβολο "$Y$" ως κεφαλή.

| **Operator "Merge NT": Merging symbols X and Y** | | | | |
|---|---|---|---|---|
| ***Initial Grammar ($G_{In}$)*** | | | ***Final Grammar ($G_{Fin}$)*** | |
| X | $\rightarrow A_1\ B_1\ C_1$ | | Z | $\rightarrow A_1\ B_1\ C_1$ |
| X | $\rightarrow A_2\ B_2\ C_2$ | | Z | $\rightarrow A_2\ B_2\ C_2$ |
| X | $\rightarrow A_3\ B_3\ C_3$ | | Z | $\rightarrow A_3\ B_3\ C_3$ |
| Y | $\rightarrow D_1\ E_1\ F_1$ | $\Rightarrow$ | Z | $\rightarrow D_1\ E_1\ F_1$ |
| Y | $\rightarrow D_2\ E_2\ F_2$ | | Z | $\rightarrow D_2\ E_2\ F_2$ |
| K | $\rightarrow L_1\ M_1\ N_1$ | | K | $\rightarrow L_1\ M_1\ N_1$ |
| … | | | … | |

**Πίνακας 38: Συγχωνεύοντας σύνολα από κανόνες.**

Προκειμένου να υπολογιστεί η συμβολή των συμβόλων "$X$" και "$Y$" στο $DDL$ της αρχικής γραμματικής $G_{In}$, πρέπει να μετρήσουμε την συχνότητα εμφάνισης κάθε συμβόλου στα σώματα των κανόνων. Συνεπώς, η συνολική συμβολή των συμβόλων "$X$" και "$Y$" στο $DDL$ είναι:

$$C_{In} = \alpha \cdot log(F_X) + \beta \cdot log(F_Y)$$

Όπου:

- $\alpha = \sum_{j \in G_{In}} N_j^X \cdot F_j^S$, $\beta = \sum_{j \in G_{In}} N_j^Y \cdot F_j^S$
- Το $F_j^S$ συμβολίζει τον αριθμό των προτάσεων που περιλαμβάνουν τον κανόνα $j$ στην παράγωγή/ανάλυσή τους.
- Το $j$ αντιπροσωπεύει έναν κανόνα της αρχικής γραμματικής $G_{In}$.
- Το $N_j^X$ συμβολίζει τον αριθμό εμφανίσεων του συμβόλου "$X$" στο σώμα του κανόνα $j$.
- Το $N_j^Y$ αντιπροσωπεύει τον αριθμό εμφανίσεων του συμβόλου "$Y$" στο σώμα του κανόνα $j$.

Όμοια, η συμβολή του νέου συμβόλου "$Z$" στο $DDL$ της διάδοχης γραμματικής $G_{Fin}$ μπορεί να υπολογιστεί ως εξής:

$$C_{Fin} = (\alpha + \beta) \cdot log(F_X + F_Y)$$

Συνεπώς, η αλλαγή $\Delta C$ του $DDL$ λόγω της συγχώνευσης των δύο συνόλων κανόνων είναι:

$$\Delta C = \alpha \cdot log\left(1 + \frac{F_Y}{F_X}\right) + \beta \cdot log\left(1 + \frac{F_X}{F_Y}\right). \quad (5.6d)$$

**Αφαίρεση διπλότυπων κανόνων**

Η αφαίρεση των κανόνων που έχουν γίνει ίδιοι είναι μια παρενέργεια της εφαρμογής του τελεστή MergeNT, και σε μεγάλο μέρος εξαρτάται από τα χαρακτηριστικά της υπό εξέταση γραμματικής. Σε περιπτώσεις όπου η αφαίρεση κανόνων κρίνεται αναγκαία, η συνολική συνεισφορά στο $DDL$ είναι:

$$\Delta C = \sum_{j \in G_{Fin}} \left( \sum_{k \in \Theta, j} M_k \right) \cdot F_j^S \quad (5.6e)$$

Όπου:

- Το $\Theta$ είναι το σύνολο των συμβόλων κεφαλής όλων των κανόνων που αφαιρέθηκαν.
- Το $k$ αντιπροσωπεύει κάθε σύμβολο στο σώμα ενός κανόνα $j$, το οποίο είναι επίσης μέλος του συνόλου $\Theta$. Πρέπει να σημειωθεί ότι αφού το $k$ είναι μέλος του $\Theta$, δεν μπορεί να είναι το νέο σύμβολο "$Z$".
- Το $M_k$ συμβολίζει τη μείωση του αριθμού των δυφίων που απαιτούνται για να κωδικοποιηθεί μια εμφάνιση του συμβόλου $k$. $\left(M_k = \log\left(\frac{Number\ of\ rules\ having\ k\ as\ head\ in\ G_{Fin}}{Number\ of\ rules\ having\ k\ as\ head\ in\ G_{In}}\right)\right)$. Αξίζει να σημειωθεί ότι το $M_k$ είναι πάντα αρνητικό.
- Το $F_j^S$ συμβολίζει τον αριθμό των προτάσεων που περιλαμβάνουν τον κανόνα $j$ στη παραγωγή/ανάλυσή τους.

### 5.4.9 Συνολική επίδραση του τελεστή "Merge NT" στο μήκος περιγραφής μοντέλου γραμματικής

Η συνολική επίδραση του τελεστή MergeNT στο μήκος περιγραφής μοντέλου μιας γραμματικής $\Delta ML$ μπορεί να υπολογιστεί με το συνδυασμό των εξισώσεων (5.6b), (5.6d) και (5.6e):

$$\begin{aligned}\Delta ML = & \left(A_{NT} + A_R - A_S + 2\right)\cdot\log\left(\frac{A_{UNT}}{A_{UNT}+1}\right) - \\ & \left(\begin{array}{l}\sum_{j\in\Omega_1}\left(L_j+1\right)\cdot\log\left(A_{UNT}\right) + \left|\Omega_2\right|\cdot\left(\log\left(A_{UNT}\right)+\log\left(A_{UT}\right)\right) + \\ \sum_{j\in\Omega_3}\left(L_j+2\right)\cdot\log\left(A_{UNT}\right)\end{array}\right) + \\ & \left(\sum_{j\in G_{In}} N_j^X\cdot F_j^S\right)\cdot\log\left(1+\frac{F_Y}{F_X}\right) + \left(\sum_{j\in G_{In}} N_j^Y\cdot F_j^S\right)\cdot\log\left(1+\frac{F_X}{F_Y}\right) + \\ & \sum_{j\in G_{Fin}}\left(\sum_{k\in\Theta,j} M_k\right)\cdot F_j^S\end{aligned} \tag{5.6}$$

Οι πρώτοι δύο όροι της εξίσωσης (5.6) αντιστοιχούν στην αλλαγή στο $GDL$. Όπως αναφέρεται στην παράγραφο 0, η μείωση του $GDL$ είναι μεγαλύτερη όσο ο αριθμός των διπλών κανόνων που αποβάλλονται από τη γραμματική αυξάνει, και όσο αυξάνει και το μήκος τους (σε αριθμό συμβόλων). Ο τρίτος και τέταρτος όρος της εξίσωσης (5.6) αντιστοιχούν στην αλλαγή στο $DDL$. Ο τρίτος όρος αντιπροσωπεύει τη συγχώνευση των συνόλων από κανόνες και είναι πάντα θετικός. Οι ανεξάρτητες παράμετροι είναι τέσσερις, η σχετική συχνότητα των δύο συμβόλων "$X$" και "$Y$" στις κεφαλές των κανόνων $\left(\alpha = {F_X}/{F_Y}\right)$, οι συχνότητες εμφάνισης των συμβόλων "$X$" και "$Y$" στα σώματα κανόνων $\left(N_j^X, N_j^Y\right)$, και η συχνότητα του κανόνα $F_j^S$. Οι δύο λογάριθμοι του τρίτου όρου στην εξίσωση 5.6 φθάνουν στην ελάχιστη τιμή $\alpha = 1\ F_X = F_Y$. Συνεπώς, αυτός ο όρος είναι μικρός όταν το $\alpha$ είναι ίσο με την μονάδα, και τα με α το ένα και τα "$X$" και "$Y$" δεν εμφανίζονται συχνά στα σώματα κανόνων που σχετίζονται με υψηλές συχνότητες $F_j^S$. Τέλος, ο τέταρτος όρος αντιπροσωπεύει την αποβολή των διπλότυπων κανόνων και είναι πάντα αρνητικός ή μηδέν, ανάλογα με την ύπαρξη/ανυπαρξία των διπλότυπων κανόνων.

### 5.4.10 Επιταχύνοντας τον τελεστή MergeNT

Ο τελεστής MergeNT μπορεί να δημιουργήσει καλύτερες διάδοχες γραμματικές από την μητρική, μόνο εάν $\Delta ML < 0$. Ο πρώτος όρος της εξίσωσης 5.6 είναι σταθερός (για μια δεδομένη μητρική γραμματική $G$) και πάντα αρνητικός. Ο δεύτερος και τέταρτοw όρος είναι επίσης πάντα αρνητικοί ή μηδέν, ανάλογα με το εάν υπάρχουν στη διάδοχη γραμματική διπλότυποι κανόνες. Τέλος, ο τρίτος όρος είναι πάντα θετικός. Συνεπώς, προκειμένου να ικανοποιηθεί η συνθήκη $\Delta ML < 0$, το άθροισμα του πρώτου, δεύτερου, και τέταρτου όρου πρέπει να είναι μεγαλύτερο, σε απόλυτη τιμή, από τον τρίτο όρο.

Η διαίσθηση σε αυτήν την περίπτωση προτείνει την εφαρμογή του τελεστή MergeNT σε μη-τερματικά σύμβολα που εμφανίζονται συχνά σε παρόμοιο περιβάλλον. Στην πραγματικότητα, ένας τέτοιος ευριστικό χρησιμοποιήθηκε στο σύστημα SNPR [93] για τη μείωση του αριθμού γραμματικών που πρέπει να εξεταστούν. Σαν «παρόμοιο

περιβάλλον» μπορεί να θεωρηθούν κανόνες με μικρές διαφορές, πέρα από τα δύο σύμβολα που θα συγχωνευθούν: αυτοί οι κανόνες είναι ιδιαίτερα πιθανό κα καταλήξουν διπλότυποι και κάποιοι να αφαιρεθούν από την γραμματική, όταν συγχωνευθούν πραγματικά τα δύο σύμβολα. Κατά συνέπεια, αυτή η ιδιότητα του τελεστή MergeNT που προκύπτει από «κοινή λογική», μπορεί να συνδεθεί με τη αποβολή διπλότυπων κανόνων στο πλαίσιό του egGRIDS+. Ωστόσο, από την εξίσωση (5.6) μπορούμε εύκολα να καταλήξουμε στο συμπέρασμα ότι αν και η αποβολή κανόνων είναι ένας σημαντικός παράγοντας, δεν είναι ο μοναδικός. Ένας σημαντικός όρος για την επιτυχή εφαρμογή αυτού του τελεστή είναι μια συναλλαγή (trade-off) μεταξύ του αριθμού, της συχνότητας και του μήκους των διπλότυπων κανόνων, καθώς και της *κατανομής των δύο συγχωνευμένων συμβόλων στη γραμματική* (τρίτος όρος της εξίσωσης 5.6). Συνεπώς, η ύπαρξη απλά συμβόλων σε παρόμοια περιβάλλοντα δεν μπορεί να χρησιμοποιηθεί ως κριτήριο για την οδήγηση της αναζήτησης, καθώς αυτή η ιδέα είναι βασισμένη πρώτιστα στην αποβολή κανόνων, αλλά αποτυγχάνει να διαχειριστεί έναν εξίσου σημαντικό παράγοντα: τις σχετικές συχνότητες των δύο συμβόλων ως κεφαλές κανόνων και τις κατανομές τους στη γραμματική. Αντί αυτού του «ανακριβούς» ευριστικού, ο egGRIDS+ χρησιμοποιεί την εξίσωση (5.6) για να κατευθύνει την αναζήτηση στον χώρο των πιθανών γραμματικών, με τον υπολογισμό του μήκους περιγραφής του μοντέλου γραμματικής χωρίς πραγματικά να παραγάγει τις αντίστοιχες γραμματικές. Ο υπολογισμός της εξίσωσης (5.6) είναι σημαντικά ταχύτερος από την παράγωγή και αποτίμηση της αντίστοιχης διάδοχης γραμματικής.

Στην ενότητα 5.4.6 η πολυπλοκότητα ενός βήματος της κατάστασης του MergeNT βρέθηκε για να είναι $C_{MergeNT} = O(1/2 \cdot N^3 \cdot (2 \cdot S + 1) + N \cdot S)$. Η πρόβλεψη του μήκους περιγραφής μοντέλου, αντί της παραγωγής και αποτίμησης μιας διάδοχης γραμματικής, εμφανίζει πολυπλοκότητα $C_{MergeNT} = O(1/2 \cdot N^2 \cdot M + N \cdot S)$, όπου $M$ ο αριθμός των κανόνων που θα αποβληθούν από τη γραμματική ($M = |\Omega_1| + |\Omega_2| + |\Omega_3|$), ο οποίος είναι συνήθως ένας μικρός αριθμός. Αν και υπάρχει μείωση μιας τάξης μεγέθους (από κυβική σε τετραγωνική διεργασία), εντούτοις, η πρόβλεψη του μήκος του μοντέλου είναι ακόμα μια τετραγωνική διαδικασία, η οποία μπορεί να αποτελέσει πρόβλημα εάν ο αριθμός των παραδειγμάτων εκπαίδευσης είναι μεγάλος. Σε τέτοιες περιπτώσεις απαιτούνται πρόσθετα ευριστικά να εφαρμοστούν, τα οποία πιθανώς να συνδυάζουν την ιδέα του «παρόμοιου περιβάλλοντος» με τη γενική κατανομή των διάφορων συμβόλων στη γραμματική.

### *5.4.11* Ο τελεστής “Create Optional NT”

Όπως έχει ήδη αναφερθεί, ο αλγόριθμος egGRIDS+ εισάγει έναν νέο τελεστή, τον “Create Optional NT”, καθώς και την σχετική κατάσταση λειτουργίας στη διαδικασία αναζήτησης. Η εισαγωγή αυτού του νέου τελεστή ήταν επιβεβλημένη, δεδομένου ότι οι υπάρχοντες τελεστές σε συνδυασμό με την αναζήτηση δέσμης δεν ήταν ικανοί να εξάγουν σωστά μερικούς τύπους γραμματικών, συγκλίνοντας συχνά σε γραμματικές λιγότερο γενικές από τη σωστή. Αυτή η ανεπάρκεια αποδίδεται κυρίως στο γεγονός ότι οι δύο παλαιότεροι τελεστές δε μπορούσαν να δημιουργήσουν τους κανόνες που περιέχουν ένα προαιρετικό σύμβολο, δηλαδή ένα σύμβολο που είτε υπάρχει είτε όχι, αλλά μόνο μια εμφάνισή του μπορεί να επεκταθεί από τον κανόνα, εάν το σύμβολο υπάρχει (σε αντίθεση με μια πλήρη αναδρομή σε αυτό το σύμβολο). Τέτοια «προαιρετικά» σύμβολα θα μπορούσαν να υπάρξουν στις εξαχθείσες γραμματικές μόνο ως υπολείμματα σε κανόνες που ξεκινούν με το σύμβολο εκκίνησης ή μέσω αναδρομής. Αφήνοντας τα σύμβολα αυτά στους κανόνες του υποσυνόλου συμβόλου έναρξης, δεν επιτρέπει στους κανόνες αυτούς να συγχωνευτούν (οδηγώντας έτσι σε πιο συμπαγείς γραμματικές), ενώ η εισαγωγή αναδρομής στην περιγραφή τέτοιων συμβόλων οδηγεί σε πιο γενικές γραμματικές, οι οποίες συχνά απορρίπτονται. Η εισαγωγή του τελεστή

*CreateOptionalNT* προσπαθεί να εξαλείψει αυτή την ανεπάρκεια με την εισαγωγή κανόνων στη γραμματική που καθιστούν τα σύμβολα προαιρετικά. Οι κανόνες αυτοί δεν θα ήταν εύκολο να έχουν εισαχθεί από τους άλλους τελεστές.

Ο τελεστής *CreateOptionalNT* επιδιώκει να επεκτείνει έναν κανόνα που δημιουργείται από τον τελεστή CreateNT, προσθέτοντας ένα επιπλέον μη τερματικό σύμβολο από τα ήδη υπάρχοντα. Αυτή η επέκταση δεν έχει επιπτώσεις στον αρχικό κανόνα, δεδομένου ότι ο CreateOptionalNT επεκτείνει έναν νέο κανόνα – που είναι ένα αντίγραφο του αρχικού κανόνα – με ένα μη τερματικό σύμβολο στο τέλος του σώματος του κανόνα, όπως φαίνεται και στον ακόλουθο πίνακα (Πίνακας 39).

Η επίδραση του νέου τελεστή στη γραμματική είναι ότι το επισυναπτόμενο σύμβολο γίνεται προαιρετικό όσον αφορά τον αρχικό κανόνα, όπως δημιουργείται από τον τελεστή CreateNT. Το γεγονός ότι ο CreateOptionalNT δεν αλλάζει τον αρχικό κανόνα κι επεκτείνει έναν νέο κανόνα με την ίδια επικεφαλίδα είναι ένα βήμα γενίκευσης.

Θεωρητικά, το αποτέλεσμα της επίδρασης του CreateOptionalNT θα μπορούσε να επιτευχθεί και από τους τελεστές CreateNT και MergeNT (Πίνακας 40).

| **Operator "Create Optional NT": Making optional symbol ADJ** | | |
|---|---|---|
| NP → AP1 NOUN<br>NP → AP1 ADJ NOUN<br>NP → AP1 ADJ ADJ NOUN<br>AP1→ ART ADJ<br><br>X → A1 AP1<br>Y → X ADJ NOUN | ⇒ | NP → AP1 NOUN<br>NP → AP1 ADJ NOUN<br>AP1→ ART ADJ<br>AP1→ ART ADJ ADJ<br>X → A1 AP1<br>Y → X NOUN |

**Πίνακας 39: Η επίδραση του τελεστή CreateOptionalNT.**

| **Operator "Create NT": Creating symbol AP2** | | |
|---|---|---|
| AP1→ ART ADJ | ⇒ | AP1→ ART ADJ<br>AP2→ ART ADJ ADJ |
| **Operator "Merge NT": Merging symbols AP1 & AP2** | | |
| AP1→ ART ADJ<br>AP2→ ART ADJ ADJ | ⇒ | AP1→ ART ADJ<br>AP1→ ART ADJ ADJ |

**Πίνακας 40: Αναλύοντας το αποτέλεσμα της επίδρασης του τελεστή CreateOptionalNT.**

Στην πράξη όμως αυτό δεν είναι δυνατό να συμβεί για τους εξής λόγους:

- Ο τελεστής CreateNT δε μπορεί να δημιουργήσει έναν κανόνα με τρία σύμβολα στο σώμα του κανόνα, καθώς λειτουργεί αποκλειστικά με διγράμματα.
- Η δημιουργία του *"AP1"* από τον τελεστή CreateNT, ο μόνος τελεστής που μπορεί να δημιουργήσει νέους κανόνες, θα εξαλείψει όλες τις εμφανίσεις του διγράμματος *"ART ADJ"* από τη γραμματική. Ως αποτέλεσμα, ένας κανόνας της μορφής *"AP2 → ART ADJ ADJ"* δε γίνεται να υπάρξει, έτσι ώστε ο τελεστής MergeNT να μπορεί να οδηγήσει στη γραμματική του ακόλουθου πίνακα (Πίνακας 41).

Η επίδραση αυτού του τελεστή δεν είναι μια απλή αλλαγή αναπαράστασης, καθώς η προσθήκη ενός νέου κανόνα που μοιράζεται την ίδια κεφαλή με έναν υπάρχοντα

κανόνα αυξάνει πάντα την κάλυψη της γραμματικής. Επιπλέον, η χρήση αυτού του τελεστή μπορεί να προκαλέσει την εξάλειψη διπλότυπων κανόνων από τη γραμματική.

| | | |
|---|---|---|
| S → … X Y … | | S → … X … |
| S → … Z Y … | | S → … Z … |
| S → … W Y … | | S → … W … |
| W → … X | ⇒ | W → … X |
| Z → … W | | Z → … W |
| X → X1 X2 | | X → X1 X2 |
| | | X → X1 X2 Y |

**Πίνακας 41: Αντικαθιστώντας όλες τις εμφανίσεις του $W_X Y$, όπου το $W_X$ μπορεί να επεκταθεί με το προαιρετικό σύμβολο $Y$.**

Γενικά, η επίδραση του τελεστή *CreateOptionalNT* σε μια γραμματική μπορεί να συνοψιστεί ως εξής (υποθέτοντας ότι ο κανόνας "$X \rightarrow X1\ X2$" πρόκειται να αυξηθεί με το μη τερματικό σύμβολο "$Y$"):

- Όλες οι εμφανίσεις της ακολουθίας "$X\ Y$" αντικαθιστώνται από το μη τερματικό σύμβολο "$Y$".
- Όλες οι εμφανίσεις της ακολουθίας $W_X Y$ αντικαθιστούνται από το $X$, εάν το σώμα του κανόνα του οποίου το $W_X$ είναι η κεφαλή τελειώνει με $X$, ή το τελευταίο σύμβολο του σώματος του κανόνα μπορεί να επεκταθεί στο τέλος με το σύμβολο $X$ (Ένα παράδειγμα φαίνεται στον πίνακα: Πίνακας 41).
- Ένας νέος κανόνας της μορφής "$X \rightarrow X1\ X2\ Y$" προστίθεται στη γραμματική.
- Κανόνες είναι δυνατόν να εξαλειφθούν από τη γραμματική, καθώς οι αντικαταστάσεις των διγραμμάτων μπορεί να οδηγήσουν σε διπλούς κανόνες.
- Η κάλυψη της γραμματικής πάντοτε αυξάνεται.
- Το GDL της γραμματικής τροποποιείται ως εξής:

a) Ένας νέος κανόνας της μορφής "X → X1 X2 Y" προστίθεται.
b) Η εξάλειψη του συμβόλου *"$Y$"* από τα σώματα κάποιων κανόνων ενδέχεται να προκαλέσει τη συγχώνευσή τους, ελαττώνοντας έτσι το συνολικό αριθμό εμφάνισης συμβόλων στη γραμματική.

- Το $DDL$ της γραμματικής τροποποιείται επίσης λόγω των ακολούθων αιτίων:

a) Ο αριθμός των αναθεωρημένων κανόνων που καθορίζουν το σύμβολο *"$X$"* αυξάνεται κατά ένα. Ως αποτέλεσμα, περισσότερα διφυία απαιτούνται –σε σχέση με πριν– για να κωδικοποιηθεί κάθε εμφάνιση του συμβόλου *"$X$"*, αυξάνοντας έτσι το $DDL$.
b) Η εξάλειψη κανόνων ενδέχεται να ελαττώσει τον αριθμό των περιπτώσεων που κάποιο μη τερματικό σύμβολο (συμπεριλαμβανομένου του *"$X$"*) εμφανίζεται ως κεφαλή κανόνα, ελαττώνοντας τον αριθμό των δυφίων που απαιτούνται για να κωδικοποιηθεί κάθε εμφάνιση των συμβόλων αυτών, ελαττώνοντας έτσι και το $DDL$.

### 5.4.12 Η πολυπλοκότητα της κατάστασης "Create Optional NT"

Σε αυτήν την παράγραφο υπολογίζουμε την πολυπλοκότητα του τελεστή CreateOptionalNT και την πολυπλοκότητα ενός πλήρους βήματος «δημιουργίας προαιρετικού συμβόλου» (δηλ. μιας ολόκληρης κατάστασης λειτουργίας του αλγορίθμου egGRIDS+ όπου εφαρμόζεται μόνο ο τελεστής CreateOptionalNT). Αυτό το βήμα χαρακτηρίζεται από την συνεχή εφαρμογή του τελεστή σε όλα τα πιθανά διγράμματα,

των οποίων το πρώτο σύμβολο είναι ένα μη τερματικό σύμβολο που έχει δημιουργηθεί από τον τελεστή CreateNT, και του υπολογισμού του μήκους του μοντέλου όλων των διάδοχων γραμματικών. Η διαδικασία πραγματοποίησης διαδικασία ενός πλήρους βήματος «δημιουργίας προαιρετικού συμβόλου» μπορεί να πραγματοποιηθεί με τον αλγόριθμο που παρουσιάζεται στην Εικόνα 28.

Σαν πρώτο βήμα, πρέπει να προσδιοριστούν όλα τα πιθανά διγράμματα των οποίων το πρώτο σύμβολο έχει δημιουργηθεί από τον τελεστή CreateNT, το οποίο απαιτεί μια επανάληψη σε όλα τα σύμβολα σε όλα τα σώματα των κανόνων. Η πολυπλοκότητα γι' αυτό είναι $O(N \cdot S)$, υποθέτοντας ότι η διαδικασία αποθήκευσης των συμβόλων έχει σταθερή πολυπλοκότητα. Τα παραγόμενα δίγραμμα είναι το πολύ $K = N \cdot S$.

Σαν δεύτερο βήμα, ο τελεστής CreateOptionalNT πρέπει να εφαρμοστεί σε όλα τα εντοπισμένα διγράμματα. Για κάθε δίγραμμα πρέπει να προσδιοριστεί το σύνολο $S_H$,. Αυτό είναι μια επαναληπτική διαδικασία που μπορεί να γίνει το πολύ σε $N - 1$ βήματα (εάν όλες οι κεφαλές των κανόνων, εκτός από το σύμβολο έναρξης γραμματικής έχουν επισυναφθεί στο $S_H$). Καθώς όλοι οι κανόνες πρέπει να εξεταστούν κατά τη διάρκεια κάθε επανάληψης, η συνολική πολυπλοκότητα του προσδιορισμού του $S_H$, είναι $(N - 1) \cdot N$. Ο συνολικός αριθμός των στοιχείων του $S_H$ είναι το πολύ $L = N - 1$.

```
for each Rule in Grammar {
  for i=0, i < Rule body symbol number, i=i+1 {
    if symbol[i] created by the CreateNT operator {
    Store_bigram (symbol[i], symbol[i+1])
    }
  }
}
for each stored Bigram {
  S_H= Bigram(0)
  Y = Bigram(1)
  do {
    for each Rule in Grammar {
      if last symbol in rule body is in S_H {
        Append rule head to S_H
      }
    }
  } while S_H modified
  for each symbol X in S_H {
    for each Rule in Grammar {
      for each Symbol in Rule {
        Replace bigram (X,Y) with X
      }
    }
  }
    Measure Grammar Model Length O(N · (1 + S))
}
```

**Εικόνα 28: Ψευδοκώδικας ενός βήματος (κατάστασης λειτουργίας του egGRIDS+) του τελεστή CreateOptionalNT.**

Τέλος, το τρίτο βήμα είναι η εφαρμογή του τελεστή CreateOptionalNT και η αποτίμηση του μήκους μοντέλου της διάδοχης γραμματικής. Η εφαρμογή του τελεστή απαιτεί μια

επανάληψη σε όλα τα σύμβολα του $S_H$: για κάθε σύμβολο στο $S_H$, τα σώματα από όλους τους κανόνες στη γραμματική πρέπει να εξεταστούν προκειμένου να εξαλείψουν το σύμβολο "$Y$". Αυτή η διαδικασία έχει πολυπλοκότητα $L \cdot N \cdot S$, υποθέτοντας ότι η διαδικασία αντικατάστασης ενός διγράμματος έχει σταθερή πολυπλοκότητα. Εξετάζοντας όλα τα βήματα, η συνολική πολυπλοκότητα για τον τελεστή CreateOptionalNT $\left(C_{CreateOptionalNT}\right)$ είναι κυβική:

$$C_{CreateOptionalNT} = O\left(N \cdot S + K \cdot N\left(N + L \cdot S + S\right)\right) = O\left(N \cdot S + N^2 \cdot S \cdot \left(N + \left(N-1\right) \cdot S + S\right)\right) = O\left(N^3 \cdot S + N^2 \cdot \left(N-1\right) \cdot S^2 + N^2 \cdot S^2 + N \cdot S\right) \approx O\left(2 \cdot N^3 + N^2\right)$$

### 5.4.13 Η επίδραση του τελεστή "Create Optional NT" στο μήκος περιγραφής γραμματικής

Όπως και στην περίπτωση του τελεστή MergeNT, το αρχικό $GDL_{In}$ της γραμματικής $G$ (προτού να εφαρμοστεί ο τελεστής) μπορεί να υπολογιστεί όπως:

$$GDL_{In} = (A_{NT} + A_R - A_S + 2) \cdot \log(A_{UNT} + 1) + A_T \cdot \log(A_{UT}).$$

Ομοίως, το $GDL_{Fin}$ μετά την εφαρμογή του τελεστή μπορεί να υπολογιστεί ως εξής:

$$GDL_{Fin} = (A_{NT} + A_R - A_S + 7) \cdot \log(A_{UNT} + 1) + A_T \cdot \log(A_{UT}) - \mathrm{E}_1 - \mathrm{E}_2$$

Ο όρος $E_1$ στην παραπάνω εξίσωση αντιπροσωπεύει την αναμενόμενη μείωση στο $GDL$ που μπορεί να εμφανιστεί εάν η εφαρμογή του τελεστή οδηγήσει σε διπλότυπους κανόνες, οι οποίοι πρέπει να διαγραφούν. Από την ενότητα 0 έχουμε:

$$\mathrm{E}_1 = \sum_{j \in \Omega_1} \left(L_j + 1\right) \cdot \log(A_{UNT} + 1) + \sum_{j \in \Omega_3} \left(L_j + 2\right) \cdot \log(A_{UNT} + 1)$$

Όπου:

- $\Omega_1$: το σύνολο κανόνων από το υποσυνόλου συμβόλων έναρξης, που αποβάλλεται από την $G$.
- $\Omega_3$: το σύνολο κανόνων από το υποσύνολο μη-τερματικών συμβόλων, που αποβάλλεται από την $G$.
- $L_j$: ο αριθμός μη-τερματικών συμβόλων στο σώμα του κανόνα $j$.

Ο όρος $E_2$ αναπαριστά την αναμενόμενη μείωση στο $GDL$ που προκύπτει από την αφαίρεση του συμβόλου "$Y$":

$$\mathrm{E}_2 = N_Y \cdot \log(A_{UNT} + 1)$$

Όπου $N_Y$ είναι ο αριθμός των διαγραφών του συμβόλου "$Y$" από τη γραμματική. Σαν αποτέλεσμα, η αλλαγή $\Delta_{GDL}$ του $GDL$ εξαιτίας αυτού του τελεστή είναι:

$$\Delta_{GDL} = 5 \cdot \log(A_{UNT} + 1) - N_Y \cdot \log(A_{UNT} + 1) - \left( \sum_{j \in \Omega_1} (L_j + 1) \cdot \log(A_{UNT}) + \sum_{j \in \Omega_3} (L_j + 2) \cdot \log(A_{UNT}) \right). \quad (5.7a)$$

Ο πρώτος όρος στην εξίσωση 5.7a είναι μια μικρή θετική σταθερά, ενώ οι άλλοι δύο παράγοντες είναι πάντα αρνητικοί. Εξαιτίας του γεγονότος ότι η μόνη θετική ποσότητα είναι μια μικρή θετική σταθερά, ακόμη και η διαγραφή πέντε εμφανίσεων του συμβόλου "$Y$" αρκεί για να μειωθεί το $GDL$ της διάδοχης γραμματικής. Η μείωση στο $GDL$ είναι μεγαλύτερη, όσο αυξάνεται ο αριθμός των διπλότυπων κανόνων που διαγράφονται από τη γραμματική, μαζί με τον αριθμό των εμφανίσεων του συμβόλου "$Y$" που διαγράφονται.

### 5.4.14 Η επίδραση του τελεστή “Create Optional NT” στο μήκος περιγραφής παραγωγών

Όπως και ο τελεστής MergeNT, έτσι και ο τελεστής CreateOptionalNT τροποποιεί το μήκος περιγραφής παραγωγών $DDL$ με δύο τρόπους: συγχωνεύοντας σύνολα από κανόνες και εξαλείφοντας κανόνες από τη γραμματική. Όσον αφορά την επίδραση στο $DDL$ λόγω της προσθήκης ενός νέου κανόνα, που μοιράζεται την ίδια κεφαλή με έναν υπάρχοντα κανόνα, είναι ίδια με τη συγχώνευση δύο συνόλων από κανόνες με το δεύτερο σύνολο που περιέχει έναν μονό κανόνα. Συνεπώς, εάν θέσουμε $\beta = 0$ και $F_Y = 1$ στην εξίσωση (5.6d), έχουμε:

$$\Delta C = \alpha \cdot \log\left(1 + \frac{1}{F_X}\right) \quad (5.7b)$$

Τέλος, σχετικά με την εξάλειψη διπλότυπων κανόνων από τη γραμματική, η επίδραση του τελεστή CreateOptionalNT είναι ίδια με εκείνη του MergeNT, όπως περιγράφεται από την εξίσωση (5.6e) (με τα σύμβολα "$X$" και "$Z$" του τελεστή MergeNT να αντικαθίστανται πλέον με τα σύμβολα "$X$" και "$Y$" στον τελεστή CreateOptionalNT).

### 5.4.15 Συνολική επίδραση του "Create Optional NT" στο μήκος περιγραφής μοντέλου γραμματικής

Η συνολική επίδραση του CreateOptionalNT στο μήκος περιγραφής μοντέλου μιας γραμματικής $\Delta ML$ μπορεί να υπολογιστεί με το συνδυασμό των εξισώσεων (5.7a), (5.7b) και (5.6e).

$$\begin{aligned}\Delta ML = {} & 5\cdot\log\left(A_{UNT}+1\right) - N_Y\cdot\log\left(A_{UNT}+1\right) - \\ & \left(\sum_{j\in\Omega_1}\left(L_j+1\right)\cdot\log\left(A_{UNT}\right)+\sum_{j\in\Omega_3}\left(L_j+2\right)\cdot\log\left(A_{UNT}\right)\right)+ \\ & \left(\sum_{j\in G_{In}} N_j^{CNT-X}\cdot F_j^S\right)\cdot\log\left(1+\frac{1}{F_{CNT-X}}\right)+\sum_{j\in G_{Fin}}\left(\sum_{k\in\Theta,j} M_k\right)\cdot F_j^S\end{aligned} \tag{5.7}$$

Οι πρώτοι τρεις όροι της παραπάνω εξίσωσης αντιστοιχούν στην αλλαγή στο $GDL$. Ο τέταρτος και πέμπτος όρος της εξίσωσης αντιστοιχούν στην αλλαγή στο $DDL$: Ο τέταρτος όρος αντιπροσωπεύει τη συγχώνευση συνόλων από κανόνες και είναι πάντα θετικός. Ο τελευταίος όρος αντιπροσωπεύει την εξάλειψη διπλότυπων κανόνων και είναι πάντα αρνητικός ή μηδέν, ανάλογα με την ύπαρξη των τέτοιων κανόνων.

## 5.5 Πειραματική αξιολόγηση και αποτελέσματα

Σε αυτή την ενότητα αξιολογούμε πειραματικά τον αλγόριθμο egGRIDS+, εστιάζοντας στην απόδοσή του σε συγκεκριμένα πεδία εφαρμογής, καθώς και στην *κλιμακωσιμότητα* του *(scalability)*. Επιπλέον, εξετάζουμε το ρόλο ορισμένων παραμέτρων του αλγορίθμου, όπως το μέγεθος της δέσμης κατά την διαδικασία αναζήτησης *(beam search)*. Για την πειραματική αξιολόγηση του αλγορίθμου θα χρησιμοποιηθούν τόσο προτάσεις που παράγονται από τεχνητές γραμματικές, όσο και προτάσεις από ένα μεγάλο σώμα κειμένων. Οι τεχνητές γραμματικές μας επιτρέπουν να αξιολογήσουμε συγκεκριμένα χαρακτηριστικά του egGRIDS+, δεδομένου ότι μπορούμε να ελέγξουμε τις διάφορες πτυχές μιας γραμματικής, ενώ ταυτόχρονα μας παρέχουν την δυνατότητα να μετρήσουμε την ακρίβεια των γραμματικών που έχουν εξαχθεί. Από την άλλη πλευρά, το μεγάλο σώμα κειμένων, μας επιτρέπει να εξετάσουμε την κλιμακωσιμότητα του egGRIDS+ σε πιο σύνθετες (και περισσότερο πρακτικές) γραμματικές περιοχές.

### 5.5.1 Πειράματα σε Τεχνητές Γραμματικές

#### Μετρικές αξιολόγησης σε τεχνητές γραμματικές

Η αξιολόγηση στην περιοχή της *επαγωγικής εξαγωγής γραμματικών (grammatical inference)* παρουσιάζει μερικές ιδιαιτερότητες, καθώς δημοφιλείς μετρικές που χρησιμοποιούνται συχνά στις τεχνικές επιβλεπόμενης μάθησης, όπως η *ανάκληση (recall)* και η *ακρίβεια (precision)*, δεν είναι δυνατό να χρησιμοποιηθούν άμεσα. Εναλλακτικά, για να αξιολογηθεί μια παραγόμενη γραμματική πρέπει να συγκριθεί με τη «σωστή» γραμματική, ώστε να προσδιοριστεί η ομοιότητά τους. Ωστόσο, ακόμα κι αν η «σωστή» γραμματική είναι γνωστή, υπόθεση που συχνά δεν είναι ρεαλιστική, ο προσδιορισμός για το εάν δύο *ανεξάρτητες από συμφραζόμενα γραμματικές είναι ισοδύναμες* δεν είναι μια καθόλου εύκολη εργασία: Δοθέντων δύο ανεξάρτητων από τα συμφραζόμενα γραμματικών $G_1$ και $G_2$, δεν υπάρχει αλγόριθμος που να μπορεί να προσδιορίσει εάν η $G_1$ είναι πιο γενικευμένη από την $G_2$ (π.χ. $L(G_1)\supseteq L(G_2)$ ή $L(G_1)\cap L(G_2)=\emptyset$, όπου $L(G)$ είναι η γλώσσα της γραμματικής $G$) [110], [124].

Συνεπώς, κατά τη διάρκεια της αξιολόγησης εστιάζουμε κυρίως στη μέτρηση τριών πτυχών της εξαχθείσας γραμματικής [13]:

- Λάθη παραλείψεων (errors of omission): (αποτυχίες στην ανάλυση προτάσεων που έχουν παραχθεί από τη «σωστή» γραμματική), που δείχνουν κατά πόσο έχει εξαχθεί μια υπερβολικά συγκεκριμένη γραμματική.
- Λάθη διάπραξης (errors of commission): (αποτυχίες της «σωστής» γραμματικής να αναλύσει προτάσεις που παράγονται από την εξαχθείσα γραμματική), που δείχνουν κατά πόσο έχει εξαχθεί μια υπερβολικά γενική γραμματική.
- Δυνατότητα της εξαχθείσας γραμματικής να αναλύσει σωστά προτάσεις που είναι μεγαλύτερες (σε αριθμό λέξεων) από εκείνες που χρησιμοποιούνται κατά τη διάρκεια της εκπαίδευσης, κάτι που δείχνει την πρόσθετη εκφραστικότητα της εξαχθείσας γραμματικής.

Σε πειράματα με τεχνητές γραμματικές, όπου η «σωστή» γραμματική είναι γνωστή, για τον προσδιορισμό των παραπάνω μετρικών χρησιμοποιούμε τόσο την αρχική (ή «σωστή») γραμματική $G_O$, όσο και την εξαχθείσα γραμματική $G_L$, για να παράγουμε έναν μεγάλο αριθμό προτάσεων. Τα λάθη παραλείψεων μπορούν να υπολογιστούν ως το κλάσμα του αριθμού των προτάσεων που έχουν παραχθεί από την $G_O$ και οι οποίες δεν έχουν παραχθεί από την $G_L$ , προς το συνολικό αριθμό των προτάσεων που παράχθηκαν από την $G_O$. Τα λάθη διάπραξης μπορούν να υπολογιστούν ως το κλάσμα του αριθμού προτάσεων που έχουν παραχθεί από την $G_L$ και οι οποίες δεν έχουν παραχθεί από την $G_O$, προς το συνολικό αριθμό των προτάσεων που έχουν παραχθεί από την $G_L$. Τα λάθη παραλείψεων και τα λάθη διάπραξης μετρούν την επικάλυψη των δύο γραμματικών. Στην ιδανική περίπτωση, και οι δύο μετρικές πρέπει να είναι μηδέν, σηματοδοτώντας ότι όλες οι προτάσεις που παράγονται από την μια γραμματική μπορούν να αναλυθούν από την άλλη. Εάν αυτό συμβαίνει, και η αποτίμηση περιλαμβάνει έναν αρκετά μεγάλος αριθμός από προτάσεις, μπορούμε να καταλήξουμε στο συμπέρασμα ότι οι δύο γραμματικές επικαλύπτονται σημαντικά.

Προκειμένου να υπολογιστεί η τρίτη μετρική, προτάσεις πρέπει να παραχθούν από την $G_O$, οι οποίες να έχουν μεγαλύτερο μήκος από αυτές που χρησιμοποιήθηκαν κατά την φάση της εκπαίδευσης. Η μετρική μπορεί έτσι να υπολογιστεί σαν το κλάσμα του αριθμού προτάσεων που αναλύθηκαν επιτυχώς από την $G_L$ , προς το συνολικό αριθμό παραγμένων προτάσεων από την $G_O$.

### Περιγραφή πειραμάτων

Το πρώτο σύνολο πειραμάτων αξιολογεί τον egGRIDS+ σε παραδείγματα που παράγονται από απλές *αναδρομικές γραμματικές (recursive grammars)* οι οποίες χρησιμοποιούν ένα μικρό σύνολο τερματικών συμβόλων, δεδομένου ότι ενδιαφερόμαστε κυρίως για την εξέταση της δυνατότητας του egGRIDS+ να συμπεράνει αναδρομικές γραμματικές. Οι δύο γραμματικές που χρησιμοποιήθηκαν στα πειράματά μας παρουσιάζονται στον πίνακα: Πίνακας 42.

Η πρώτη γραμματική (α) περιλαμβάνει δηλωτικές προτάσεις με απροσδιόριστο αριθμό διαδοχικών επιθέτων, μαζί με μεταβατικά και αμετάβατα ρήματα, αλλά χωρίς δευτερεύουσες προτάσεις, εμπρόθετες φράσεις, επιρρήματα ή κλίσεις. Η δεύτερη γραμματική (β) περιέχει δηλωτικές προτάσεις με απροσδιόριστο αριθμό δευτερευουσών προτάσεων, αλλά δίχως επίθετα, επιρρήματα, εμπρόθετες φράσεις ή κλίσεις. Και οι δύο γραμματικές συνδέονται με ένα μικρό λεξικό τερματικών συμβόλων (λέξεων).

Αν και αυτές οι γραμματικές είναι απλοϊκές έναντι γραμματικών από την περιοχή της επεξεργασίας φυσικής γλώσσας, και οι δύο περιλαμβάνουν αναδρομή και περιγράφουν μέσω αυτής μια άπειρη γλώσσα. Η πρώτη γραμματική χρησιμοποιεί έναν απλό τύπο αναδρομής, όπου τα επίθετα επαναλαμβάνονται μπροστά από ένα ουσιαστικό. Μόλις

ομαδοποιηθούν όλα τα επίθετα σε ένα τερματικό σύμβολο από τον egGRIDS+, οι ακολουθίες συμβόλων που ανήκουν σε αυτήν την κατηγορία παρέχουν συνήθως αρκετές πληροφορίες για να ανιχνεύσουν την αναδρομή. Από την άλλη πλευρά, η δεύτερη γραμματική περιλαμβάνει αναδρομή σε μια δευτερεύουσα πρόταση: η δευτερεύουσα πρόταση πρέπει πρώτα να προσδιοριστεί σωστά, προτού παρουσιαστούν αρκετά στοιχεία για να ανιχνευθεί η αναδρομή. Το κύριο σημείο ενδιαφέροντος και στις δύο γραμματικές είναι να εξεταστεί η δυνατότητα του egGRIDS+ να γενικεύσει σωστά, με το να συμπεράνει γραμματικές που περιέχουν αναδρομή.

| (a) | | (b) | |
|---|---|---|---|
| S | → NP VP | S | → NP VP |
| VP | → VERBI | VP | → VERB NP |
| VP | → VERBT NP | NP | → ART NOUN |
| NP | → the NOUN | NP | → ART NOUN RC |
| NP | → the AP NOUN | RC | → REL VP |
| AP | → ADJ | VERB | → saw |
| AP | → ADJ AP | VERB | → heard |
| VERBI | → ate | NOUN | → cat |
| VERBI | → slept | NOUN | → dog |
| VERBT | → saw | NOUN | → mouse |
| VERBT | → heard | ART | → a |
| NOUN | → cat | ART | → the |
| NOUN | → dog | REL | → that |
| ADJ | → big | | |
| ADJ | → old | | |

**Πίνακας 42: Δύο τεχνητές γραμματικές: η γραμματική (a) περιλαμβάνει τυχαίες σειρές επιθέτων, και η (b) υποστηρίζει τυχαίες δευτερεύουσες προτάσεις (relative clauses) [13].**

Στα πειράματα που πραγματοποιήθηκαν, εξήχθη ένας μεγάλος αριθμός προτάσεων από τις γραμματικές (α) και (β), χρησιμοποιώντας μια από *επάνω προς τα κάτω (top-down)* διαδικασία. Για κάθε γραμματική, παρήχθη ένας μεγάλος αριθμός παραδειγματικών προτάσεων (πάνω από 10.000), χρησιμοποιώντας μια ομοιόμορφη κατανομή (normal distribution) για την επιλογή τυχαίων κανόνων όταν επεκτείνονταν διφορούμενα μη-τερματικά σύμβολα. Οι εξαχθείσες προτάσεις ανακατεύθηκαν τυχαία. Για κάθε γραμματική ορίσαμε ένα αυθαίρετο μέγιστο μήκος $L_{max}$ τόσο για τα παραδείγματα εκπαίδευσης, όσο και για τα παραδείγματα αξιολόγησης. Το σύνολο που προέκυψε, χρησιμοποιήθηκε για την αξιολόγηση σύμφωνα με τις δύο πρώτες μετρικές, δηλαδή τα λάθη παραλείψεων και τα λάθη διάπραξης. Δεδομένου ότι θελήσαμε να μελετήσουμε κατά πόσο ο egGRIDS+ μπορεί να συμπεράνει αναδρομικές γραμματικές, θελήσαμε επίσης να αξιολογήσουμε τις εξαχθείσες γραμματικές στις παραδειγματικές προτάσεις οι οποίες ήταν ακόμα μακρύτερες από εκείνες που χρησιμοποιήθηκαν κατά την εκπαίδευση. Για το σκοπό αυτό, ένα δεύτερο σύνολο αξιολόγησης δημιουργήθηκε, που περιείχε παραδειγματικές προτάσεις με μήκη μεγαλύτερα από $L_{max}$, αλλά μικρότερα από ένα δεύτερο αυθαίρετα μέγιστο μήκος. Αυτό το δεύτερο σύνολο είναι εκείνο που χρησιμοποιήθηκε προκειμένου να υπολογιστεί η τρίτη μετρική, δηλαδή η δυνατότητα να αναλυθούν προτάσεις που είναι μεγαλύτερες από αυτές που χρησιμοποιήθηκαν κατά την εκπαίδευση. Όλα τα σύνολα δοκιμής εμπλουτίστηκαν με τυχαίες παραδειγματικές προτάσεις από τις εξαχθείσες γραμματικές. Ιδιαίτερη προσοχή

λήφθηκε προκειμένου να εξασφαλιστεί ότι η ίδια πρόταση δεν εμφανίστηκε ταυτόχρονα και στα δύο σύνολα (εκπαίδευσης και αξιολόγησης).

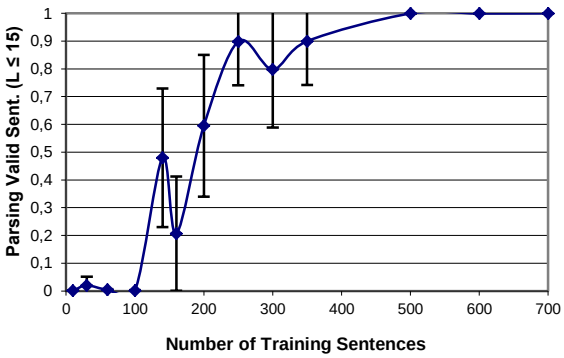
Δεδομένου ότι ενδιαφερόμαστε επίσης για το ρυθμό (rate) εκμάθησης, ο αλγόριθμος αποτιμήθηκε με δεδομένα εκπαίδευσης διαφόρων μεγεθών. Για την αποκόμιση μιας ρεαλιστικής και αμερόληπτης εκτίμησης των επιδόσεων του egGRIDS+, κάθε πείραμα επαναλήφθηκε δέκα φορές. Για κάθε προκαθορισμένο μέγεθος εκπαίδευσης, δημιουργήθηκαν δέκα σύνολα εκπαίδευσης του επιθυμητού μεγέθους, καθώς και ένα σύνολο αξιολόγησης, του ίδιου μεγέθους συν 1000 επιπλέον προτάσεις. Οι μισές από τις προτάσεις του συνόλου αξιολόγησης ήταν μακρύτερες από $L_{max}$. Καθένα από τα δέκα σύνολα εκπαίδευσης χρησιμοποιήθηκε για την εκπαίδευση του αλγορίθμου egGRIDS+, και η απόδοση της εξαχθείσας γραμματικής αξιολογήθηκε στο κοινό σύνολο αξιολόγησης. Ο μέσος όρος από τα δέκα πειράματα αναφέρεται εδώ ως τελική αξιολόγηση, συνοδευόμενη από τη *τυπική απόκλιση (standard deviation)*. Πρέπει να σημειωθεί ότι μια παραδειγματική πρόταση θεωρείται ότι αναλύθηκε επιτυχώς, όταν τουλάχιστον μια ανάλυση έχει βρεθεί, λαμβάνοντας υπόψη μια γραμματική. Καμία προσοχή δεν έχει δοθεί στην ταύτιση των *δέντρων ανάλυσης (parse trees)*.

### 5.5.2 Πείραμα 1: Αξιολόγηση σε μικρές γραμματικές

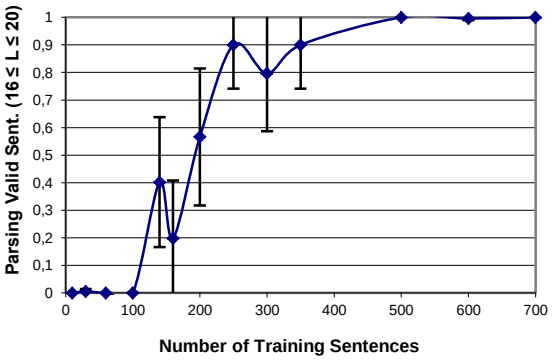
Σαν πρώτο πείραμα, έχουμε αξιολογήσει τη δυνατότητα του egGRIDS+ να συμπεράνει αναδρομικές γραμματικές με τη χρησιμοποίηση των δύο γραμματικών που παρουσιάστηκαν στον Πίνακας 42. Όσον αφορά τη γραμματική (α), παρήχθη ένα σύνολο 22.826 προτάσεων. Αυτό το σύνολο χωρίστηκε σε δύο υποσύνολα. Το πρώτο υποσύνολο (A) περιείχε τις προτάσεις με μήκος μέχρι 15 λέξεις (21.183 προτάσεις), ενώ το δεύτερο υποσύνολο (B) περιείχε τις προτάσεις με μήκος μεταξύ 16 και 20 λέξεων (1.643 προτάσεις). Η *αναζήτηση δέσμης (beam search)* χρησιμοποιήθηκε με ένα μέγεθος δέσμης 3. Το πείραμα πραγματοποιήθηκε χρησιμοποιώντας διάφορα καθορισμένα μεγέθη εκπαίδευσης, από 10 έως 700 προτάσεις. Για κάθε καθορισμένο μέγεθος εκπαίδευσης, δημιουργήθηκαν δέκα σύνολα εκπαίδευσης. Κάθε σύνολο εκπαίδευσης περιείχε έναν σταθερό αριθμό μοναδικών προτάσεων που επιλέχτηκαν τυχαία από το υποσύνολο (A). Από αυτά τα σύνολα εκπαίδευσης, δέκα γραμματικές παρήχθησαν και αξιολογήθηκαν στο κοινό σύνολο αξιολόγησης, το οποίο προέρχεται από το υποσύνολο (B) και περιέχει κατά προσέγγιση 1000 προτάσεις περισσότερες από κάθε σύνολο εκπαίδευσης. Στις εικόνες Εικόνα 29, Εικόνα 30, Εικόνα 31, παρουσιάζονται οι μέσοι όροι της απόδοσης, μαζί με τις μπάρες λάθους, που αναπαριστούν την τυπική απόκλιση.

Η Εικόνα 29 παρουσιάζει την πιθανότητα της εξαχθείσας γραμματικής να αναλύσει μια έγκυρη πρόταση που έχει μήκος μέχρι το μέγιστο μήκος των παραδειγμάτων που χρησιμοποιούνται κατά την εκπαίδευση. Η Εικόνα 30 παρουσιάζει την πιθανότητα της ανάλυσης μιας έγκυρης πρότασης που έχει μήκος μεγαλύτερο από οποιαδήποτε από τα παραδείγματα που βλέπει ο egGRIDS+ κατά τη διάρκεια της εκπαίδευσης. Τέλος, η Εικόνα 31 παρουσιάζει την πιθανότητα της παραγωγής μιας έγκυρης πρότασης. Από τις δύο πρώτες καμπύλες μπορούμε να καταλήξουμε στο συμπέρασμα ότι ο egGRIDS+ φθάνει σε μια αποδεκτή απόδοση (δηλαδή $\geq 0.9$) χωρίς να απαιτεί περισσότερα από 350 παραδείγματα εκπαίδευσης. Μετά τα 350 παραδείγματα, οι εξαχθείσες γραμματικές είναι αρκετά γενικές προκειμένου να αναγνωρίσουν ένα μεγάλο ποσοστό των προτάσεων αξιολόγησης, ακόμα κι αν οι προτάσεις είναι μεγαλύτερες από εκείνες που χρησιμοποίησε ο egGRIDS+ κατά την εκπαίδευση. Οι καμπύλες των σχημάτων Εικόνα 29 και Εικόνα 30 αντιπροσωπεύουν τη δυνατότητα του egGRIDS+ να γενικεύσει, ενώ επιπρόσθετα, η καμπύλη στην Εικόνα 30 παρέχει μια ένδειξη της δυνατότητας του egGRIDS+ να ανιχνεύσει αναδρομή. Η ομοιότητα των δύο καμπυλών δείχνει ότι ο

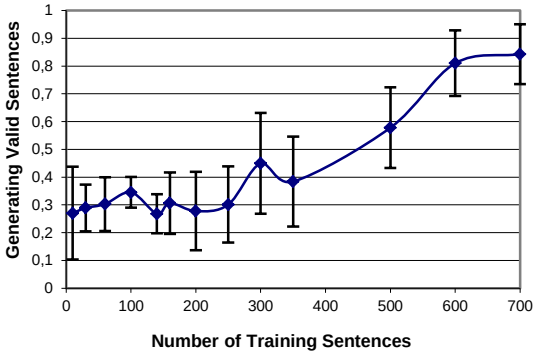
egGRIDS+ πέτυχε την ανίχνευση αναδρομής από τα επίθετα στα δεδομένα εκπαίδευσης.




**Εικόνα 29: Πιθανότητα ανάλυσης μιας έγκυρης πρότασης μήκους μέχρι 15 λέξεις. (1-errors of omission)**




**Εικόνα 30: Πιθανότητα ανάλυσης μιας έγκυρης πρότασης μήκους μεταξύ 16 και 20 λέξεων. (1-errors of omission)**




**Εικόνα 31: Πιθανότητα παραγωγής μιας έγκαιρης πρότασης. (1-errors of commission)**

Μια άλλη ενδιαφέρουσα παρατήρηση είναι οι σχετικά μεγάλες μπάρες λάθους, οι οποίες οφείλονται στην κβαντοποιημένη συμπεριφορά του αλγορίθμου. Οι εξαχθείσες γραμματικές είτε γενίκευσαν αρκετά ώστε να αναγνωρίσουν το σύνολο αξιολόγησης (συμπεριλαμβανομένων των κανόνων αναδρομής) είτε ήταν υπερβολικά συγκεκριμένες και δεν γενίκευσαν καθόλου στο σύνολο αξιολόγησης. Για παράδειγμα, για μέγεθος εκπαίδευσης 250 προτάσεων, ο egGRIDS+ συμπέρανε εννέα γραμματικές που ήταν σε θέση να αναγνωρίσουν όλες τις προτάσεις στο σύνολο αξιολόγησης και μια γραμματική που απέτυχε εντελώς στο ίδιο σύνολο.

Η καμπύλη στην Εικόνα 31 δείχνει σαφώς ότι ο egGRIDS+ υπεργενικεύει με μικρά σύνολα εκπαίδευσης. Ακόμη και με μέγεθος δεδομένων εκπαίδευσης τις 700 προτάσεις, οι εξαχθείσες γραμματικές έχουν μια πιθανότητα 0,15 να παράγουν μη γραμματικές προτάσεις. Αυτό οφείλεται στον τρόπο με τον οποίο ο egGRIDS+ επιλέγει να ταξινομήσει τα μη τερματικά σύμβολα. Για παράδειγμα, έστω ότι ο egGRIDS+ αποφασίζει να ταξινομήσει ένα ουσιαστικό κι ένα επίθετο κάτω από το ίδιο μη τερματικό σύμβολο. Αυτή η λανθασμένη κατηγοριοποίηση ίσως δε δημιουργήσει ποτέ προβλήματα στην αναγνώριση των προτάσεων αξιολόγησης, αλλά θα οδηγήσει στην παραγωγή πολλών μη-γραμματικών προτάσεων. Ο αριθμός των διαθέσιμων παραδειγμάτων δεν είναι επαρκής για την πραγματοποίηση πειραμάτων

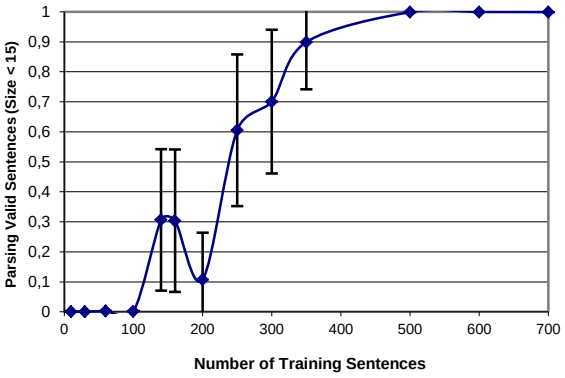
διασταυρωμένης επικύρωσης 10 βημάτων, χρησιμοποιώντας σύνολα εκπαίδευσης με μέγεθος μεγαλύτερο από 700. Εντούτοις, ένας και μόνο γύρος εκπαίδευσης/δοκιμής, δείχνει ότι η πιθανότητα παραγωγής μιας έγκυρης πρότασης φτάνει το 1,0, και σταθεροποιείται για μεγαλύτερα σύνολα εκπαίδευσης (πάνω από 700).

Όσον αφορά τη δεύτερη γραμματική (β), παρήχθη ένα σύνολο 11.500 προτάσεων, το οποίο χωρίστηκε και πάλι σε δύο υποσύνολα. Το πρώτο, περιείχε τις προτάσεις με μήκος μέχρι 15 λέξεις (8.387 προτάσεις), ενώ το δεύτερο περιείχε τις προτάσεις με μήκος μεταξύ 16 και 20 λέξεων (3.113 προτάσεις). Το μέγεθος της δέσμης ήταν πάλι 3. Το πείραμα πραγματοποιήθηκε με ποικίλα μεγέθη συνόλων εκπαίδευσης (από 10 έως 50 παραδείγματα) και τα αποτελέσματα έδειξαν ότι ακόμη και για έναν μικρό αριθμό παραδειγμάτων (20 και άνω) ο egGRIDS+ δεν κάνει κανένα λάθος. Με άλλα λόγια, και οι τρεις μετρικές αξιολόγησης παίρνουν την τιμή 1,0.

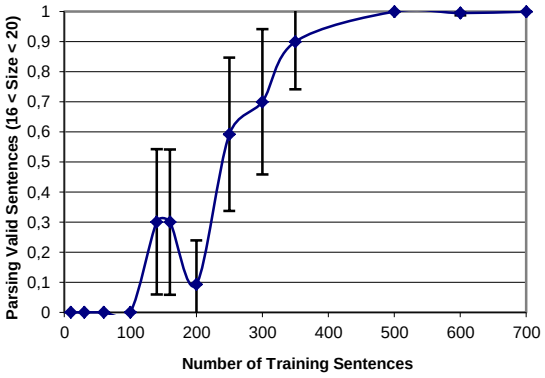
### 5.5.3 Πείραμα 2: Διαφοροποιώντας το μέγεθος της δέσμης

Στο δεύτερο πείραμα, ερευνήσαμε το ρόλο του μεγέθους της δέσμης κατά τη διαδικασία αναζήτησης. Η δέσμη στον egGRIDS+ χρησιμοποιείται προκειμένου να αποθηκευτεί σε οποιοδήποτε σημείο της διαδικασίας εκμάθησης οι $B$ πιο προεξέχουσες γραμματικές. Αυτές είναι οι γραμματικές που θα γενικευτούν περαιτέρω από τους τρεις τελεστές. Γενικά, αυτό που αναμένεται είναι ότι η απόδοση θα αυξηθεί καθώς αυξάνεται και το μέγεθος της δέσμης, και πράγματι αυτό παρατηρείται και στα πειράματα. Πραγματοποιήσαμε δυο διαφορετικές δοκιμές βασισμένες στη γραμματική (α), καθώς η γραμματική (β) φαίνεται να μαθαίνεται εύκολα, ακόμα και για $B = 3$. Στην πρώτη δοκιμή σαν μέγεθος δέσμης τέθηκε η τιμή $B = 1$, δηλαδή δεν χρησιμοποιήθηκε αναζήτηση δέσμης, ενώ στη δεύτερη δοκιμή το μέγεθος ακτίνας τέθηκε σαν $B = 10$. Τα αποτελέσματα με $B = 1$ ήταν πολύ κοντά στα αντίστοιχα με $B = 3$, κι επομένως δεν επαναλαμβάνονται εδώ.

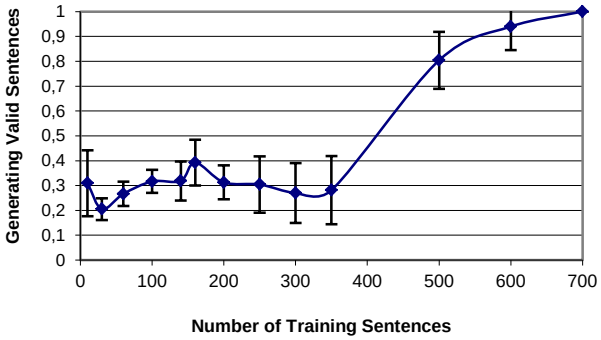
Τα αποτελέσματα της δεύτερης δοκιμής ($B = 10$) παρουσιάζονται στις Εικόνα 32, Εικόνα 33, και Εικόνα 34. Συγκρίνοντας τις γραφικές παραστάσεις με τις αντίστοιχες στις Εικόνα 29, Εικόνα 30, και Εικόνα 31, καταλήγουμε στο συμπέρασμα ότι η απόδοση του egGRIDS+ έχει βελτιωθεί. Η απόδοση σύμφωνα με τα 3 γραφήματα στο προσεγγίζει το 1.0 με σύνολα εκπαίδευσης των 600 προτάσεων. Βέβαια, το αυξημένο μέγεθος της δέσμης επιβαρύνει το συνολικό χρόνο εκπαίδευσης: ο egGRIDS+ απαιτεί περισσότερο από 10 λεπτά ώστε να συγκλίνει σε μια τελική γραμματική όταν εκπαιδεύεται με ένα σύνολο μεγέθους 700, ενώ λιγότερο από 2 λεπτά απαιτούνται όταν το μέγεθος της δέσμης είναι 3.




**Εικόνα 32: Πιθανότητα ανάλυσης μιας έγκυρης πρότασης μεγέθους μέχρι 15 λέξεις. (B = 10)**




**Εικόνα 33: Πιθανότητα ανάλυσης μιας έγκυρης πρότασης μεγέθους μεταξύ 16 και 20 λέξεων. (B = 10)**




**Εικόνα 34: Πιθανότητα παραγωγής μιας έγκυρης πρότασης. (B = 10)**

### 5.5.4 Πείραμα 3: Η γλώσσα ισορροπημένων παρενθέσεων

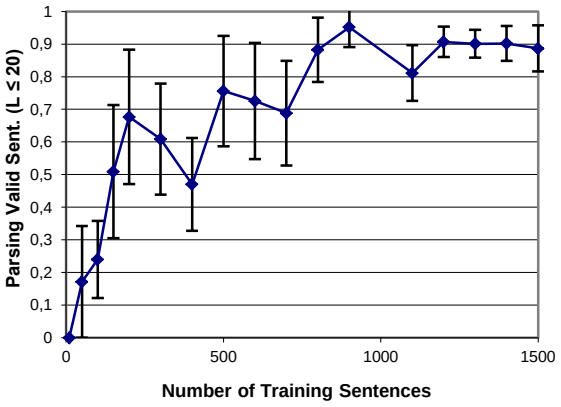
Οι γραμματικές (α) και (β) στον Πίνακας 42, μπορούν να μετατραπούν σε ισοδύναμες *κανονικές γραμματικές (regular grammars)*. Εντούτοις, δεδομένου ότι ο egGRIDS+ είναι σε θέση να συμπεράνει γραμματικές ανεξάρτητες από τα συμφραζόμενα, θα ήταν ενδιαφέρον να μελετηθεί η συμπεριφορά του σε μια τέτοια γραμματική. Σαν γλώσσα αξιολόγησης έχουμε επιλέξει τη γλώσσα Dyck με $k = 1$:

$$S \rightarrow S\ S|[\ S\ ]|\in \qquad (5.8)$$

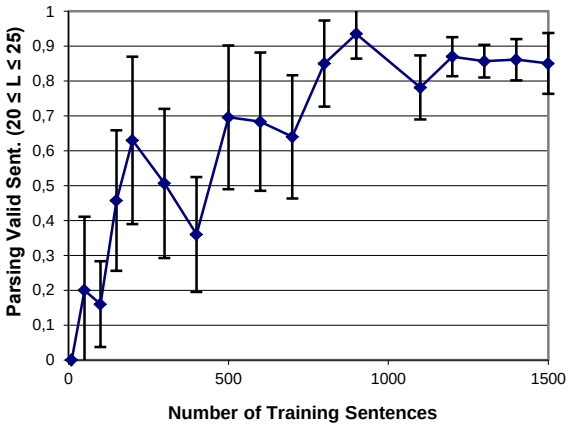
Καθώς η γλώσσα Dyck με $k = 1$ δε μπορεί να χρησιμοποιηθεί για να παραγάγει ένα μεγάλο αριθμό παραδειγματικών προτάσεων εάν περιορίσουμε το μέγιστο μήκος πρότασης, πραγματοποιήσαμε μια διαδικασία διασταυρωμένης επικύρωσης 10 βημάτων. Παρόμοια με τη διαδικασία που ακολουθήθηκε στα προηγούμενα πειράματα, το υποσύνολο (A) χωρίστηκε σε δέκα υποσύνολα του ίδιου μεγέθους και ένα σύνολο αξιολόγησης δημιουργήθηκε από το υποσύνολο (B). Το πείραμα επαναλήφθηκε 10 φορές: κάθε φορά χρησιμοποιούνται εννέα υποσύνολα του (A) για την εκπαίδευση του egGRIDS+, ενώ η εξαχθείσα γραμματική αξιολογείται στο δέκατο αχρησιμοποίητο σύνολο, επαυξημένο με το σύνολο δοκιμής που δημιουργείται από το υποσύνολο (B). Οι εικόνες Εικόνα 35, Εικόνα 36, και Εικόνα 37 παρουσιάζουν τα αποτελέσματα.

Όσον αφορά την απόδοση στην ανάλυση έγκυρων προτάσεων, είτε μεγέθους μέχρι 20 (Εικόνα 35), είτε μεγέθους από 21 έως 25 (Εικόνα 36), ο egGRIDS+ προσεγγίζει το 0.95

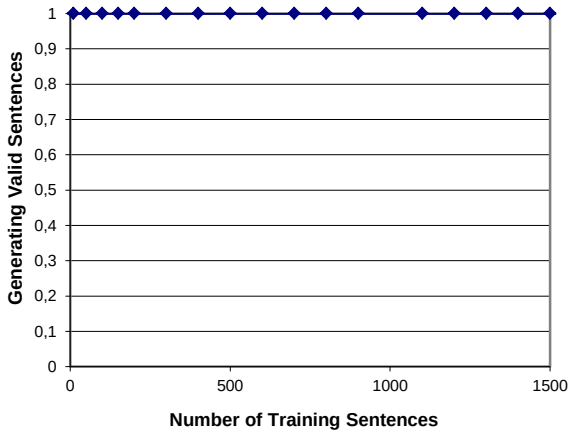
με ένα σύνολο 900 παραδειγμάτων και παραμένει περίπου στο 0.90 για μεγαλύτερα μεγέθη παραδειγμάτων. Ο λόγος πίσω από αυτή την συμπεριφορά είναι ότι ο egGRIDS+ συγκλίνει σε γραμματικές που είναι λιγότερο γενικές από τη γραμματική Dyck, σε μερικές από τις 10 επαναλήψεις του πειράματος. Ένα άλλο ενδιαφέρον σημείο είναι ότι αντίθετα από τα προηγούμενα πειράματα, ο egGRIDS+ δεν συγκλίνει στη γραμματική Dyck, όπως αυτή περιγράφεται στην εξίσωση (5.8), ακόμα και φθάνοντας σε απόδοση 1.0 και στις τρεις μετρικές αξιολόγησης. Οι εξαχθείσες γραμματικές ήταν πιο πολύπλοκες, με την έννοια ότι ένας μεγαλύτερος αριθμός κανόνων και τερματικών συμβόλων περιλήφθηκαν από ότι στη σωστή γραμματική. Εντούτοις, οι εξαχθείσες γραμματικές κωδικοποίησαν τα δύο κύρια φαινόμενα, την αναδρομή και την *κεντρική ενσωμάτωση (center embedding)* της γλώσσας Dyck.




**Εικόνα 35: Πιθανότητα ανάλυσης μιας έγκυρης πρότασης μεγέθους μέχρι 20 λέξεις. (1-errors of omission)**




**Εικόνα 36: Πιθανότητα ανάλυσης μιας έγκυρης πρότασης μεγέθους από 21 μέχρι 25 λέξεων. (1-errors of omission)**




**Εικόνα 37: Πιθανότητα παραγωγής μιας έγκυρης πρότασης. (1-errors of commission)**

### 5.5.5 Πειράματα με σώματα κειμένων μεγάλου μεγέθους

Αν και τα πειράματα με τεχνητά παραγμένες προτάσεις επέτρεψαν τη μελέτη διαφόρων χαρακτηριστικών του egGRIDS+, πειράματα με προτάσεις από γλωσσικά σώματα απαιτούνται προκειμένου να εξεταστεί η ευρωστία του egGRIDS+, καθώς και η κλιμακωσιμότητα του σε πιο σύνθετες γραμματικές περιοχές. Εντούτοις υπάρχουν πολλές δυσκολίες που σχετίζονται με την πειραματική αξιολόγηση σε πραγματικά

δεδομένα. Συνήθως η «τέλεια» γραμματική δεν είναι διαθέσιμη, και συνεπώς δεν είναι δυνατό να συγκριθούν οι εξαχθείσες γραμματικές με αυτή. Επιπλέον, μιας και η «τέλεια» γραμματική δεν είναι γνωστή, καμία επιπρόσθετη πρόταση δε μπορεί να παραχθεί προκειμένου να αξιολογηθεί η επικάλυψή της με μια εξαχθείσα γραμματική. Επομένως, όλες οι απαραίτητες για την αξιολόγηση προτάσεις πρέπει να αντικατασταθούν από επιπρόσθετες, αχρησιμοποίητες προτάσεις από το σώμα κειμένων, εάν αυτό είναι αρκετά μεγάλο. Δυστυχώς αυτό δε λύνει το πρόβλημα μέτρησης της επικάλυψης, δεδομένου ότι δε γίνεται να προσδοκούμε την ύπαρξη (στο σώμα κειμένων) προτάσεων που έχουν παραχθεί από τις εξαχθείσες γραμματικές. Το γεγονός ότι οι εξαχθείσες προτάσεις δεν υπάρχουν στο σώμα κειμένων, δεν σημαίνει απαραιτήτως ότι είναι μη γραμματικές.

Μια ακόμη δυσκολία σχετικά με αυτά τα πειράματα, είναι το γεγονός ότι το σύνολο λέξεων που χρησιμοποιούνται στις προτάσεις δεν είναι κλειστό. Για παράδειγμα, έστω μια γραμματική που προκύπτει από ένα μικρό αριθμό προτάσεων που αποτελούνται από ένα περιορισμένο σύνολο λέξεων. Η εξαχθείσα γραμματική δεν θα είναι ποτέ σε θέση να αναλύσει προτάσεις που περιέχουν λέξεις έξω από το περιορισμένο αυτό σύνολο λέξεων, δεδομένου ότι οι λέξεις δεν θα αντιπροσωπεύονται από τερματικά σύμβολα στη γραμματική. Αυτό το πρόβλημα είναι δύσκολο να επιλυθεί. Ένας λογικός συμβιβασμός είναι να κατασκευαστούν συγκεκριμένα σύνολα εκπαίδευσης και αξιολόγησης που να περιέχουν τις ίδιες λέξεις στις προτάσεις τους. Μια ευκολότερη λύση είναι να παρεμβληθεί ένα «αφαιρετικό» επίπεδο πάνω από τις μεμονωμένες λέξεις, το οποίο να αντιστοιχεί τις λέξεις αυτές σε ένα σταθερό σύνολο συμβόλων. Για παράδειγμα, μια τέτοια «αφαίρεση» είναι η χρησιμοποίηση *του μέρους του λόγου (part of speech - POS)* των λέξεων αντί των πραγματικών λέξεων.

Η εισαγωγή ενός αφαιρετικού επιπέδου μπορεί να λύσει το πρόβλημα των ανοιχτών λέξεων, αλλά οδηγεί επίσης σε ένα νέο πρόβλημα: το αφαιρετικό επίπεδο είναι το ίδιο μια *γενίκευση* πάνω στις αρχικές προτάσεις. Για παράδειγμα, έστω μια πρόταση που μετατρέπεται σε μια γραμματική. Εάν κάθε λέξη αντικατασταθεί από μια κατηγορία μέρους του λόγου (POS), η εξαχθείσα γραμματική από αυτήν την πρόταση θα είναι σε θέση να αναγνωρίσει όχι μόνο την πρόταση από την οποία προήλθε, αλλά και οποιαδήποτε άλλη πρόταση που περιέχει την ίδια ακολουθία μερών του λόγου. Το επίπεδο γενίκευσης εξαρτάται από τις μορφολογικές λεπτομέρειες που αντιπροσωπεύονται από τις κατηγορίες λέξεων. Για παράδειγμα, μια ταξινόμηση βασισμένη μόνο στην πληροφορία για τα μέρη του λόγου των λέξεων, αντιστοιχεί σε ένα σημαντικό βήμα γενίκευσης. Από την άλλη πλευρά, εάν η ταξινόμηση αυξάνεται με πρόσθετες πληροφορίες όπως το γένος, ο αριθμός και το πρόσωπο, το επίπεδο γενίκευσης είναι χαμηλότερο, δεδομένου ότι ο αριθμός κατηγοριών είναι μεγαλύτερος, με κάθε κατηγορία να αντιστοιχεί σε έναν μικρότερο αριθμό λέξεων. Ένας υψηλός βαθμός γενίκευσης μπορεί να έχει αρνητική επίδραση στην πιθανότητα παραγωγής έγκυρων προτάσεων χρησιμοποιώντας τις εξαχθείσες γραμματικές. Ακόμη και οι προτάσεις εκπαίδευσης, όταν αντιμετωπίζονται ως αρχική γραμματική, μπορούν να παραγάγουν μη γραμματικές προτάσεις. Τα πράγματα μπορούν να γίνουν χειρότερα εάν οι λέξεις καταχωρούνται λανθασμένα σε κατηγορίες από ένα αυτοματοποιημένο σύστημα. Εάν οι κατηγορίες μερών του λόγου χρησιμοποιούνται ως επίπεδο αφαίρεσης, οι λάθος ταξινομήσεις δεν είναι σπάνιες δεδομένου ότι οι ίδιες λέξεις μπορούν να ανήκουν σε διαφορετικές κατηγορίες που δύσκολα μπορούν να αποσαφηνιστούν. Δεδομένου ότι ο egGRIDS+ γενικεύει μόνο μια αρχική γραμματική που εξάγεται από τις προτάσεις εκπαίδευσης, δεν μπορεί να εξαλείψει αυτά τα λάθη, που θα συγκλίνουν σε μια τελική γραμματική που μπορεί να παραγάγει πολλές μη γραμματικές προτάσεις.

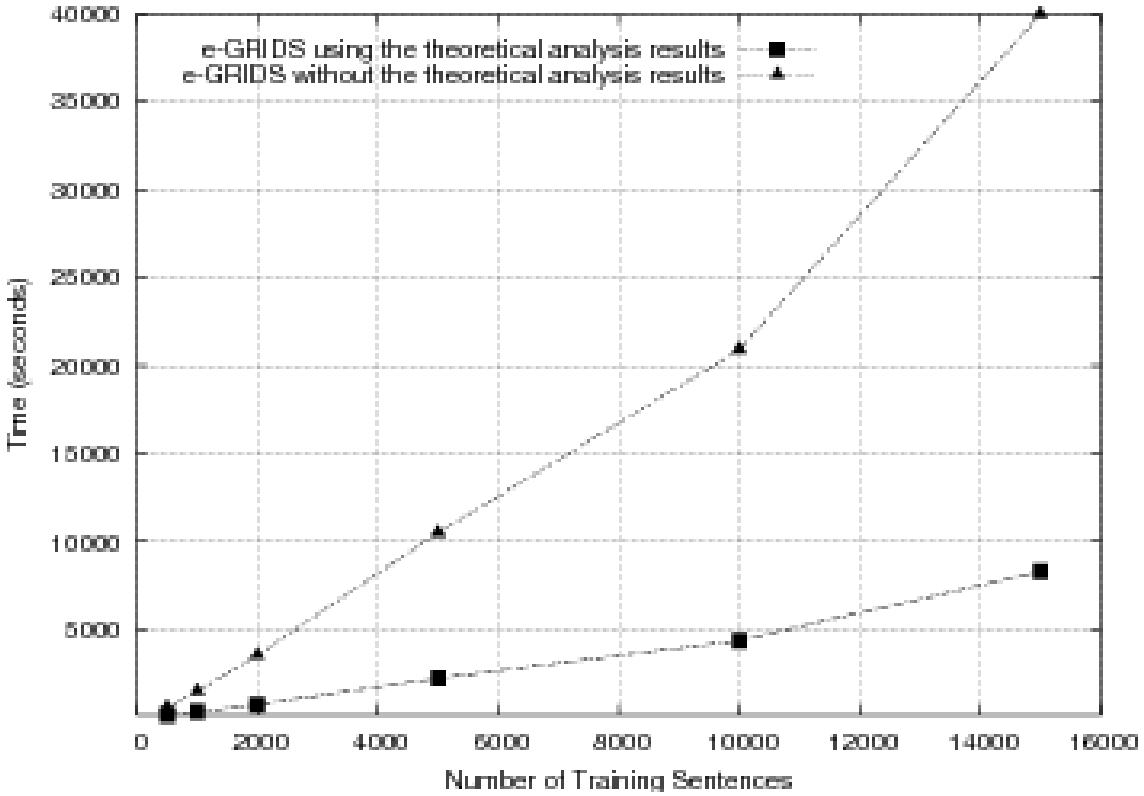
Παρά τα προβλήματα που συνδέονται με τη χρήση ενός επιπέδου αφαίρεσης, έχουμε επιλέξει να το χρησιμοποιήσουμε για την εφαρμογή του egGRIDS+ στις προτάσεις από τα σώματα κειμένων φυσικής γλώσσας. Ο λόγος που δεν έχουμε ακολουθήσει το παράδειγμα της δημιουργίας της κατασκευής συνόλων εκπαίδευσης και δοκιμής που μοιράζονται τις ίδιες λέξεις, οφείλεται κυρίως στην απρόβλεπτη φύση του. Επιπλέον, η χρήση ενός πολύ μεγάλου αριθμού τερματικών συμβόλων θα οδηγούσε σε ένα εξίσου μεγάλο υποσύνολο τερματικών κανόνων στην αρχική γραμματική. Αυτό θα οδηγούσε σε μια υπερβολική αύξηση στο μήκος περιγραφής της γραμματικής σε σχέση με το μήκος παραγωγών, κάτι που θα απαγόρευε στους τελεστές τη βελτίωση της αρχικής γραμματικής.

Το σώμα που χρησιμοποιήθηκε στο πείραμα ήταν ένα μέρος του σώματος SEMCORE [125]. Ο κύριος σκοπός αυτού του πειράματος ήταν να αξιολογηθεί η δυνατότητα του egGRIDS+ να χειριστεί μεγάλα σύνολα εκπαίδευσης, και να συγκλίνει σε μια τελική γραμματική σε ένα λογικό χρονικό διάστημα. Ο χρόνος που απαιτείται από τον egGRIDS+ για να ολοκληρώσει την εκπαίδευση παρουσιάζεται στην Εικόνα 38. Το γράφημα δείχνει το χρόνο που απαιτείται για να ολοκληρωθεί η εκπαίδευση από δύο εκδόσεις του egGRIDS+: η πρώτη έκδοση χρησιμοποιεί τα αποτελέσματα της θεωρητικής ανάλυσης, όπως παρουσιάστηκαν στις προηγούμενες ενότητες, ενώ η δεύτερη έκδοση δε χρησιμοποιεί αυτήν την ανάλυση, προσφεύγοντας στον τρόπο που ο αρχικός GRIDS πραγματοποιούσε αναζήτηση στο χώρο των πιθανών γραμματικών. Το μέγεθος της δέσμης τέθηκε στην τιμή 1. Όπως φαίνεται από τη γραφική παράσταση, ο χρόνος που απαιτείται από τον βελτιστοποιημένο egGRIDS+ είναι σημαντικά χαμηλότερος από το χρόνο που απαιτείται από τον απλό GRIDS. Αυτό είναι πολύ σημαντικό, δεδομένου ότι επιτρέπει στον egGRIDS+ να εκπαιδευτεί με σημαντικά μεγαλύτερα σύνολα από τον απλό GRIDS, συγκλίνοντας σε μια τελική γραμματική μέσα σε ένα λογικό χρονικό διάστημα.

Πρέπει επίσης να σημειωθεί ότι αυτή η βελτιστοποίηση είναι βασισμένη σε μια θεωρητική ανάλυση όπου κανένας συμβιβασμός δεν έχει γίνει σχετικά με την ακρίβεια των παραγόμενων γραμματικών, σε σχέση με τον GRIDS. Περαιτέρω σημαντικές βελτιώσεις είναι δυνατές με την ανοχή ενός μικρού βαθμού ανακρίβειας. Για παράδειγμα, ένα σημαντικό ποσοστό του χρόνου επεξεργασίας ξοδεύεται κατά τη διάρκεια του τέλους της διαδικασίας εκπαίδευσης, όπου ο egGRIDS+ παράγει πολλές παρόμοιες γραμματικές, με ασήμαντες διαφορές στο μήκος περιγραφής, προκειμένου να επιλεχτεί η καλύτερη ως τελική γραμματική. Τερματίζοντας την αναζήτηση όταν μειώνεται το οριακό κέρδος στο μήκος περιγραφής κάτω από ένα ορισμένο κατώτατο όριο, μπορεί να μειωθεί σημαντικά ο χρόνος επεξεργασίας. Πρόσθετη βελτιστοποίηση είναι δυνατή με τον υπολογισμό μόνο των πιο κυρίαρχων τμημάτων των εξισώσεων που προέκυψαν από τη θεωρητική ανάλυση.

Η αξιολόγηση της διαδικασίας εκμάθησης του egGRIDS+ σε αυτό το πείραμα δεν είναι απλή, δεδομένου ότι ο στόχος εκμάθησης δεν είναι καθορισμένος με σαφήνεια. Προκειμένου να πάρουμε να ένδειξη της δυνατότητας του egGRIDS+ να γενικεύσει από μεγάλα σύνολα εκπαίδευσης, πραγματοποιήσαμε έναν απλό έλεγχο. Μετά την εκπαίδευση του egGRIDS+ με 2000 προτάσεις, επιλέξαμε τυχαία 200 νέες προτάσεις και μετρήσαμε πόσες νέες προτάσεις μπόρεσαν να αναλυθούν από τη γραμματική που έχει μαθευτεί, καθώς επίσης και από το σύνολο παραδειγμάτων εκπαίδευσης που μετατράπηκε σε μια (αρχική) γραμματική. Η αρχική γραμματική ήταν σε θέση να αναλύσει 25 προτάσεις, ενώ η γραμματική που είχε εξαχθεί ήταν σε θέση να αναλύσει 4 προτάσεις περισσότερες, συνολικά 29. Αυτοί οι αριθμοί, παρέχουν μια αρχική ένδειξη της δυνατότητας του egGRIDS+ να συγκλίνει σε πιο γενικές γραμματικές σε σχέση με την απλή περίπτωση. Από την άλλη, περαιτέρω πειραματισμός σε ένα πιο ελεγχόμενο

πείραμα απαιτείται προκειμένου να αποδειχθεί η αξία του egGRIDS+ στην εκμάθηση γραμματικών ανεξάρτητων από συμφραζόμενα για φυσικές γλώσσες.




**Εικόνα 38: Ο χρόνος που απαιτείται από τον egGRIDS+, με και χωρίς τη χρήση αποτελεσμάτων θεωρητικής ανάλυσης.**

### 5.5.6 Ο διεθνής διαγωνισμός "Omphalos"

Ο διεθνής διαγωνισμός "Omphalos"[5] [14] ήταν ένας διαγωνισμός εκμάθησης γραμματικών ανεξάρτητων από τα συμφραζόμενα, ο οποίος διοργανώθηκε στο πλαίσιο του συνεδρίου ICGI 2004 [126]. Ο διαγωνισμός αποτελούταν από ένα σύνολο προβλημάτων της μορφής $(\Sigma, X, Y)$, όπου $\Sigma$ είναι ένα πεπερασμένο αλφάβητο, $X \subset \Sigma^* \times \{0,1\}$ είναι ένα επισημειωμένο σύνολο από προτάσεις (ανάλογα με το αν περιγράφονται ή όχι από την γραμματική που πρέπει να μαθευτεί), και $Y \subset \Sigma^*$ ένα σύνολο από προτάσεις χωρίς οποιαδήποτε επισημείωση. Οι συμμετέχοντες κλήθηκαν να επισημειώσουν το σύνολο των προτάσεων αξιολόγησης $Y$, και να υποβάλλουν την επισημείωση σε ένα ηλεκτρονικό σύστημα αποτίμησης (oracle), το οποίο απαντούσε $\{0,1\}$ ανάλογα με το αν ή αποτίμηση ήταν επιτυχής $(1)$ ή όχι $(0)$. Το κριτήριο επιτυχίας ήταν ο ακριβής χαρακτηρισμός *όλων* των προτάσεων, ανάλογα με το αν ανήκουν ή όχι στην γλώσσα στόχο. Ο αριθμός των αποτιμήσεων μέσω του ηλεκτρονικού συστήματος αποτίμησης έπρεπε να διατηρηθεί μικρός. Περισσότερες λεπτομέρειες σχετικά με την κατασκευή των γραμματικών, των παραδειγμάτων, και του τρόπου αξιολόγησης μπορούν να βρεθούν στις αναφορές [14] και [127]. Μια συνοπτική περιγραφή των προβλημάτων δίνεται στον πίνακα (Πίνακας 43).

[5] The "Omphalos" context-free language learning competition: http://www.irisa.fr/Omphalos/

Ο αλγόριθμος egGRIDS+ συμμετείχε στον διαγωνισμό, καταφέρνοντας να λύσει το πρώτο πρόβλημα, αγνοώντας το σύνολο των αρνητικών παραδειγμάτων του συγκεκριμένου προβλήματος. Ο αλγόριθμος που έλυσε το δυσκολότερο πρόβλημα στο διαγωνισμό [127] χρησιμοποιούσε ένα συνδυασμό *αμοιβαίας πληροφορίας (mutual information)* για να υπολογίσει την πιθανότητα ένα *ν-γραμμα (n-gram)* να αποτελεί παραγωγή ενός μη-τερματικού συμβόλου και ευριστικών, απαιτώντας ανθρώπινη παρέμβαση σε περιπτώσεις ισοβαθμιών. Ο αλγόριθμος αυτός υλοποιήθηκε με την μορφή ευριστικής στρατηγικής αναζήτησης, η οποία προτείνει συγχωνεύσεις συμβόλων σε μη-τερματικά σύμβολα, εντός του egGRIDS+, συνδυάζοντας τον με το ευριστικό του ελάχιστου μήκους περιγραφής του egGRIDS+. Ο συνδυασμός κατάφερε να επιλύσει τα ίδια προβλήματα με τον αλγόριθμο [127], χωρίς την ανάγκη για ανθρώπινη παρέμβαση σε περίπτωση ισοβαθμιών.

| Πρόβλημα | Δεδομένα | $\|\Sigma\|$ | Μέγεθος παραδειγμάτων | Λυμένο | $\|N\|$ |
|---|---|---|---|---|---|
| 1 | $+/-$ | 5 | 801 | Μικρό | 7 |
| 2 | $+$ | 5 | 280 | Ναι | 10 |
| 3 | $+/-$ | 24 | 924 | Μικρό | 13 |
| 4 | $+$ | 24 | 418 | Ναι | 15 |
| 5 | $+/-$ | 24 | 1238 | Ναι | 20 |
| 6 | $+$ | 24 | 498 | Μικρό | 19 |

**Πίνακας 43: Συνοπτική περιγραφή των προβλημάτων του διαγωνισμού "Omphalos".**

## 5.6 Συνεισφορά

Σε αυτό το κεφάλαιο παρουσιάστηκε ο νέος αλγόριθμος egGRIDS+ για την εξαγωγή, από θετικά παραδείγματα, γραμματικών ανεξάρτητων από τα συμφραζόμενα. Ο αλγόριθμος χρησιμοποιεί μια ευριστική αναζήτηση, βασισμένη στην αρχή του ελάχιστου μήκους περιγραφής, ώστε να αποφύγει την υπεργενίκευση, μια συνέπεια της απουσίας αρνητικών παραδειγμάτων. Ένα από τα κύρια πλεονεκτήματά του νέου αλγορίθμου είναι η υπολογιστική επάρκειά του, που διευκολύνει την κλιμακωσιμότητα του σε μεγάλα σύνολα παραδειγμάτων. Η δυναμική συμπεριφορά των τελεστών αναζήτησης έχει αναλυθεί θεωρητικά, οδηγώντας στη βελτιστοποίηση της διαδικασίας συμπερασμού (μειώνοντας την υπολογιστική πολυπλοκότητα από κυβική σε τετραγωνική), αίροντας την απαίτηση της παραγωγής όλων των γραμματικών σε κάθε επανάληψη του προκατόχου του. Η απόδοση του egGRIDS+ αξιολογήθηκε σε τεχνητά σύνολα εκπαίδευσης, καθώς επίσης και σε ένα μεγάλο σώμα κειμένων.

Όσον αφορά τη συμπεριφορά του egGRIDS+ κατά τη διαδικασία εκπαίδευσης, πραγματοποιήθηκαν πειράματα με τη βοήθεια τεχνητών παραδειγμάτων. Τα αποτελέσματα έδειξαν ότι ο αλγόριθμος είναι σε θέση να συμπεράνει γραμματικές που αποδίδουν ικανοποιητικά,, βασισμένος σε σχετικά μικρά σύνολα παραδειγμάτων εκπαίδευσης. Το ενσωματωμένο ευριστικό του ελάχιστου μήκους περιγραφής εμφανίζεται να βοηθά τον αλγόριθμο να αποφύγει την υπεργενίκευση, τουλάχιστον για τις απλές τεχνητές γλώσσες που εξετάστηκαν. Ο egGRIDS+ φαίνεται να είναι σε θέση να γενικεύσει σωστά, συμπεραίνοντας γραμματικές που μοντελοποιούν επιτυχώς τα σύνολα εκπαίδευσης και συγχρόνως δεν παράγουν πολλές μη γραμματικές προτάσεις. Ένα άλλο ενδιαφέρον χαρακτηριστικό του αλγορίθμου, είναι η ικανότητά του να εξάγει γραμματικές ικανές να αναγνωρίσουν μεγαλύτερου μήκους προτάσεις από εκείνες που χρησιμοποιήθηκαν κατά την εκπαίδευση, καθώς ο αλγόριθμος είναι σε θέση να εξάγει

αναδρομικές γραμματικές που μοντελοποιούν απλές αναδρομικές περιπτώσεις, όπως όταν μερικές κατηγορίες λέξεων ή φράσεις τείνουν να επαναλαμβάνονται.

Πιο χρήσιμη για πρακτικές εφαρμογές είναι η κλιμακωσιμότητα του egGRIDS+ σε μεγάλα σύνολα παραδειγμάτων. Όσον αφορά την υπολογιστική του επάρκεια, τα αποτελέσματα είναι πολύ ικανοποιητικά, καθώς ο αλγόριθμος μπορεί να χειριστεί μεγάλα σύνολα παραδειγμάτων σε σημαντικά μειωμένο χρονικό διάστημα, σε σχέση με τον απλό GRIDS. Ο egGRIDS+ χρησιμοποιεί τα αποτελέσματα μιας θεωρητικής ανάλυσης της δυναμικής συμπεριφοράς των τελεστών μάθησής του, όπου κανένας συμβιβασμός δεν έχει γίνει σχετικά με την ακρίβεια των παραγόμενων γραμματικών σε σχέση με τον προκάτοχό του, GRIDS. Εάν μπορούσαν να γίνουν έστω και μικροί συμβιβασμοί ακρίβειας, ο egGRIDS+ θα μπορούσε να γίνει ακόμα πιο αποδοτικός.

# 6. Εξαγωγή Πληροφορίας: Εξαγωγή Σχέσεων μεταξύ Ονομάτων Οντοτήτων

Η *εξαγωγή σχέσεων μεταξύ ονομάτων οντοτήτων (relation extraction)* είναι ένα βασικό στάδιο της εξαγωγής πληροφορίας. Έχοντας αναγνωρίσει τα ονόματα των οντοτήτων που βρίσκονται σε ένα κείμενο, το στάδιο της εξαγωγής σχέσεων μεταξύ ονομάτων οντοτήτων καλείται να αποσαφηνίσει τις σχέσεις που υπάρχουν μεταξύ όλων αυτών των οντοτήτων, έτσι ώστε να μπορέσουν να εξαχθούν ομάδες συνδεδεμένων οντοτήτων, οι οποίες συνήθως θεωρούνται ότι συμμετέχουν στο ίδιο γεγονός.

Το κεφάλαιο αυτό παρουσιάζει μια καινοτομική προσέγγιση στην εργασία της εξαγωγής σχέσεων μεταξύ ονομάτων οντοτήτων, χρησιμοποιώντας μηχανική μάθηση, και συγκεκριμένα τον αλγόριθμο επαγωγικής εξαγωγής γραμματικών egGRIDS+, ο οποίος αναπτύχθηκε στο πλαίσιο αυτής της διατριβής και περιγράφεται λεπτομερώς στο κεφάλαιο 5.

Στις επόμενες ενότητες ορίζεται το πρόβλημα της εξαγωγής σχέσεων μεταξύ ονομάτων οντοτήτων, παρουσιάζεται η σχετική διεθνή βιβλιογραφία στην οποία συμβάλει η παρούσα διατριβή, αναπτύσσεται η μεθοδολογία που ακολουθήθηκε, και τέλος αποτιμάται η προτεινόμενη μεθοδολογία σε κείμενα εκφρασμένα στην Αγγλική γλώσσα.

## 6.1 Ορισμός προβλήματος

Ο όρος *εξαγωγή συσχετίσεων ή εξαγωγή σχέσεων (relation extraction ή relationship extraction)*, αναφέρεται στην εργασία της αναγνώρισης και κατηγοριοποίησης σημασιολογικών σχέσεων, εντός ενός συνόλου από οντότητες, τυπικά σε κειμενικά δεδομένα. Συνήθως οι σχέσεις που πρέπει να εξαχθούν είναι δυαδικές, δηλ. αφορούν μόνο δύο οντότητες, και μετά την αναγνώρισή τους συχνά πρέπει να κατηγοριοποιηθούν στον κατάλληλο τύπο σχέσης, με τις κατηγορίες να ποικίλουν ανάλογα με την θεματική περιοχή και το σύστημα εξαγωγής πληροφορίας.

Η εργασία της εξαγωγής μιας μόνο κατηγορίας (τύπου) σχέσης μπορεί να περιγραφεί ως εξής: Με δεδομένο ένα σύνολο δεδομένων $D$, και μια σχέση $n$-ορισμάτων $Rel$, με ορίσματα $X, Y, \dots, Z$, πρέπει να βρεθούν όλα τα *στιγμιότυπα (instances)* $x \in X$, $y \in Y$, …, $z \in Z$ $(x, y, z \in D)$, έτσι ώστε το $Rel(x, y, \dots, z)$ να είναι αληθές [128]. Η προτεινόμενη σε αυτό το κεφάλαιο προσέγγιση επικεντρώνεται στην εξαγωγή δυαδικών σχέσεων από σώματα κειμένων, προσπαθώντας να συλλάβει τις γλωσσικές ενδείξεις του κειμένου που συνδέει δυο συσχετιζόμενες οντότητες.

## 6.2 Βιβλιογραφική επισκόπηση

Ο όρος *«εξαγωγή σχέσεων» (relation extraction)* αναφέρεται στην εργασία της αναγνώρισης συσχετίσεων που πιθανώς υπάρχουν μεταξύ οντοτήτων σε κειμενικά δεδομένα. Όντας μια απαιτητική υπο-εργασία της εξαγωγής πληροφοριών, εξάγει την απαιτουμένη γνώση για να περάσει από την αναγνώριση ονομάτων οντοτήτων στην ερμηνεία των δεδομένων και στην κατανόηση τους. Λόγω αυτού, έχει γίνει μια από τις κυρίες περιοχές ερευνάς στο πεδίο της υπολογιστικής επεξεργασίας φυσικής γλώσσας. Οι αρχικές προσπάθειες ήταν κυρίως βασισμένες σε κανόνες [19] χειρωνακτικά κατασκευασμένους, βασισμένους στα αποτελέσματα συντακτικής ανάλυσης. Η τρέχουσα ερευνά εστιάζει κυρίως στη χρήση τεχνικών μηχανικής μάθησης. Οι *επιβλεπόμενες (supervised)* τεχνικές μηχανικής μάθησης έχουν αποδειχτεί αρκετά αποτελεσματικές για την εργασία αυτή [129], [130], [131], ενώ αρκετές προσεγγίσεις απασχολούν ημι-επιβλέπομενες ή μάθηση χωρίς επίβλεψη, ( [132], [133], [134], [135], [136], [137]), χρησιμοποιώντας επίσης το Διαδίκτυο σαν ένα σώμα κειμένων.

Από όσο γνωρίζουμε, υπάρχει πολύ περιορισμένη έρευνα στην περιοχή της εξαγωγής σχέσεων μεταξύ ονομάτων οντοτήτων με χρήση επαγωγικής εξαγωγής γραμματικών. Στην εργασία [138] παρουσιάζεται μια ημι-αυτόματοποιημένη προσέγγιση, η οποία εκμεταλλεύεται τα αποτελέσματα στατιστικής ανάλυσης των σωμάτων κειμένων (κυρίως *συνταυτίσεις ρημάτων – verb concordances*) για την εξαγωγή *προτύπων (patterns)*. Αυτά τα πρότυπα, αφού επικυρωθούν από ειδικό, μετατρέπονται σε ένα *σύνολο από αυτόματα πεπερασμένων καταστάσεων (set of finite state automata)*. Παρομοίως, στην αναφορά [139] χρησιμοποιούνται επίσης αυτόματα, τα οποία δημιουργήθηκαν από χειρωνακτικά κατασκευασμένα πρότυπα. Ωστόσο, και οι δυο προσεγγίσεις λειτουργούν έχοντας σαν είσοδο δέντρα συντακτικής ανάλυσης, τα οποία έχουν εξαχθεί εφαρμόζοντας συντακτική ανάλυση, και περιλαμβάνουν την εφαρμογή προτύπων που έχουν εξαχθεί είτε χειρωνακτικά είτε ημι-αυτοματοποιημένα, για την εξαγωγή συσχετίσεων μεταξύ ονομάτων οντοτήτων.

Από την άλλη πλευρά, υπάρχουν ορισμένες προσεγγίσεις που παρουσιάζουν αρκετή ομοιότητα με αλγορίθμους *γραμματικού συμπερασμού (grammatical inference)*, με την έννοια ότι προσπαθούν να γενικεύσουν πρότυπα/κανόνες που έχουν εξορυχτεί [140], ή να τροποποιήσουν κανόνες αναγνώρισης, εφαρμόζοντας τελεστές παρόμοιους με αυτούς που χρησιμοποιούνται από αλγόριθμους επαγωγικής εξαγωγής γραμματικών, όπως ο egGRIDS+ [136]. Το σύστημα εξαγωγής πληροφορίας LearningPinocchio [141] έχει κτιστεί πάνω στον αλγόριθμο $LP^2$ [140], και το οποίο δημιουργεί ένα αρχικό σύνολο κανόνων από θετικά παραδείγματα, τα οποία γενικεύονται με την αξιοποίηση αποτελεσμάτων από γλωσσική ανάλυσης/ρηχή συντακτική μελέτη, ώστε να αφαιρεθούν περιορισμοί από τους κανόνες. Η υπεργενίκευση ελέγχεται μέσω *αρνητικών παραδειγμάτων*, τα οποία αποκτώνται αυτόματα από το σώμα κειμένων, χρησιμοποιώντας την παραδοχή ότι οτιδήποτε δεν έχει αναγνωριστεί σαν θετικό παράδειγμα είναι ένα αρνητικό παράδειγμα.

Ακολουθώντας μια παρόμοια προσέγγιση, το σύστημα DARE [136] εκκινεί απαιτώντας έναν μικρό αριθμό από *ενδεικτικούς κανόνες (seed rules)*, οι οποίοι χρησιμοποιούνται για να επισημειώσουν ένα σώμα κειμένων. Έχοντας σαν είσοδο δέντρα συντακτικής ανάλυσης, το σύστημα DARE ακολουθεί *προσέγγιση από κάτω προς τα πάνω (bottom-up approach)*, σε μια προσπάθεια κατασκευής περισσότερο γενικευμένων κανόνων, με τη συγχώνευση κόμβων από τα δέντρα συντακτικής ανάλυσης των προτάσεων, μια λειτουργία που αποτελεί επίσης μέρος της διαδικασίας αναζήτησης στον χώρο των πιθανών γραμματικών του egGRIDS+, ως ένας από τους τελεστές γενίκευσης. Η υπεργενίκευση ελέγχεται με την προσπάθεια εξισορρόπησης μεταξύ του αριθμού των σχέσεων που αναγνωρίζονται σε σχετικά έγγραφα με την θεματική περιοχή, σε σχέση με τον αριθμό σχέσεων που εξάγονται από κείμενα άσχετα με την θεματική περιοχή από τους κανόνες.

Η προσέγγιση που ερευνήθηκε στο πλαίσιο αυτής της διατριβής για την εργασία της εξαγωγής σχέσεων μεταξύ ονομάτων οντοτήτων, διαφέρει από αυτά δύο συστήματα, μη απαιτώντας δέντρα συντακτικής ανάλυσης σαν είσοδο (τα οποία χρησιμοποιούνται είτα σαν σημείο εκκίνησης για την εξαγωγή κανόνων στο DARE ή για καθοδήγηση της διαδικασίας γενίκευσης στο LearningPinocchio). Η προσέγγιση που προτείνεται βασίζεται σε ευριστικά (και κυρίως την ελαχιστοποίηση του μήκους περιγραφής – MDL) για την αποφυγή της υπεργενίκευσης, και δεν απαιτεί την παρουσία αρνητικών παραδειγμάτων ή κειμένων από διαφορετικές θεματικές περιοχές.

## 6.3 Η προσέγγιση

Η αξιοποίηση της μηχανικής μάθησης σε διάφορα στάδια της εξαγωγής πληροφορίας από κείμενα αποτελεί βασικό πυλώνα αυτής της διδακτορικής διατριβής, οπότε είναι

αναμενόμενη η πρόταση χρήσης μηχανικής μάθησης και για την εργασία της αναγνώρισης συσχετίσεων μεταξύ ονομάτων οντοτήτων. Υποθέτοντας την ύπαρξη ενός αναγνωριστή ονομάτων οντοτήτων (NERC), η προτεινόμενη προσέγγιση εξάγει δυαδικές σχέσεις μεταξύ ονομάτων οντοτήτων τα οποία έχουν ήδη αναγνωριστεί σε κειμενικά δεδομένα. Λειτουργώντας στο επίπεδο των προτάσεων, η προσέγγιση εξάγει μια γραμματική ανεξάρτητη από τα συμφραζόμενα (CFG), η οποία καταγράφει *πρότυπα (patterns)* συνδέσμων μεταξύ συνδεδεμένων ονομάτων οντοτήτων, η οποία συνάγεται μονό από θετικά παραδείγματα. Για την επαγωγική εξαγωγή γραμματικών ανεξάρτητων από τα συμφραζόμενα από θετικά παραδείγματα, χρησιμοποιείται ο αλγόριθμος egGRIDS+, ο οποίος αναπτύχθηκε στο πλαίσιο αυτής της διατριβής και ο οποίος παρουσιάζεται λεπτομερώς στο κεφάλαιο 5. Όπως έχει ήδη αναφερθεί, η ανάγκη για αρνητικά παραδείγματα τα οποία θα βοηθούσαν στον έλεγχο της υπεργενίκευσης, εξαλείφεται μέσω της χρήσης ευριστικού, που βασίζεται στο ελάχιστο μήκος περιγραφής (MDL) [123].

Ο κύριος στόχος αυτής της προσέγγισης, είναι να εξετάσει την καταλληλότητα της επαγωγικής εξαγωγής γραμματικών για το έργο της εξαγωγής συσχετίσεων μεταξύ ονομάτων οντοτήτων. Ένα μεγάλο μέρος της έρευνας που έχει γίνει στην περιοχή, αξιοποιεί τα αποτελέσματα της συντακτικής ανάλυσης, μαζί με στατιστικές πληροφορίες που προέρχονται από μεγάλα σώματα κειμένων, για να εξάγει/γενικεύσει κανόνες/πρότυπα που να μπορούν να εφαρμοστούν στην εξαγωγή σχέσεων ( [138], [139], [136], [141]). Αρχίζοντας από ένα δέντρο συντακτικής ανάλυσης, το οποίο μπορεί να γενικευθεί με την συγχώνευση κόμβων [136], ή από ακολουθίες λέξεων που μπορούν να μετατραπούν σε κανόνες με την εξαγωγή πληροφοριών από δέντρα συντακτικής ανάλυσης [140], διάφορα ευριστικά έχουν προταθεί για την καθοδήγηση της διαδικασίας γενίκευσης και τον έλεγχο του επιπέδου της γενίκευσης που εκτελείται, προκειμένου να αποφευχθεί η υπερ/υπο-γενίκευση.

Ένας αλγόριθμος επαγωγικής εξαγωγής γραμματικών γενικής χρήσης, από την άλλη πλευρά, περιλαμβάνει ήδη την απαιτουμένη στρατηγική για την καθοδήγηση της γενίκευσης, μαζί με τα απαιτούμενα κριτήρια τερματισμού. Επιπροσθέτως, ένας αλγόριθμος επαγωγικής εξαγωγής γραμματικών αναμένεται να είναι σε θέση να συλλάβει τη συντακτική δομή της γλώσσας, ελαχιστοποιώντας την ανάγκη της εκτέλεσης συντακτικής ανάλυσης, καθιστώντας την προσέγγιση καταλληλότερη για θεματικές περιοχές όπου η συντακτική ανάλυση εμφανίζει μειωμένη απόδοση ή για γλώσσες όπου η απαιτούμενοι γλωσσικοί πόροι δεν είναι διαθέσιμοι.

Τα κριτήρια που οδήγησαν στην επιλογή του egGRIDS+ για την εξαγωγή συσχετίσεων, περιλαμβάνουν την ικανότητα του να εξάγει γραμματικές μονό από θετικά παραδείγματα, την ποικιλία των υλοποιηθέντων στρατηγικών αναζήτησης, αλλά και την απόδοση του αλγόριθμου στον διαγωνισμό εξαγωγής γραμματικών, ανεξάρτητων από τα συμφραζόμενα, “Omphalos” [14]. Τα αποτελέσματα αξιολόγησης του egGRIDS+ καταδεικνύουν ότι η προτεινόμενη μέθοδος αποδίδει συγκρίσιμα με τις καλύτερες προσεγγίσεις της διεθνούς βιβλιογραφίας, ενώ επιδεικνύει και μια τάση προς την διατήρηση της *ακρίβειας (precision)* έναντι της *ανάκλησης (recall)*, γεγονός που μπορεί να αποδοθεί στην συντηρητική στρατηγική γενίκευσης που υλοποιεί ο αλγόριθμος egGRIDS+. Πρωτοποριακές πτυχές της προτεινόμενης μεθόδου περιλαμβάνουν την ικανότητα αυτόνομης εκμάθησης γραμματικής, χωρίς να βασίζεται στην διαθεσιμότητα γλωσσικών πόρων όπως αναγνωριστών μέρων-του-λογού ή συντακτικούς αναλυτές. Για παράδειγμα, πολλές υπάρχουσες προσεγγίσεις χρησιμοποιούν τα αποτελέσματα της συντακτικής ανάλυσης για να γενικεύσουν μια αρχική υπόθεση, ή χρησιμοποιούν συντακτικά δέντρα ως την αρχική υπόθεση, η οποία γενικεύεται μέσω συγχωνεύσεων κόμβων. Η προτεινόμενη προσέγγιση, σε αντίθεση με τις υπάρχουσες προσεγγίσεις,

ελαχιστοποιεί αυτές τις εξαρτήσεις σε επεξεργαστικούς πόρους, με τίμημα την εξαγωγή της απαιτουμένης γνώσης κατευθείαν από τα δεδομένα. Αντί για την εφαρμογή ευριστικών στην προσαρμογή μιας γραμματικής γενικής χρήσης, όπως η γραμματική ενός συμβατικού συντακτικού αναλυτή, σε μια εξειδικευμένη γραμματική για εξαγωγή σχέσεων, η προτεινόμενη προσέγγιση επικεντρώνεται στην απευθείας εξαγωγή της εξειδικευμένης γραμματικής.

Εξίσου σημαντικό είναι το γεγονός ότι η προτεινόμενη προσέγγιση δεν στηρίζεται σε καμία μορφή αρνητικής πληροφορίας, είτε άμεσης, όπως η απαίτηση για αρνητικά παραδείγματα ή άσχετα έγγραφα, είτε έμμεσης, όπως υποθέτοντας ότι όλα τα δεδομένα που δεν αναγνωρίζονται ως θετικά είναι αρνητικά παραδείγματα. Το πλεονέκτημα της μη απαίτησης επιπρόσθετων πόρων και αρνητικών πληροφοριών, αυξάνει τη δυνατότητα εφαρμογής της προτεινομένης προσέγγισης όχι μονό σε νέες θεματικές περιοχές και γλώσσες, αλλά ίσως ακόμα και σε διαφορετικές οργανώσεις μάθησης, όπως για παράδειγμα *ελάχιστα επιβλεπόμενες προσεγγίσεις (minimally supervised approaches)*: απαιτώντας μόνο ένα περιορισμένο αριθμό θετικών παραδειγμάτων (ή κανόνων), ο στόχος είναι να εξαχθεί μια γραμματική μέσω της *μεθόδου επανεκκίνησης (bootstrapping)* σε σχέση με ένα σώμα κειμένων.

### 6.3.1 Εξάγοντας συσχετίσεις

Στη φάση της εκπαίδευσης η μέθοδος απαιτεί σαν είσοδο ένα σύνολο από παραδείγματα εκπαίδευσης. Τα απαιτούμενα παραδείγματα μπορούν να αποκτηθούν εύκολα, αν υποτεθεί η ύπαρξη ενός σώματος κειμένου, επισημειωμένου με ονόματα οντοτήτων και σχέσεις μεταξύ των ονομάτων οντοτήτων. Κάθε παράδειγμα εκπαίδευσης αποτελείται από ένα σύνολο συμβόλων (λέξεις) που βρίσκονται μεταξύ των δυο συσχετισμένων ονομάτων οντοτήτων $x$, $y$ (συμπεριλαμβανομένων των σημείων στίξης), ενώ είναι χαρακτηρισμένο από τον τύπο σχέσης $Rel(x, y)$. Εάν ένα οποιοδήποτε όνομα οντότητας $w$ περιέχεται μέσα στα σύμβολα ενός παραδείγματος εκπαίδευσης, τότε όλα τα σύμβολα που απαρτίζουν το όνομα της οντότητας $w$ αντικαθιστώνται από την κατηγορίας της οντότητας (π.χ. αν βρεθεί το όνομα οντότητας «Ηνωμένες Πολιτείες», αντικαθίσταται με την κατηγορία «Τοποθεσία»), καθώς είναι επιθυμητή η καταγραφή της πληροφορίας μεταξύ οντοτήτων και όχι η γλωσσική δομή των ίδιων των οντοτήτων, το οποίο είναι το έργο ενός αναγνωριστή ονομάτων οντοτήτων.

Στη συνέχεια, από το σύνολο των παραδειγμάτων εκπαίδευσης, εξάγεται ένα σύνολο από γραμματικές ανεξάρτητες από τα συμφραζόμενα, όπου μία γραμματική αντιστοιχεί σε ένα τύπο (σημασιολογική κατηγορία) σχέσης. Το αποτελέσματα της φάσης εκπαίδευσης είναι ένα σύνολο γραμματικών ανεξάρτητων από συμφραζόμενα, μία για κάθε τύπο σχέσης που πρέπει να εξαχθεί. Κάθε γραμματική ανεξάρτητη από συμφραζόμενα μετατρέπεται σε έναν λογικό *ταξινομητή (classifier)*, με την βοήθεια της βιβλιοθήκης Boost.Xpressive, η οποία αποτελεί μια βιβλιοθήκη προτύπων για την γλώσσα προγραμματισμού C++ (C++ template library) [142]. Ένας τέτοιος ταξινομητής επιστρέφει «αληθές» αν το περιεχόμενο (σύμβολα) μεταξύ δυο ονομάτων οντοτήτων μπορεί να αναλυθεί από τη γραμματική, ενώ επιστρέφει «ψευδές» εάν δεν μπορεί να αναλυθεί από την γραμματική.

### 6.3.2 Ο αλγόριθμος egGRIDS+: η τάση προς απλές γραμματικές

Ο αλγόριθμος egGRIDS+ εξάγει γραμματικές ανεξάρτητες από συμφραζόμενα αποκλειστικά από θετικά παραδείγματα. Χρησιμοποιώντας ένα περιορισμένο σύνολο τελεστών γενίκευσης, ο egGRIDS+ ακολουθεί μια επαναληπτική προσέγγιση με στόχο τη γενίκευση μιας αρχικής «επίπεδης» γραμματικής που εξάγεται από τα (θετικά) παραδείγματα εκπαίδευσης. Σε κάθε επανάληψη, οι υποψήφιες γραμματικές

βαθμολογούνται σύμφωνα με το ευριστικό $MDL$, ενώ ο χώρος των πιθανών γραμματικών μπορεί να ερευνηθεί με διάφορες στρατηγικές αναζήτησης (όπως η αναζήτηση δέσμης ή η γενετική εξέλιξη) και ευριστικά, τα οποία προσπαθούν να μειώσουν τον χρόνο εκπαίδευσης μέσω του εντοπισμού συγκεκριμένων γραμματικών δομών.

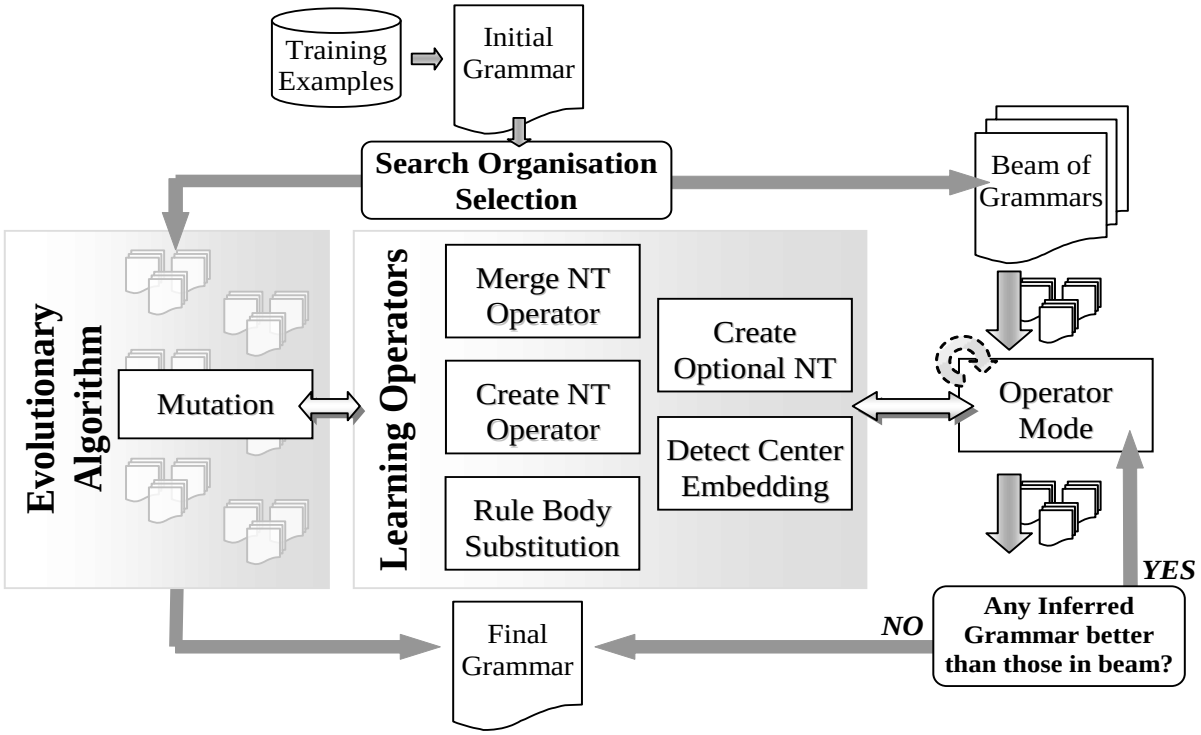
Καθώς ο egGRIDS+ δεν χρησιμοποιεί αρνητική πληροφόρηση, χρειάζεται ένα επιπρόσθετο κριτήριο για να κατευθύνει την έρευνα στον χώρο των πιθανών γραμματικών και να αποφευχθούν οι υπερβολικά γενικές γραμματικές. Το ευριστικό του ελάχιστου μήκους περιγραφής ($MDL$) έχει υιοθετηθεί στον egGRIDS+, που κατευθύνει την έρευνα προς «συμπαγείς» γραμματικές, δηλ. εκείνες που η κωδικοποίησή τους απαιτεί λίγα *δυφία (bits)*, ενώ ταυτόχρονα κωδικοποιούν και το σύνολο των παραδειγμάτων με ένα «συμπαγή» τρόπο, δηλ. απαιτούνται λίγα δυφία για την κωδικοποίηση των παραδείγματα χρησιμοποιώντας την γραμματική. Υποθέτοντας μια γραμματική $G$ και ένα σύνολο παραδειγμάτων (προτάσεις) $T$, τα οποία μπορούν να αναγνωριστούν (αναλυθούν) από την γραμματική $G$, το συνολικό μήκος περιγραφής της γραμματικής (εφεξής $ML$), είναι το άθροισμα δυο ανεξάρτητων μηκών:

- Το *μήκος περιγραφής της γραμματικής (grammar description length – $GDL$)*, δηλ. τα δυφία που απαιτούνται για να κωδικοποιήσουν τους γραμματικούς κανόνες και να τους διαβιβάσουν σε έναν παραλήπτη που έχει ελαχίστη γνώση της γραμματικής αναπαράστασης, και
- Το *μήκος περιγραφής των παραγωγών (derivations description length – $DDL$)*, δηλ. τα δυφία που απαιτούνται για να κωδικοποιηθούν και να μεταβιβαστούν όλα τα παραδείγματα του συνόλου $T$, όπως κωδικοποιούνται από την γραμματική $G$, με την προϋπόθεση ότι ο παραλήπτης ήδη γνωρίζει την $G$.

Το πρώτο συστατικό του $ML$ κατευθύνει την έρευνα μακριά από την απλοϊκή γραμματική που έχει ένα ξεχωριστό κανόνα για κάθε παράδειγμα εκπαίδευσης, καθώς αυτή η γραμματική θα έχει ένα μεγάλο $GDL$. Όμως, το ίδιο συστατικό οδηγεί σε άλλο είδος απλοϊκής γραμματικής, μια γραμματική που αποδέχεται όλα τα παραδείγματα (π.χ. η πιο γενική γραμματική *"*$S \rightarrow S\ T;\ T \rightarrow (οποιαδήποτε\ λέξη \mid e)$*"*). Για να αποφευχθεί αυτό, το δεύτερο συστατικό εκτιμά την *δύναμη παραγωγής* της γραμματικής, μετρώντας τον τρόπο που τα παραδείγματα εκπαίδευσης προκύπτουν από την γραμματική, και βοηθά στην αποφυγή της υπεργενίκευσης, τιμωρώντας τις γενικές γραμματικές.

### 6.3.3 Αρχιτεκτονική και τελεστές μάθησης

Η αρχιτεκτονική του egGRIDS+ συνοψίζεται στην Εικόνα 39. Ο egGRIDS+ χρησιμοποιεί τα δεδομένα εκπαίδευσης προκειμένου να κατασκευάσει μια αρχική, «επίπεδη» γραμματική. Αυτή η αρχική γραμματική κατασκευάζεται με την απλή μετατροπή κάθε ενός από τα παραδείγματα εκπαίδευσης σε ένα γραμματικό κανόνα. Αυτή η αρχική γραμματική είναι υπερβολικά συγκεκριμένη, καθώς μπορεί να αναγνωρίσει μόνο τις προτάσεις που περιέχονται στο σύνολο εκπαίδευσης. Στην συνέχεια, ο egGRIDS+ γενικεύει αυτή την αρχική γραμματική, χρησιμοποιώντας κάποια από τις διαθέσιμες στρατηγικές αναζήτησης, όπως αναζήτηση δέσμης, γενετική αναζήτηση, ευριστική αναζήτηση, κλπ. Όλες οι στρατηγικές αναζήτησης χρησιμοποιούν τους ίδιους τελεστές αναζήτησης ώστε να παραχθούν περισσότερο γενικές γραμματικές.




**Εικόνα 39: Η αρχιτεκτονική του egGRIDS+.**

Προς το παρών, ο egGRIDS+ υποστηρίζει πέντε τελεστές αναζήτησης:

- **Merge NT:** συγχωνεύει δυο μη-τερματικά σύμβολα σε ένα νέο σύμβολο.
- **Create NT:** δημιουργεί ένα νέο μη-τερματικό σύμβολο $X$, το οποίο καθορίζεται ως μια ακολουθία από δυο ή περισσότερα μη-τερματικά σύμβολα.
- **Create Optional NT:** επαναλαμβάνει ένα κανόνα που δημιουργήθηκε από τον τελεστή «CreateNT**»** και επισυνάπτει ένα μη-τελικό σύμβολο στο τέλος του σώματος του κανόνα, κάνοντας αυτό το σύμβολο προαιρετικό.
- **Detect Center Embedding:** στοχεύει στην σύλληψη του φαινόμενου του center embedding. Αυτός ο τελεστής προσπαθεί να εντοπίσει το πιο συχνό τετρά-γραμμα της μορφής *"$A\ A\ B\ B$"*. Μόλις αυτό εντοπιστεί, ο τελεστής δημιουργεί ένα νέο μη-τελικό σύμβολο $X$, όπως θα έκανε και ο τελεστής «CreateNT**»**. Όμως, υποθέτοντας ότι αυτό το τετρά-γραμμα δημιουργήθηκε μέσω center embedding που περιλαμβάνει το σύμβολο $X$, αυτός ο τελεστής θα δημιουργήσει έναν ακόμα κανόνα της μορφής *"$X \rightarrow A\ A\ X\ B\ B$"*, ενώ θα αντικαταστήσει όλες τις ακολουθίες σύμβολων που ταιριάζουν στο μοτίβο *"$A+\ X?\ B+$"* με το $X$.
- **Rule Body Substitution:** εξετάζει αν το σώμα ενός κανόνα $R$ περιέχεται στα σώματα άλλων κανόνων. Σε αυτή την περίπτωση, κάθε εμφάνιση του σώματος του κανόνα $R$ σε αλλά σώματα κανόνων αντικαθιστώνται από την κεφαλή του κανόνα $R$.

Οι πέντε τελεστές δημιουργούν γραμματικές που έχουν είτε την ιδία ή μεγαλύτερη εκφραστικότητα από την μητρική γραμματική. Καθώς οι τελεστές ποτέ δεν αφαιρούν κανόνες από μια γραμματική, η τελική γραμματική έχει τουλάχιστον την ιδία κάλυψη με την αρχική γραμματική, δηλ. μπορούν να αναγνωρίσουν τουλάχιστον το ίδιο σύνολο προτάσεων.

## 6.4 Τα σώματα κειμένων (δεδομένα εκπαίδευσης)

Για τους σκοπούς της αξιολόγησης της προτεινόμενης προσέγγισης στην εργασία της εξαγωγής συσχετίσεων μεταξύ ονομάτων οντοτήτων, χρησιμοποιήθηκε ένα επισημειωμένο σώμα κειμένων, το οποίο ήταν αποτέλεσμα ενός ευρωπαϊκού ερευνητικού έργου, χρηματοδοτούμενου από την ΕΕ, γνωστού με την επωνυμία BOEMIE[6]. Το σώμα κειμένων περιέχει 800 HTML σελίδες στην Αγγλική γλώσσα, οι οποίες ανακτήθηκαν από επίσημους ιστότοπους διαφόρων αθλητικών οργανώσεων όπως οι IAAF[7], EAA[8] και η USATF[9]. Η θεματική περιοχή των κειμένων αφορά νέα σχετικά με τα αθλήματα του στίβου, αποτελέσματα από διεθνείς διοργανώσεις σχετικές με στίβο, καθώς και βιογραφίες αθλητών στίβου.

Όλες οι σελίδες έχουν επισημειωθεί χειρωνακτικά, σύμφωνα με ένα σημασιολογικό μοντέλο που περιλαμβάνει πληροφορίες σχετικές με τα αθλήματα, τις διοργανώσεις, τις επιδόσεις, του αθλητές και τις συμμέτοχες τους σε αθλητικούς διαγωνισμούς, σχετικούς με το *στίβο (athletics)*. Αυτό το σημασιολογικό μοντέλο αποτέλεσε την βάση, από την οποία εξήχθηκαν τα ονόματα οντοτήτων τις συγκεκριμένης περιοχής καθώς και οι συσχετίσεις μεταξύ των ονομάτων οντοτήτων, που χρησιμοποιήθηκαν για την αξιολόγηση της προτεινόμενης προσέγγισης.

## 6.5 Πειραματική αξιολόγηση και αποτελέσματα

Η πειραματική αξιολόγηση περιλαμβάνει την εξέταση του αλγορίθμου egGRIDS+ στην εργασία της αναγνώρισης συσχετίσεων μεταξύ ονομάτων οντοτήτων, σε κείμενα της Αγγλικής γλώσσας με θεματική περιοχή αθλητικά νέα και βιογραφικά αθλητών από αθλήματα σχετικά με τον στίβο. Η αξιολόγηση αυτή έχει διπλό στόχο:

- Να εξετάσει την εφαρμοσιμότητα του νέου αλγορίθμου egGRIDS+ σε πραγματικά δεδομένα, και την *επεκτασιμότητα (scalability)* του αλγορίθμου στα μεγέθη παραδειγμάτων εκπαίδευσης που απαιτούνται για την εφαρμογή του egGRIDS+ σε πρακτικά προβλήματα.
- Να εξετάσει την απόδοση του egGRIDS+ στην εργασία της εξαγωγής σχέσεων.

Για την εκπλήρωση του δεύτερου στόχου, η απόδοση του egGRIDS+ συγκρίνεται με έναν ακόμα αλγόριθμο μηχανικής μάθησης στα ίδια δεδομένα, αλλά και με άλλες προσεγγίσεις που αναφέρονται στην διεθνή βιβλιογραφία για διαφορετικά σώματα κειμένων.

### 6.5.1 Δημιουργία παραδειγμάτων εκπαίδευσης/αποτίμησης

Το σώμα κειμένων που παραχωρήθηκε από το ερευνητικό έργο BOEMIE είναι επισημειωμένο με ένα σημασιολογικό μοντέλο, οι οντότητες του οποίου πρέπει να εντοπιστούν στα κείμενα, ώστε να είναι δυνατή η αξιοποίησή τους για τις ανάγκες της πειραματικής αξιολόγησης. Για τον σκοπό αυτό κατασκευάστηκε έναν αναγνωριστής ονομάτων οντοτήτων, βασισμένος σε μηχανική μάθηση, και συγκεκριμένα στον αλγόριθμο των *Conditional Random Fields* [143]. Ο αναγνωριστής ονομάτων οντοτήτων που κατασκευάστηκε αξιολογήθηκε, εμφανίζοντας *ακρίβεια (precision)* 90% και *ανάκτηση (recall)* η οποία προσεγγίζει το 86%. Αυτό το σύστημα αναγνώρισης

[6] BOEMIE: http://www.boemie.org/.

[7] International Association of Athletics Federations – http://www.iaaf.org/.

[8] European Athletics Association – http://www.european-athletics.org/.

[9] USA Track and Field – http://www.usatf.org/.

ονομάτων οντοτήτων εφαρμόστηκε στο σώμα κειμένων, το οποίο επισημειώθηκε με ονόματα οντοτήτων. Στην συνέχεια τα ονόματα των αναγνωρισθέντων οντοτήτων ταιριάστηκαν (χρησιμοποιώντας *διαδικασίες ταιριάσματος (matching techniques)*) με τις οντότητες του σημασιολογικού μοντέλου. Έχοντας μια ευθυγράμμιση μεταξύ αναγνωρισμένων ονομάτων οντοτήτων στα κείμενα, και των οντοτήτων του σημασιολογικού μοντέλου, οι σχέσεις μεταξύ οντοτήτων από το σημασιολογικό μοντέλο προβλήθηκαν στο σώμα κειμένων, οδηγώντας σε μια αρχική επισημείωση των δυαδικών σχέσεων μεταξύ των αναγνωρισμένων ονομάτων οντοτήτων στα κείμενα. Ως επόμενο βήμα στην προετοιμασία των δεδομένων, οι σχέσεις που αφορούν ονόματα πρόσωπων και ιδιότητες πρόσωπων όπως το φύλο, η ηλικία, η εθνικότητα, η επίδοση και κατάταξη του αθλητή, επαληθευτήκαν χειρωνακτικά, και διορθωθήκαν όπου χρειάστηκε, ώστε να δημιουργηθεί ένα αξιόπιστο σύνολο δεδομένων εκπαίδευσης/αξιολόγησης.

Η αξιολόγηση περιορίστηκε σε σχέσεις που συμβαίνουν μέσα στα όρια προτάσεων, προκειμένου να κρατηθεί η πολυπλοκότητα των γραμματικών που θα παραχθούν από την διαδικασία της εκπαίδευσης, αλλά και ο απαιτούμενος χρόνος για αυτό, σε χαμηλά επίπεδα. Αυτός είναι και ο κύριος λόγος που επιλέχθηκαν μόνο σχέσεις που αφορούν ονόματα και ιδιότητες σχετικά με αθλητές, καθώς η τεράστια πλειοψηφία τους δεν ξεπερνά τα όρια των προτάσεων. Αντίθετα σχέσεις που αφορούν αθλητές και αθλητικές διοργανώσεις ή αθλητικές εκδηλώσεις εκτείνονται συχνά εκτός των ορίων των προτάσεων, η αξιολόγηση των οποίων θα ήταν δύσκολη στο πειραματικό πλαίσιο που έχουμε ορίσει. Έτσι, σαν τελικό βήμα, οι σχέσεις που ξεπερνούν τα όρια των προτάσεων αφαιρέθηκαν από το σώμα κειμένων, καταλήγοντας σε ένα σώμα κειμένων που περιέχει 8.497 σχέσεις μεταξύ ονομάτων πρόσωπων και διάφορες ιδιότητες των προσώπων αυτών.

Από αυτό το σώμα κειμένων, δημιουργήθηκαν 8.497 παραδείγματα εκπαίδευσης. Για να μειωθεί η *διάσπαση (sparseness)* των δεδομένων, χρησιμοποιήθηκαν τα *θέματα (stems)* των λέξεων αντί για τις ίδιες τις λέξεις. Κάθε παράδειγμα εκπαίδευσης περιέχει όλα τα θέματα των λέξεων, καθώς και τα σημεία στίξης, που βρέθηκαν στο κείμενο, μεταξύ των δυο συσχετιζόμενων ονομάτων οντοτήτων, με τη σειρά που εμφανιστήκαν στο κείμενο. Κάθε όνομα οντότητας που βρέθηκε εντός του παραδείγματος εκπαίδευσης, αντικαταστάθηκε από τον τύπο της οντότητας, ενώ κάθε παράδειγμα χαρακτηρίστηκε από τον τύπο της σχέσης. Ένα παράδειγμα πρότασης, επισημειωμένης με ονόματα οντοτήτων, φαίνεται στην Εικόνα 40, ενώ τα παραγόμενα παραδείγματα εκπαίδευσης φαίνονται στην Εικόνα 41.

**Kenya**=[country]'s **Richard Limo**=[name] the World **5000m**=[sport_name] champion (eventual **third**=[ranking] **26:50.20**=[performance]) came the nearest during the first 300m of the lap, until in the finishing straight, **Ethiopia**=[country]'s Olympic **bronze**=[ranking] **Assefa Mezegebu**=[name] started a drive to the line which took **second**=[ranking] place (**26:49.90**=[performance]).

**Εικόνα 40: Παράδειγμα πρότασης επισημειωμένης με ονόματα οντοτήτων.**

Για την αποκόμιση μιας ρεαλιστικής και αμερόληπτης εκτίμησης των επιδόσεων της μεθόδου, χρησιμοποιήθηκε *δεκαπλή διασταυρωμένη επικύρωση (10-fold cross validation)* (ενότητα 2.3). Σύμφωνα με αυτή τη μέθοδο αξιολόγησης, το σώμα κειμένων χωρίζεται σε δέκα, ίσου μεγέθους υποσώματα, με το τελικό αποτέλεσμα να είναι ο μέσος όρος της επίδοσης στα δέκα πειράματα. Σε κάθε πείραμα, εννέα από τα δέκα

επιμέρους υποσώματα κειμένων χρησιμοποιούνται για την εκπαίδευση του αναγνωριστή συσχετίσεων μεταξύ ονομάτων οντοτήτων, ενώ το δέκατο χρησιμοποιείται για την αποτίμηση της απόδοσης (αξιολόγηση). Η απόδοση της προσέγγισης αποτιμήθηκε με όρους *ακριβείας (precision)*, *ανάκλησης (recall)* και *F-measure*. Κατά την διαδικασία της αποτίμησης, κάθε παράδειγμα του συνόλου αξιολόγησης αναλύθηκε από όλες τις παραχθείσες γραμματικές: αν το παράδειγμα αναλύθηκε σωστά *μόνο από την γραμματική* που αντιστοιχεί στο σωστό τύπο σχέσης, το παράδειγμα θεωρείται σωστό. Σε όλες τις άλλες περιπτώσεις, συμπεριλαμβανομένης και της περίπτωσης όπου ένα παράδειγμα αναλύθηκε από περισσότερες από μια εξαχθείσα γραμματική, το παράδειγμα θεωρείται λανθασμένο. Τα αποτελέσματα της αξιολόγησης παρουσιάζονται στον πίνακα: Πίνακας 44.

| Word stems | Relation label |
|---|---|
| ' s | name-country |
| the world entity:sport_name champion ( eventual | name-ranking |
| the world entity:sport_name champion ( eventual entity:ranking | name-performance |
| ' s | name-country |
| start a drive to the line which take | name-ranking |
| start a drive to the line which take entity:ranking place ( | name-performance |

**Εικόνα 41: Παραδείγματα εκπαίδευσης που εξήχθησαν από την πρόταση του παραδείγματος (Εικόνα 40).**

| | Ακρίβεια | Ανάκληση | F-measure |
|---|---|---|---|
| **Name-Ranking** | 95.05 % | 54.07 % | 68.57 % |
| **Name-Performance** | 92.14 % | 49.26 % | 64.17 % |
| **Name-Country** | 98.85 % | 88.88 % | 93.58 % |
| **Name-Gender** | 99.21 % | 79.17 % | 88.00 % |
| **Name-Age** | 100.00 % | 98.11 % | 99.04 % |
| **Overall** | **96.48 %** | **65.96 %** | **78.32 %** |

**Πίνακας 44: Τα αποτελέσματα της αξιολόγησης του egGRIDS+ στην εργασία της αναγνώρισης συσχετίσεων μεταξύ ονομάτων οντοτήτων.**

Τα αποτελέσματα της αξιολόγησης δείχνουν ότι η προτεινόμενη προσέγγιση αποδίδει καλά σε σύγκριση με προσεγγίσεις της διεθνούς βιβλιογραφίας, παρά τη δυσκολία σύγκρισης των αποτελεσμάτων που αποκτήθηκαν σε διάφορα σώματα. Για παράδειγμα, στην εργασία [136], η προτεινόμενη προσέγγιση, η οποία γενίκευσε ξεκινώντας από 55 χειρωνακτικά κατασκευασμένους κανόνες, παρουσιάζει μια απόδοση γύρω στο 88% (ακρίβεια) και 43% (ανάκληση) σε 1032 άρθρα νέων σχετικά με τα βραβεία Νόμπελ, από ειδησεογραφικούς οργανισμούς, όπως New York Times, BBC και CNN.

Το γεγονός ότι η προσέγγιση μας χρησιμοποιεί σαν είσοδο μόνο θέματα λέξεων έχει δυο ενδιαφέρουσες επιπτώσεις:

a) Αν ένα παράδειγμα αξιολόγησης περιέχει ένα θέμα άγνωστο στην εξαχθείσα γραμματική, αυτό το παράδειγμα θα χαρακτηριστεί σαν αποτυχία, αφού δεν θα μπορεί να αναλυθεί από καμία εξαχθείσα γραμματική, και
b) οποιαδήποτε γενίκευση πρέπει να αποδοθεί στην επιτυχή λειτουργία του egGRIDS+, ο οποίος έμαθε τις σωστές συντακτικές δομές, προκειμένου να επιτραπεί η χρήση «παρομοίων» θεμάτων αντί για ένα συγκεκριμένο στέλεχος.

Μια εύκολη λύση στο (a), η οποία ακολουθείται από πολλές προσεγγίσεις (π.χ. [140]) στη βιβλιογραφία, είναι να προστεθεί άλλο ένα αφαιρετικό επίπεδο στις λέξεις, όπως ετικέτες σχετικές με το μέρος του λόγου. Το γεγονός ότι η παρουσιαζόμενη προσέγγιση δεν κάνει χρήση ενός τέτοιου αφαιρετικού επιπέδου, μας επιτρέπει να αποκτήσουμε μια εκτίμηση της γενίκευσης που επιτυγχάνεται αποκλειστικά από τον αλγόριθμο γραμματικού συμπερασμού που χρησιμοποιείται. Για αυτό το λόγο, το ίδιο πείραμα επαναλήφτηκε με μια μικρή αλλαγή: οι δίπλες καταχωρήσεις αφαιρέθηκαν από το σύνολο των παραδειγμάτων εκπαίδευσης, μετατρέποντας όλα τα παραδείγματα εκπαίδευσης σε μοναδικά. Αυτό μείωσε το σύνολο των παραδειγμάτων εκπαίδευσης σχεδόν κατά 2/3, αλλά εξασφαλίζει ότι όλα τα παραδείγματα που θα χρησιμοποιηθούν για την αποτίμηση, δεν θα έχουν εμφανιστεί ποτέ στο σύνολο των δεδομένων εκπαίδευσης. Χρησιμοποιήθηκε επίσης δεκαπλή διασταυρωμένη επικύρωση, ενώ τα αποτελέσματα αξιολόγησης εμφανίζονται στον πίνακα: Πίνακας 45.

Παρά το γεγονός ότι τα αποτελέσματα του Πίνακας 45 είναι μια απαισιόδοξη προσέγγιση (αφού τα παραδείγματα αξιολόγησης που περιέχουν άγνωστες λέξεις σε σχέση με τα παραδείγματα εκπαίδευσης δεν έχουν αφαιρεθεί), ο egGRIDS+ κατάφερε να πετύχει μια γενίκευση περίπου 10 μονάδων σε όρους ανάκλησης, το οποίο είναι σημαντικό αν αναλογιστεί κανείς ότι τα παραδείγματα αξιολόγησης περιλαμβάνουν χρήση λέξεων με σειρά που δεν έχει παρατηρηθεί στην διάρκεια της εκπαίδευσης, ακόμα και αν η απώλεια στην ακρίβεια προσεγγίζει τις 29 μονάδες.

Όσον αφόρα τον χρόνο εκτέλεσης κατά την διαδικασία εκπαίδευσης, ο αλγόριθμος egGRIDS+ είναι σε θέση να συγκλίνει σε μια τελική γραμματική μέσα σε λίγα λεπτά (από 5 ως 15 λεπτά στις περισσότερες περιπτώσεις), όταν η εκπαίδευση πραγματοποιείται στο σύνολο των παραδειγμάτων εκπαίδευσης του πρώτου πειράματος (8.497 παραδείγματα εκπαίδευσης). Ωστόσο, η μετατροπή της εξαχθείσας ανεξάρτητης από συμφραζόμενα γραμματικής σε έναν κατηγοριοποιητή, (χρησιμοποιώντας την C++ βιβλιοθήκη Boost.Xpressive) απαιτούσε σημαντικά χρονικά διαστήματα στο στάδιο της μεταγλώττισης[10]: συνήθως από 45 λεπτά μέχρι λίγο περισσότερα από 60 λεπτά για κάθε γραμματική.

[10] Το πείραμα πραγματοποιήθηκε σε έναν προσωπικό υπολογιστή με λειτουργικό σύστημα Windows Vista (64bit), με επεξεργαστή Intel 6700 και 4 GB RAM. Ο μεταγλωττιστής που χρησιμοποιήθηκε ήταν ο VC++ 2005 της Microsoft.

| | Ακρίβεια | Ανάκληση | F-measure |
|---|---|---|---|
| **Name-Ranking** | 50.04 % | 6.79 % | 11.90 % |
| **Name-Performance** | 67.16 % | 11.87 % | 20.13 % |
| **Name-Country** | 100.00 % | 16.05 % | 27.20 % |
| **Name-Gender** | 74.83 % | 7.04 % | 12.73 % |
| **Name-Age** | 80.00 % | 47.12 % | 55.00 % |
| **Overall** | **67.58 %** | **10.46 %** | **18.09 %** |

**Πίνακας 45: Τα αποτελέσματα της αξιολόγησης του egGRIDS+ στην εργασία της αναγνώρισης συσχετίσεων μεταξύ ονομάτων οντοτήτων (μοναδικά παραδείγματα εκπαίδευσης).**

### 6.5.2 Συγκριτική αξιολόγηση με έτερο αλγόριθμο μηχανικής μάθησης

Η προσέγγιση για την αναγνώριση συσχετίσεων μεταξύ ονομάτων οντοτήτων που παρουσιάστηκε στις προηγούμενες ενότητες, αξιολογήθηκε εμφανίζοντας μια απόδοση μεγαλύτερη από αντίστοιχα συστήματα της διεθνούς βιβλιογραφίας, σε διαφορετικά σώματα κειμένων. Στην ενότητα αυτή θα εφαρμοστεί μια διαφορετική προσέγγιση, βασισμένη σε μηχανική μάθηση, η οποία θα αποτιμηθεί στο ίδιο σώμα κειμένων όπως και η βασισμένη στον egGRIDS+ προσέγγιση. Κίνητρο αποτελεί η συγκριτική αξιολόγηση της προτεινόμενης προσέγγισης με έναν διαφορετικό αλγόριθμο μηχανικής μάθησης, στα ίδια δεδομένα.

Για τους σκοπούς της συγκριτικής αξιολόγησης χρησιμοποιήθηκαν τα *Conditional Random Fields*[11] [143], ένας στοχαστικός αλγόριθμος ο οποίος έχει αποδειχθεί ιδιαίτερα αποδοτικός στην κατασκευή πιθανοτικών μοντέλων για την κατηγοριοποίηση σειραϊκών δεδομένων. Ο εναλλακτικός αυτός αναγνωριστής λειτουργεί επίσης σε επίπεδο προτάσεων, χρησιμοποιώντας μια αναπαράσταση όσο το δυνατόν πλησιέστερη εκείνης των προηγούμενων ενοτήτων. Για κάθε πρόταση, δημιουργείται ένα διάνυσμα χαρακτηριστικών το οποίο περιέχει τα θέματα όλων των λέξεων, καθώς και τα σημεία στίξης, της πρότασης, με τη σειρά που εμφανιστήκαν στο κείμενο. Κάθε όνομα οντότητας που βρέθηκε εντός του διανύσματος, αντικαταστάθηκε από τον τύπο της οντότητας. Ταυτόχρονα, σε κάθε όνομα οντότητας αποδόθηκε ένα μοναδικό χαρακτηριστικό, ένας προσδιοριστής: το ζητούμενο από την αλγόριθμο μηχανικής μάθησης είναι να συσχετίσει τους προσδιοριστές μεταξύ τους, ώστε να τους συσχετίσει. Υποθέτοντας την ακόλουθη πρόταση:

**Kenya**=[country]'s **Paul Korir**=[name] **won**=[ranking] the **Emsley Carr Mile**=[event] in the fastest time in the world this year – **3:49.84**=[performance] – pulling **Ivan Heshko**=[name] to a new **Ukrainian**=[nationality] record of **3:50.04**=[performance] in **second**=[ranking].

**Εικόνα 42: Παράδειγμα πρότασης επισημειωμένης με ονόματα οντοτήτων.**

[11] Για τις ανάγκες της συγκριτικής αξιολόγησης χρησιμοποιήθηκε η υλοποίηση CRF++ του Taku Kudo, η οποία μπορεί να βρεθεί στον ακόλουθο σύνδεσμο: http://crfpp.sourceforge.net/.

Προκύπτει το ακόλουθο διάνυσμα χαρακτηριστικών (εκπαίδευσης):

| Διάνυσμα χαρακτηριστικών | | Κατηγορία |
|---|---|---|
| **:nationality:** | :no: | NR1 |
| **'s** | :no: | :no: |
| **:name:** | NR1 | :no: |
| **:ranking:** | :no: | NR1 |
| **the** | :no: | :no: |
| **:event_name:** | ER1 | :no: |
| **in** | :no: | :no: |
| **the** | :no: | :no: |
| **fastest** | :no: | :no: |
| **time** | :no: | :no: |
| **in** | :no: | :no: |
| **the** | :no: | :no: |
| **world** | :no: | :no: |
| **this** | :no: | :no: |
| **year** | :no: | :no: |
| **-** | :no: | :no: |
| **:performance:** | :no: | NR1 |
| **-** | :no: | :no: |
| **pull** | :no: | :no: |
| **:name:** | NR2 | :no: |
| **to** | :no: | :no: |
| **a** | :no: | :no: |
| **new** | :no: | :no: |
| **:nationality:** | :no: | NR2 |
| **record** | :no: | :no: |
| **of** | :no: | :no: |
| **:performance:** | :no: | NR2 |
| **in** | :no: | :no: |
| **:ranking:** | :no: | NR2 |

**Εικόνα 43: Παράδειγμα διανύσματος εκπαίδευσης.**

Όπως είναι εμφανές και από την Εικόνα 43, στην πρόταση (Εικόνα 42) εμφανίζονται έξι συσχετίσεις μεταξύ ονομάτων οντοτήτων:

- Τρεις συσχετίσεις σχετικές με το όνομα οντότητας *NR1*, των τύπων "Name-Country", "Name-Ranking", και "Name-Performance".
- Τρεις συσχετίσεις σχετικές με το όνομα οντότητας *NR2*, επίσης των τύπων "Name-Country", "Name-Ranking", και "Name-Performance".

Τα αποτελέσματα της αξιολόγησης της προσέγγισης που βασίζεται στα *Conditional Random Fields (CRF++)* εμφανίζονται στον πίνακα: Πίνακας 46. Στον ίδιο πίνακα

εμφανίζονται τα αποτελέσματα της προτεινόμενης προσέγγισης που βασίζεται στον egGRIDS+, τα οποία προέρχονται από τον σχετικό πίνακα αξιολόγησης (Πίνακας 44). Για την πειραματική αξιολόγηση χρησιμοποιήθηκε *πενταπλή διασταυρωμένη επικύρωση (5-fold cross validation)*.

| | CRF++ | | | egGRIDS+ | | |
|---|---|---|---|---|---|---|
| | **Ακρίβεια** | **Ανάκληση** | **F-measure** | **Ακρίβεια** | **Ανάκληση** | **F-measure** |
| **Name-Ranking** | 77.40 % | 60.47 % | 67.80 % | 95.05 % | 54.07 % | 68.57 % |
| **Name-Performance** | 84.42 % | 84.18 % | 84.93 % | 92.14 % | 49.26 % | 64.17 % |
| **Name-Country** | 88.78 % | 84.63 % | 86.70 % | 98.85 % | 88.88 % | 93.58 % |
| **Name-Gender** | 65.22 % | 36.78 % | 42.43 % | 99.21 % | 79.17 % | 88.00 % |
| **Name-Age** | 79.88 % | 56.03 % | 64.28 % | 100.00 % | 98.11 % | 99.04 % |
| **Overall** | **79.88 %** | **60.47 %** | **67.80 %** | **96.48 %** | **65.96 %** | **78.32 %** |

**Πίνακας 46: Τα αποτελέσματα της αξιολόγησης της προσέγγισης που βασίζεται στον CRF++, στην εργασία της αναγνώρισης συσχετίσεων μεταξύ ονομάτων οντοτήτων.**

Τα αποτελέσματα της αξιολόγησης δείχνουν ότι η προσέγγιση που βασίζεται στον egGRIDS+ αποδίδει καλύτερα από την προσέγγιση που βασίζεται στον αλγόριθμο *CRF++*. Ο λόγος για αυτή την υστέρηση του *CRF++* μπορεί να αποδοθεί στους ακόλουθους λόγους:

- Το ιστορικό της υλοποίησης *CRF++* μπορεί να φτάσει μέχρι τις προηγούμενες τέσσερις λέξεις, το οποίο είναι πολύ μικρό για κάποιες συσχετίσεις, οι οντότητες των οποίων μπορεί να απέχουν περισσότερο από τέσσερις λέξεις.
- Ίσως η απόδοση στις εμπλεκόμενες οντότητες του προσδιοριστή μιας άλλης οντότητας να ήταν μια σύνθετη απαίτηση για τον συγκεκριμένο αλγόριθμο.

Για το λόγο αυτό, δοκιμάστηκε και μια δεύτερη εναλλακτική προσέγγιση, χρησιμοποιώντας πάλι τον αλγόριθμο *CRF++*, αλλά αυτή την φορά με χρησιμοποίηση διαφορετικής αναπαράστασης. Από την στιγμή που το ιστορικό του αλγορίθμου δεν μπορεί να υπερβεί τις τέσσερις λέξεις, δημιουργήθηκε ένα διάνυσμα χαρακτηριστικών για κάθε πιθανή συσχέτιση εντός κάθε πρότασης, αντί για ένα διάνυσμα για κάθε λέξη της πρότασης. Το ζητούμενο πλέον από τον αλγόριθμο *CRF++* είναι η κατηγοριοποίηση το αν το διάνυσμα είναι συσχέτιση ή όχι. Κάθε διάνυσμα απαρτίζεται από ενεά χαρακτηριστικά, τα οποία επιλέχθηκαν με βάση την μεγιστοποίηση της απόδοσης, και τα οποία παρουσιάζονται στην ακόλουθη λίστα:

1. Ο τύπος της υποψήφιας σχέσης, π.χ. Name-Ranking.
2. Η τιμή “normal” αν το υποκείμενο της υποψήφιας σχέσης εμφανίζεται στην πρόταση πριν το αντικείμενο, διαφορετικά η τιμή “reverse”.
3. Ο συνολικός αριθμός από λέξεις που παρεμβάλλονται μεταξύ του υποκειμένου και του αντικειμένου της υποψήφιας σχέσης.
4. Ο αριθμός από ουσιαστικά που υπάρχουν μεταξύ του υποκειμένου και του αντικειμένου της υποψήφιας σχέσης.
5. Ο αριθμός από ρήματα που υπάρχουν μεταξύ του υποκειμένου και του αντικειμένου της υποψήφιας σχέσης.
6. Ο αριθμός από επιρρήματα που υπάρχουν μεταξύ του υποκειμένου και του αντικειμένου της υποψήφιας σχέσης.

7. Ο αριθμός από επίθετα που υπάρχουν μεταξύ του υποκειμένου και του αντικειμένου της υποψήφιας σχέσης.
8. Ο αριθμός από αντωνυμίες που υπάρχουν μεταξύ του υποκειμένου και του αντικειμένου της υποψήφιας σχέσης.
9. Ο αριθμός από λέξεις που υπάρχουν μεταξύ του υποκειμένου και του αντικειμένου της υποψήφιας σχέσης τα οποία δεν ανήκουν στις περιπτώσεις των χαρακτηριστικών 4 έως 8.

Αξίζει να σημειωθεί ότι δεν είναι δυνατόν να περιληφθούν τα θέματα των λέξεων σε αυτή την αναπαράσταση, η έλλειψη των οποίων έχει αντικατασταθεί από διάφορα στατιστικά βασισμένα στο μέρος του λόγου των σχετικών λέξεων. Και η προσέγγιση αυτή αξιολογήθηκε στο ίδιο σώμα κειμένων, χρησιμοποιώντας *πενταπλή διασταυρωμένη επικύρωση (5-fold cross validation)*, ενώ τα αποτελέσματα παρουσιάζονται στον ακόλουθο πίνακα (Πίνακας 47):

| | CRF++ | | | egGRIDS+ | | |
|---|---|---|---|---|---|---|
| | **Ακρίβεια** | **Ανάκληση** | **F-measure** | **Ακρίβεια** | **Ανάκληση** | **F-measure** |
| **Name-Ranking** | 59.30 % | 46.68 % | 52.79 % | 95.05 % | 54.07 % | 68.57 % |
| **Name-Performance** | 84.62 % | 81.40 % | 82.38 % | 92.14 % | 49.26 % | 64.17 % |
| **Name-Country** | 85.52 % | 85.47 % | 85.66 % | 98.85 % | 88.88 % | 93.58 % |
| **Name-Gender** | 46.89 % | 46.29 % | 44.63 % | 99.21 % | 79.17 % | 88.00 % |
| **Name-Age** | 84.79 % | 81.14 % | 83.04 % | 100.00 % | 98.11 % | 99.04 % |
| **Overall** | **81.13 %** | **77.28 %** | **79.16 %** | **96.48 %** | **65.96 %** | **78.32 %** |

**Πίνακας 47: Τα αποτελέσματα της αξιολόγησης της εναλλακτικής (δεύτερης) προσέγγισης που βασίζεται στον CRF++, στην εργασία της αναγνώρισης συσχετίσεων μεταξύ ονομάτων οντοτήτων.**

## 6.6 Συνεισφορά

Οι μέθοδοι εξαγωγής συσχετίσεων μεταξύ ονομάτων οντοτήτων τυπικά βασίζονται στην απόκτηση κανόνων εξαγωγής και γραμματικών: από τις εκφράσεις/πρότυπα Hearst [132], τα οποία προσπαθούν να ανιχνεύσουν ιεραρχικές σχέσεις όπως τα *υπερώνυμα (hypernyms)*, μέχρι πολύπλοκες λεξικό-συντακτικές γραμματικές [136] που στοχεύουν στην εξαγωγή $n$-αδικών σχέσεων με $n > 2$. Όντας κυρίως μέθοδοι με επίβλεψη ή ημί-επίβλεψη, συχνά συνδυάζουν συντακτικά δέντρα που αποκτιούνται μέσω συντακτικής ανάλυσης με ευριστικά που βασίζονται σε διαφορές στατιστικές μετρήσεις, προκειμένου να γενικεύσουν μια αρχική υπόθεση που σχηματίζεται από τα δεδομένα εκπαίδευσης.

Σε μια προσπάθεια να διευκολύνουμε τις απαιτήσεις που τίθενται από τέτοιες προσεγγίσεις, εξετάσαμε την καταλληλότητα ενός γενικής χρήσης αλγορίθμου γραμματικού συμπερασμού για την εργασία, στοχεύοντας στην εκτίμηση της καταλληλότητας για την αντικατάσταση, τόσο της ανάγκης για συντακτική ανάλυση, όσο και των ευριστικών που απαιτούνται στην καθοδήγηση της διαδικασίας γενίκευσης. Η προτεινόμενη προσέγγιση έχει αποτιμηθεί με την βοήθεια ενός χειρωνακτικά επισημειωμένου σώματος κειμένων, και τα αποτελέσματα της αξιολόγησης έδειξαν ότι η προσέγγιση αποδίδει συγκρίσιμα με άλλες προσεγγίσεις που απαντώνται στην διεθνή βιβλιογραφία, χωρίς ωστόσο να υπάρχει απαίτηση προηγούμενης επεξεργασίας όπως συντακτική ανάλυση ή αναγνώριση μερών του λόγου.

Επιπλέον, το γεγονός ότι η προτεινόμενη προσέγγιση δεν εισάγει κάποιο αφαιρετικό επίπεδο, πέρα από τη γενίκευση που πετυχαίνεται από τον αλγόριθμο γραμματικού συμπερασμού, μας επιτρέπει να έχουμε μια εκτίμηση του βαθμού γενίκευσης που μπορεί να επιτευχτεί από τον αλγόριθμο. Αυτή μετρήθηκε τουλάχιστον 10 μονάδες, συνοδευόμενη από μια υποβάθμιση της ακριβείας 29 περίπου μονάδων.

Δεδομένου ότι τα επιτευχθέντα αποτελέσματα είναι ικανοποιητικά, φαίνεται ενδιαφέρουσα η προσπάθεια εξάλειψης ακόμα και της απαίτησης για ένα επισημειωμένο σώμα κειμένου. Οι μέθοδοι χωρίς επίβλεψη έχουν προσελκύσει ένα σημαντικό ερευνητικό ενδιαφέρον, καθώς η χειρωνακτική επισημείωση είναι μια χρονοβόρα διαδικασία, η οποία χρειάζεται πόρους. Ακολουθώντας τις πρόσφατες εξελίξεις στον τομέα, η προτεινόμενη προσέγγιση μπορεί να τροποποιηθεί ώστε να δέχεται ένα σύνολο από ενδεικτικές γραμματικές ανεξάρτητες από τα συμφραζόμενα, με κάθε μια να περιέχει μόνο μερικούς κανόνες που στοχεύουν έναν συγκεκριμένο τύπο σχέσης. Χρησιμοποιώντας μια *διαδικασία επανεκκίνησης (bootstrapping)*, το σύστημα μπορεί να γενικεύσει αυτές τις γραμματικές-σπόρους σε σχέση με ένα σύνολο εγγράφων σχετικών με την θεματική περιοχή ενδιαφέροντος.

# 7. Συμπεράσματα και Μελλοντική Εργασία

Η επεξεργασία φυσικής γλώσσας είναι μια ερευνητική περιοχή με σημαντικότατο ερευνητικό ενδιαφέρον και ταυτόχρονα τόσο παλιά, όση σχεδόν και η ιστορία των ηλεκτρονικών υπολογιστών. Η επεξεργασία φυσικής γλώσσας (από τις πιο απλοϊκές ενέργειες μέχρι τις προσπάθειες «κατανόησης» κειμένων) ήταν και παραμένει ακρογωνιαίος λίθος της τεχνητής νοημοσύνης, και τη σημερινή εποχή, όπου λόγω του παγκοσμίου ιστού και του τεράστιου όγκου της πληροφορίας που διαρκώς προστίθεται σε αυτόν, κάνει επιτακτική την ανάγκη καλύτερης και κυρίως αυτοματοποιημένης οργάνωσης και διαχείρισης όλης αυτής της κειμενικής πληροφορίας.

Ωστόσο, η επεξεργασία φυσικής γλώσσας δέχθηκε μια σημαντική βοήθεια από έναν ακόμα βασικό πυλώνα της τεχνητής νοημοσύνης, την μηχανική μάθηση, η οποία στα τέλη της δεκαετίας του 1990 έκανε δειλά την εμφάνισή της για την αξιοποίησή της στην επεξεργασία κειμένων σε φυσική γλώσσα. Σήμερα, η χρήση μηχανικής μάθησης είναι αρκετά διαδεδομένη σε διάφορες εργασίες σχετικές με την επεξεργασία φυσικής γλώσσας, προσφέροντας ικανοποιητικές λύσεις λόγω της διαθεσιμότητας μεγάλων όγκων δεδομένων στα οποία μπορεί να εφαρμοστεί στατιστική ανάλυση, αλλά και λόγω της διαρκούς ωρίμανσης των αλγορίθμων μηχανικής μάθησης.

Σε αυτή την ενδιαφέρουσα μίξη επεξεργασίας φυσικής γλώσσας και μηχανικής μάθησης, εντάσσεται η παρούσα διατριβή. Ξεκινώντας σε μια εποχή όπου η μηχανική μάθηση άρχισε να βρίσκει ένα ενδιαφέρον πεδίο εφαρμογής στην επεξεργασία φυσικής γλώσσας, η διατριβή αυτή εξέτασε την εφαρμοσιμότητα αρκετών μεθόδων μηχανικής μάθησης, σε διάφορες εργασίες της εξαγωγής πληροφορίας από κείμενα. Κύριοι στόχοι της διατριβής αυτής ήταν από την μια η χρήση της μηχανικής μάθησης σαν εργαλείο για την εύκολη κατασκευή συστημάτων εξαγωγής πληροφορίας, εύκολα προσαρμόσιμων σε νέες θεματικές περιοχές, ακόμη και γλώσσες, και από την άλλη η χρήση της μηχανικής μάθησης με τέτοιο τρόπο ώστε να προκύψουν συστήματα που να μπορούν να εφαρμοστούν στην πράξη.

Το κεφάλαιο αυτό συνοψίζει την συμβολή της παρούσας διδακτορικής διατριβής, παρουσιάζοντας τα συμπεράσματα, αλλά και πιθανές κατευθύνσεις για μελλοντική έρευνα. Συγκεκριμένα, στην ενότητα 7.1 συνοψίζονται τα βασικά συμπεράσματα της παρούσας διατριβής, ενώ στην ενότητα 7.2 συζητούνται πιθανές βελτιώσεις και επεκτάσεις των προσεγγίσεων που παρουσιάστηκαν στα προηγούμενα κεφάλαια της διατριβής αυτής.

## 7.1 Συμπεράσματα

Η παρούσα διατριβή κινήθηκε σε δύο κύριους άξονες: α) την αξιοποίηση και αποτίμηση υπαρχόντων αλγορίθμων μηχανικής μάθησης κυρίως στα στάδια της προ-επεξεργασίας (όπως η αναγνώριση μερών του λόγου) και της αναγνώρισης ονομάτων οντοτήτων, και β) την ανάπτυξη ενός νέου αλγορίθμου μηχανικής μάθησης και αποτίμησής του, τόσο σε συνθετικά δεδομένα, όσο και σε πραγματικά δεδομένα που αφορούσαν το στάδιο της εξαγωγής σχέσεων μεταξύ ονομάτων οντοτήτων.

### 7.1.1 Αξιοποίηση και αποτίμηση υπαρχόντων αλγορίθμων μηχανικής μάθησης

Παραδοσιακά, η εξαγωγή πληροφορίας από κείμενα είναι μια απαιτητική σε πόρους εργασία, τόσο γλωσσικούς, όσο και ανθρώπινους. Η ανυπαρξία αυτών των πόρων σε γλώσσες όπως η Ελληνική, σύντομα μας οδήγησε στην αναζήτηση εναλλακτικών τρόπων αντιμετώπισης της έλλειψης πόρων, κυρίως μέσω της μηχανικής μάθησης. Ξεκινώντας πρακτικά από το μηδέν, η εργασία που περιγράφεται στην διατριβή αυτή

προσπάθησε να κατασκευάσει ένα σύστημα εξαγωγής πληροφορίας το οποίο θα χρησιμοποιεί μηχανική μάθηση σε όσο το δυνατόν περισσότερα υποσυστήματα του.

Μια από τις πρώτες εργασίες ενός συστήματος εξαγωγής πληροφορίας, είναι η γλωσσική προ-επεξεργασία, η οποία συνήθως περιλαμβάνει κάποιες βασικές εργασίες, όπως η αναγνώριση λέξεων, προτάσεων, μερών του λόγου των λέξεων, ενώ συχνά περιλαμβάνει και πιο σύνθετη μορφολογική ανάλυση, όπως η εύρεση θεμάτων ή λημμάτων λέξεων. Λύνοντας εύκολα το θέμα της αναγνώρισης λέξεων και προτάσεων για την Ελληνική γλώσσα, οδηγούμαστε στο πρώτο στάδιο της μορφολογικής ανάλυσης το οποίο δεν μπορεί να αντιμετωπιστεί εύκολα, την αναγνώριση μερών του λόγου.

Στα πλαίσια της διατριβής αυτής εξετάστηκε το πρόβλημα της αναγνώρισης μερών του λόγου, τόσο από ερευνητική σκοπιά, όσο και από πρακτική, δημιουργώντας ένα σύστημα αναγνώρισης μερών του λόγου για την Ελληνική γλώσσα, το οποίο μπορεί να εφαρμοστεί στην πράξη. Τα συμπεράσματα που προέκυψαν σχετικά με αυτό το πεδίο εφαρμογής είναι αρκετά, και μπορούν να συνοψιστούν ως εξής:

- Προσαρμόστηκε και εξετάστηκε η συμπεριφορά της *μάθησης στηριζόμενη σε κανόνες μετασχηματισμού καθοδηγούμενη από σφάλματα (transformation-based error-driven learning – TBED)* για την εργασία της αναγνώρισης μερών του λόγου της Ελληνικής γλώσσας. Παρότι η συγκεκριμένη μέθοδος σχεδιάστηκε και αποτιμήθηκε για την Αγγλική γλώσσα, η πειραματική αξιολόγηση έδειξε ότι μπορεί να εφαρμοστεί και σε κλιτικές γλώσσες, όπως η Ελληνική.
- Δημιουργήθηκε ένα σύνολο ετικετών για την Ελληνική γλώσσα, το οποίο παρέχει μια ικανοποιητική σχέση μεταξύ της παρεχόμενης μορφολογικής πληροφορίας και των απαιτήσεων σε μέγεθος δεδομένων εκπαίδευσης.
- Εξετάστηκε η απόδοση της τεχνικής TBED σε δύο διαφορετικές θεματικές περιοχές: μία αρκετά περιορισμένη και μια πιο ευρεία. Η πειραματική αξιολόγηση έδειξε ότι η απόδοση δεν εξαρτάται σημαντικά από την θεματική περιοχή.
- Η απόδοση της μεθόδου TBED για την Ελληνική γλώσσα μετρήθηκε σε ικανοποιητικά για πρακτική χρήση επίπεδα, παρουσιάζοντας συγκρίσιμες επιδόσεις με τις καλύτερες προσεγγίσεις που έχουν εμφανιστεί στην διεθνή βιβλιογραφία για την αναγνώριση μερών του λόγου σχετικά με την Ελληνική γλώσσα.
- Η προσέγγιση που περιγράφεται σε αυτή την διατριβή αποτέλεσε το πρώτο πρακτικό σύστημα αναγνώρισης μερών του λόγου για την Ελληνική γλώσσα, ο οποίος διατέθηκε ελεύθερα για κάθε χρήση, με άδεια χρήσης ανοικτού λογισμικού (LGPL).
- Η μέθοδος TBED συνδυάστηκε με ένα μορφολογικό λεξικό σε ηλεκτρονική μορφή. Η πειραματική αξιολόγηση μέτρησε την υψηλότερη επίδοση που έχει αναφερθεί μέχρι σήμερα, για την αναγνώριση ονομάτων οντοτήτων της Ελληνικής γλώσσας.

Ένα ακόμα πρόβλημα που εξετάστηκε στα πλαίσια της διατριβής αυτής, είναι το πρόβλημα της αναγνώρισης ονομάτων οντοτήτων, τόσο από ερευνητική σκοπιά, όσο και από πρακτική, δημιουργώντας συστήματα αναγνώρισης ονομάτων οντοτήτων για διάφορες θεματικές περιοχές και γλώσσες, τα οποία μπορούν να εφαρμοστούν στην πράξη. Τα συμπεράσματα που προέκυψαν σχετικά με αυτό το πεδίο εφαρμογής είναι αρκετά, και μπορούν να συνοψιστούν ως εξής:

- Εξετάστηκε και συγκρίθηκε η συμπεριφορά δύο υπαρχόντων αλγορίθμων μηχανικής μάθησης, ενός συμβολικού (δέντρο αποφάσεων) και ενός αριθμητικού (νευρωνικό δίκτυο). Αντίστοιχα δοκιμάστηκαν δύο διαφορετικές αναπαραστάσεις, μια συμβολική και μια αριθμητική. Η αξιολόγηση έδειξε ότι οι δύο αλγόριθμοι

έχουν συγκρίσιμη απόδοση, κυρίως όταν χρησιμοποιούν την ίδια αναπαράσταση.

- Η αριθμητική αναπαράσταση δεδομένων που εξετάστηκε περιείχε λιγότερη πληροφορία από την συμβολική, αφού δεν κωδικοποιούσε την σειρά των λέξεων στο κείμενο. Η αξιολόγηση έδειξε ότι η σειρά των λέξεων δεν παίζει σημαντικό ρόλο για την εργασία της αναγνώρισης ονομάτων οντοτήτων. Επιπρόσθετα, η αξιολόγηση έδειξε ότι η αριθμητική αναπαράσταση απέδιδε καλύτερα από την συμβολική, έχοντας σαν αλγόριθμο μηχανικής μάθησης το δέντρο αποφάσεων, που μπορεί να χειριστεί και τις δύο αναπαραστάσεις, παρότι περιείχε λιγότερη πληροφορία.
- Η απόδοση συστημάτων αναγνώρισης ονομάτων οντοτήτων βασισμένων σε μηχανική μάθηση, για τους αλγορίθμους και τις αναπαραστάσεις που δοκιμάστηκαν, μετρήθηκε να είναι χαμηλότερη συμβατικών συστημάτων με σημαντική επένδυση χρόνου (σύστημα MITOS, Ελληνικά), αλλά σημαντικά υψηλότερη από συμβατικά συστήματα με λιγότερη επένδυση χρόνου στην κατασκευή τους (σύστημα VIE, Αγγλικά).
- Προτάθηκε μια προσέγγιση που συνδυάζει δύο διαφορετικούς τύπους ταξινομητών, λειτουργώντας σε επίπεδο λέξεων και σε επίπεδο φράσεων. Και οι δύο τύποι ταξινομητών βασίζονται σε μηχανική μάθηση, με τον ταξινομητή που λειτουργεί σε επίπεδο λέξεων να αποτελείται από διάφορους αλγορίθμους μηχανικής μάθησης και διάφορες αναπαραστάσεις, συνδυασμένους μέσω πλειοψηφικής ψηφοφορίας. Η αξιολόγηση της προτεινόμενης προσέγγισης έδειξε ότι το σύστημα αναγνώρισης ονομάτων οντοτήτων μπορεί να εφαρμοστεί επιτυχώς σε διάφορες γλώσσες (εξετάστηκαν η Αγγλική και η Ελληνική γλώσσα) και διάφορες θεματικές περιοχές (περιγραφές προϊόντων ηλεκτρονικών καταστημάτων, αγγελίες προσφοράς εργασίας).
- Η προσέγγιση που συνδυάζει δύο διαφορετικούς τύπους ταξινομητών αξιολογήθηκε συγκριτικά με αντίστοιχο σύστημα από το Πανεπιστήμιο του Εδιμβούργου (το οποίο έχει μεγάλη εμπειρία στην ανάπτυξη συστημάτων αναγνώρισης ονομάτων οντοτήτων για την Αγγλική γλώσσα) σε κείμενα της Αγγλικής γλώσσας, και αποτιμήθηκε ελαφρώς καλύτερο σε σχέση με το σύστημα του Πανεπιστημίου του Εδιμβούργου.
- Εξετάστηκε η διαφοροποίηση της απόδοσης για την προσέγγιση που συνδυάζει δύο διαφορετικούς τύπους ταξινομητών, με την προσθήκη γνώσης για την θεματική περιοχή. Η αξιολόγηση έδειξε ότι η απόδοση του συστήματος βελτιώνεται σημαντικά, κυρίως στην *ανάκληση (recall)*.
- Εξετάστηκε η αυτόματη προσαρμογή/εμπλουτισμός του υποσυστήματος του λεξικού, χρησιμοποιώντας ένα σύστημα αναγνώρισης ονομάτων οντοτήτων βασισμένο σε μηχανική μάθηση και ταυτόχρονα προσανατολισμένο να εμφανίζει μια προτίμηση προς την ακρίβεια έναντι της ανάκλησης. Η προτεινόμενη προσέγγιση που εξετάστηκε κατάφερε να αυξήσει ένα *αρχικό (seed)* λεξικό κατά 36 %, πετυχαίνοντας ακρίβεια 94.03 % και καταφέρνοντας να εντοπίσει το 72.94 % των ονομάτων οντοτήτων που περιέχονταν στο σώμα κειμένων αξιολόγησης και ταυτόχρονα δεν περιέχονταν στο λεξικό.
- Τέλος, εξετάσαμε την ενημέρωση ενός συστήματος αναγνώρισης ονομάτων οντοτήτων, μέσω της κατασκευής ενός παράλληλου συστήματος, το οποίο επιτηρεί το αρχικό σύστημα μέσω του εντοπισμού των διαφορών τους. Η προσέγγιση αυτή χρησιμοποιεί την μηχανική μάθηση με έναν καινοτόμο τρόπο (σαν επιτηρητή ενός άλλου συστήματος), και το αποτέλεσμα της προσέγγισης μπορεί να χρησιμοποιηθεί για να προσδιοριστεί τόσο η χρονική στιγμή που απαιτείται η ενημέρωση του συστήματος (π.χ. βάζοντας ένα άνω όριο στις

διαφορές μεταξύ των δύο συστημάτων, του συστήματος επιτήρησης και του επιτηρούμενου συστήματος), αλλά και να καθοδηγήσει την ενημέρωση του επιτηρούμενου συστήματος, εστιάζοντας στην βελτίωση του συστήματος στις περιπτώσεις που εμφανίζονται διαφωνίες.

Η διδακτορική διατριβή επίσης συνέβαλε στην χρήση της μηχανικής μάθησης στην επεξεργασία φυσικής γλώσσας, εξετάζοντας την εφαρμοσιμότητα αρκετών μεθόδων μηχανικής μάθησης σε διάφορες εργασίες της εξαγωγής πληροφορίας, και παρουσιάζοντας τα αποτελέσματα σε διεθνή καταξιωμένα συνέδρια, εκδόσεις επιστημονικών περιοδικών και βιβλίων.

### 7.1.2 egGRIDS+: ένας νέος αλγόριθμος επαγωγικής εξαγωγής γραμματικών

Η χρήση γραμματικών για διάφορες εργασίες στην επεξεργασία φυσικής γλώσσας, δεν είναι μια άγνωστη πρακτική. Το αντίθετο μάλιστα: οι γραμματικές αποτελούν ένα από τα κύρια μέσα «βαθιάς» γλωσσικής ανάλυσης (σε αντιδιαστολή με την «ρηχή» ανάλυση που συνήθως επιτυγχάνεται με μηχανική μάθηση ή στατιστικές μεθόδους), κυρίως στο συντακτικό και σημασιολογικό επίπεδο. Ωστόσο, η εφαρμογή μηχανικής μάθησης για την αυτόματη εξαγωγή γραμματικών δεν είναι μια διαδεδομένη πρακτική στο χώρο της επεξεργασίας φυσικής γλώσσας (σε αντίθεση με άλλους χώρους, όπως η αποκωδικοποίηση ακολουθιών DNA), γεγονός που αποτέλεσε ένα σημαντικό κίνητρο για την ενασχόληση με αυτή την ερευνητική περιοχή στο πλαίσιο αυτής της διατριβής.

Η εφαρμογή μηχανικής μάθησης για την επαγωγική εξαγωγή γραμματικών στην περιοχή της επεξεργασίας φυσικής γλώσσας, έχει να αντιμετωπίσει μια σειρά από αρνητικά θεωρητικά αποτελέσματα. Ένα από τα σημαντικότερα, είναι η αδυναμία εξαγωγής ακόμα και της απλούστερης των γραμματικών της ιεραρχίας του Chomsky – των κανονικών γραμματικών – μόνο από θετικά παραδείγματα. Συνεπώς, οι περισσότεροι αλγόριθμοι επαγωγικής εξαγωγής γραμματικών απαιτούν την παρουσία αρνητικών παραδειγμάτων, μια σημαντική απαίτηση για ένα πεδίο εφαρμογής, όπως η επεξεργασία φυσικής γλώσσας, όπου τα αρνητικά παραδείγματα σπανίζουν.

Μία από τις σημαντικότερες συνεισφορές της παρούσας διατριβής, είναι ένας νέος αλγόριθμος επαγωγικής εξαγωγής γραμματικών, γνωστός με την ονομασία egGRIDS+. Προοριζόμενος για πρακτικές εφαρμογές στην περιοχή της επεξεργασίας φυσικής γλώσσας, ο egGRIDS+ δεν απαιτεί την παρουσία αρνητικών παραδειγμάτων: είναι σε θέση να εξάγει γραμματικές ανεξάρτητες από τα συμφραζόμενα μόνο από θετικά παραδείγματα. Αυτό επιτυγχάνεται με την χρήση ευριστικών, τα οποία προτιμούν «απλές» γραμματικές, μέσω της *ελαχιστοποίησης του μήκους περιγραφής (minimum description length – MDL)*. Τα συμπεράσματα που προέκυψαν από την αξιολόγηση του νέου αλγορίθμου είναι αρκετά, και μπορούν να συνοψιστούν ως εξής:

- Ο αλγόριθμος egGRIDS+ βασίστηκε σε προγενέστερα συστήματα, και κυρίως στον GRIDS [13] και στον SNPR [93]. Ωστόσο, ο αλγόριθμος egGRIDS+ είναι υπολογιστικά αποδοτικότερος, επιτρέποντας την εφαρμογή του σε σημαντικά μεγαλύτερους όγκους δεδομένων, από ότι οι προκάτοχοί του. Ο λόγος της σημαντικής βελτίωσης των υπολογιστικών απαιτήσεων βρίσκεται στην ιδέα ότι είναι υπολογιστικά οικονομικότερο να προβλέψεις το μήκος περιγραφής του μοντέλου, από το να κατασκευάσεις και να αποτιμήσεις το μοντέλο. Η ιδέα αυτή επιβεβαιώθηκε και θεωρητικά, μετατρέποντας την διαδικασία αναζήτησης από κυβική (όσον αφορά τους προκατόχους του egGRIDS+), σε τετραγωνική όσον αφορά τον απαιτούμενο αριθμό των παραδειγμάτων εκπαίδευσης.
- Ο αλγόριθμος SNPR χρησιμοποιούσε ευριστικά για την επιτάχυνση της αναζήτησης στον χώρο των πιθανών γραμματικών. Συγκεκριμένα, ομαδοποιούσε διγράμματα με υψηλή συχνότητα εμφάνισης. Η θεωρητική ανάλυση του αλγόριθμου egGRIDS+ έδειξε ότι αυτό το ευριστικό είναι σωστό, οδηγώντας σε «απλούστερες» ή «καλύτερες» γραμματικές σύμφωνα με το MDL. Ένα δεύτερο ευριστικό αφορούσε τη συγχώνευση συμβόλων που εμφανίζονται συχνά σε παρόμοια περιβάλλοντα. Σε αυτή την περίπτωση, η θεωρητική ανάλυση του egGRIDS+ έδειξε ότι αυτό το ευριστικό είναι λαθεμένο, και δεν οδηγεί πάντα σε «απλούστερες» ή «καλύτερες» γραμματικές σύμφωνα πάντα με το MDL.
- Ο αλγόριθμος egGRIDS+ αποτιμήθηκε σε τεχνητές γλώσσες, όπου εμφάνισε ικανοποιητική απόδοση.

- Η βελτίωση της υπολογιστικής απόδοσης του egGRIDS+, η οποία είχε προβλεφθεί θεωρητικά, αποτιμήθηκε και πρακτικά, εφαρμόζοντας τον αλγόριθμο σε ένα πραγματικό σώμα κειμένων σημαντικού μεγέθους.
- Η απόδοση του egGRIDS+ εξετάστηκε εμπειρικά στην *αναγνώριση συσχετίσεων μεταξύ ονομάτων οντοτήτων (relation extraction)*. Η απόδοση του αλγορίθμου μετρήθηκε σε επίπεδα συγκρίσιμα με τα αποτελέσματα που έχουν αναφερθεί στην διεθνή βιβλιογραφία για αυτή την εργασία.

## 7.2 Προοπτικές μελλοντικές έρευνας

Η ερευνητική εργασία που παρουσιάστηκε σε αυτή τη διατριβή προετοιμάζει το έδαφος για περαιτέρω βελτιώσεις και επεκτάσεις. Λαμβάνοντας υπόψη τα θετικά αποτελέσματα από την αξιοποίηση του αλγόριθμου egGRIDS+ σε πραγματικό σώμα κειμένων σημαντικού μεγέθους, πιστεύουμε ότι ορισμένες βελτιώσεις μπορούν να αυξήσουν την απόδοσή του ακόμα περισσότερο. Οι πιο ενδιαφέρουσες επεκτάσεις είναι εκείνες που απαιτούνται για να αντιμετωπιστούν πολύπλοκα γλωσσικά προβλήματα. Προς αυτή την κατεύθυνση, ο χειρισμός των *γραμματικών με χαρακτηριστικά (attribute grammars)* παρουσιάζει ιδιαίτερο ενδιαφέρον. Η εκμάθηση των γραμματικών χαρακτηριστικών είναι σημαντική για διάφορα προβλήματα όπου τα χαρακτηριστικά σχετίζονται με χαρακτηριστικά των δεδομένων εκπαίδευσης. Για παράδειγμα, στην αναγνώριση ονομάτων οντοτήτων, κάθε λέξη συνδέεται συνήθως με πρόσθετες πληροφορίες -πέρα από το μέρος του λόγου- όπως μορφολογικές πληροφορίες ή πληροφορίες σχετικές με το εάν η λέξη έχει προσδιοριστεί ως τμήμα μιας γνωστής ονοματικής οντότητας που περιλαμβάνεται σε ένα γεωγραφικό λεξικό. Αυτές οι πληροφορίες είναι σημαντικές για το πρόβλημα της αναγνώρισης ονομάτων οντοτήτων, συνεπώς η εισαγωγή της υποστήριξης χαρακτηριστικών στον egGRIDS+ θα βοηθήσει σημαντικά τη μοντελοποίηση γλωσσικών προβλημάτων.

Επιπλέον, ο egGRIDS+ μπορεί να αξιοποιηθεί περαιτέρω, και να συγκριθεί με άλλους υπάρχοντες αλγορίθμους. Ένας στόχος για μελλοντική εργασία είναι να συγκριθεί η υπολογιστική επάρκεια του egGRIDS+ με άλλους αλγορίθμους που έχουν εφαρμοστεί σε προβλήματα που περιλαμβάνουν πολύπλοκες γλώσσες με μεγάλα αλφάβητα (π.χ. διάφορα προβλήματα επεξεργασίας φυσικής γλώσσας). Μια τέτοια αξιολόγηση θα επιτρέψει επίσης τη σύγκριση της ακρίβειας των γραμματικών που προκύπτουν από τον egGRIDS+ με εκείνες που προκύπτουν από άλλους αλγορίθμους, κάτω από τις ίδιες πειραματικές συνθήκες.

Μια άλλη ενδιαφέρουσα πτυχή που μπορεί να εξεταστεί είναι η δυνατότητα του egGRIDS+ να μαθαίνει κατά τρόπο επαυξητικό. Η τρέχουσα υλοποίηση του αλγορίθμου προσφέρει τη δυνατότητα να φορτωθεί μια γραμματική -που έχει προκύψει νωρίτερα από τον egGRIDS+- και να γενικευτεί σε σχέση με ένα νέο σύνολο παραδειγμάτων. Θα ήταν ενδιαφέρον να εξεταστεί η συμπεριφορά εκμάθησης του egGRIDS+, όταν αυτός μαθαίνει με τρόπο επαυξητικό σε σχέση με όταν μαθαίνει με τον κλασσικό (μη επαυξητικό) τρόπο.

Μια πρόσθετη δυνατότητα που παρέχεται από τον egGRIDS+, είναι ο καθορισμός εναλλακτικών στρατηγικών για τους τελεστές μάθησης. Για παράδειγμα, ο egGRIDS+ μπορεί να λειτουργήσει με τέτοιο τρόπο όπου οι τρεις τελεστές παρεμβάλλονται, αντί καθένας να εφαρμόζεται συνεχώς μέχρι να αδυνατεί να οδηγήσει σε περαιτέρω βελτίωση. Σε αυτό το κεφάλαιο έχει γίνει σημαντική εργασία ώστε να καταστεί κατανοητή η επίδραση κάθε τελεστή χωριστά στην εξαχθείσα γραμματική. Μελλοντική έρευνα θα μπορούσε να εστιάσει στην επίδραση εναλλακτικών στρατηγικών για τους τρεις τελεστές.

Ολοκληρώνοντας, η έρευνα που παρουσιάστηκε στη διατριβή αυτή έχει οδηγήσει σε έναν νέο αλγόριθμο που εξαλείφει αρκετές από τις ανεπάρκειες των προκατόχων του, ενώ δίνεται ιδιαίτερη προσοχή στην ευρωστία και την υπολογιστική του επάρκεια. Πιστεύουμε ότι ο egGRIDS+ θα είναι χρήσιμος στη μοντελοποίηση διαφόρων υπο-προβλημάτων φυσικής γλώσσας που δε θα μπορούν εύκολα να μοντελοποιηθούν από άλλους αλγορίθμους μηχανικής μάθησης, τουλάχιστον εκείνους που απαιτούν στην είσοδό τους διανύσματα χαρακτηριστικών σταθερού μήκους. Ακόμα κι αν κάποιοι αλγόριθμοι μηχανικής μάθησης μπορούν να εφαρμοστούν σε τέτοια υπο-προβλήματα, οι γραμματικές θα έχουν ακόμα κάποιο πλεονέκτημα, καθώς είναι ο πιο φυσικός τρόπος αναζήτησης τμημάτων κειμένου. Γενικά, ενδιαφέροντα προβλήματα που θα μπορούσαν να εξεταστούν αξιοποιώντας το αλγόριθμο egGRIDS+ είναι η *κατάτμηση ονοματικών φράσεων (noun phrase chunking)*, η αναγνώριση ονομάτων οντοτήτων, καθώς και σε διάφορες άλλες υπο-εργασίες της εξαγωγής πληροφορίας.

# ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

| **Ξενόγλωσσος όρος** | **Ελληνικός Όρος** |
|---|---|
| 10-fold cross validation | δεκαπλή διασταυρωμένη επικύρωση |
| 5-fold cross validation | πενταπλή διασταυρωμένη επικύρωση |
| abstraction layer | επίπεδο απόκρυψης |
| accuracy | ορθότητα |
| active learning | ενεργητική μάθηση |
| ambiguous | διφορούμενες |
| artifact | αντικείμενο |
| artificial neural networks | τεχνητά νευρωνικά δίκτυα |
| athletics | στίβος |
| attribute grammars | γραμματικές με χαρακτηριστικά |
| automaton | αυτόματο |
| back propagation | ανάστροφη μετάδοση |
| bag-of-words representation | αναπαράσταση «συνόλου λέξεων» |
| beam search | αναζήτηση δέσμης |
| beam search | αναζήτηση δέσμης |
| bigrams | διγράμματα |
| bits | δυφία |
| bootstrapping | μέθοδος επανεκκίνησης |
| boundary | όριο |
| center embedding | κεντρική ενσωμάτωση |
| classification | ταξινόμηση |
| classifier | ταξινομητής |
| clustering | κατηγοριοποίηση |
| company designators | προσδιοριστές ονομάτων εταιριών |
| consistent | συνεπής |
| context free grammars | γραμματικές ανεξάρτητες από τα συμφραζόμενα |
| context sensitive grammars | γραμματικές συμφραζομένων |
| contextual rules | κανόνες συμφραζομένων |
| contingency table | πίνακας συνάφειας |
| co-reference | συν-αναφορά |
| co-reference resolution | επίλυση καθορισμού συν-αναφορών |
| corpora | σωμάτων κειμένων |
| corpus | σώμα κειμένων |
| decision tree | δέντρο αποφάσεων |
| delimitation | αρχική οριοθέτηση (π.χ. ονομάτων οντοτήτων) |
| derivation power | δυναμικότητα παραγωγής |
| derivations | παραγωγές |
| derivations description length | μήκος περιγραφής παραγωγών |
| descriptor | περιγραφέας |
| dimensionality | διαστατικότητα |
| empiricism | εμπειρική μέθοδος |
| entities | οντότητες |
| errors of commission | λάθη διάπραξης |
| errors of omission | λάθη παραλείψεων |
| evaluation | αποτίμηση |
| event | σενάριο γεγονότος |

| | |
|---|---|
| expert systems | έμπειρο σύστημα |
| exhaustive enumeration | εξαντλητική απαρίθμηση |
| field | πεδίο |
| filtering | φιλτράρισμα |
| formal language | τυπική γλώσσα |
| gazetteer | αναγνωριστής γνωστών ονομάτων οντοτήτων |
| gazetteer list lookup | αναγνωριστής γνωστών ονομάτων οντοτήτων |
| gazetteer lists | κατάλογοι (λίστες) γνωστών ονομάτων οντοτήτων |
| global approximators | σύστημα προσέγγισης οποιαδήποτε συνάρτησης |
| gradient descent | κλιμακωτή κάθοδος |
| grammar description length | μήκος περιγραφής γραμματικής |
| grammatical inference | επαγωγική εξαγωγή γραμματικών, αλγόριθμος γραμματικού συμπερασμού |
| greedy | εξαντλητικός |
| heuristics | ευριστικά |
| hidden layer | κρυμμένο επίπεδο απόφασης |
| hidden Markov models | κρυφά μοντέλα Markov |
| hill-climbing | αναζήτηση «αναρρίχησης λόφου» |
| hypernyms | υπερώνυμα |
| incrementally | επαυξητικά |
| inductive grammar learning | επαγωγική εξαγωγή γραμματικών |
| inference process | συμπερασματική διαδικασία |
| information extraction | εξαγωγή πληροφορίας |
| information filtering | φιλτράρισμα πληροφορίας |
| information overloading | υπερπληροφόρηση |
| information retrieval | ανάκτηση πληροφορίας |
| information theory | θεωρία πληροφοριών |
| instances | στιγμιότυπα |
| k-nearest-neighbours | ταξινομητής k-κοντινότερων γειτόνων |
| learning operators | τελεστές εκμάθησης |
| lexical rules | λεκτικοί κανόνες |
| lexicalisations | λεκτικές μορφές/αναφορές |
| lexico-syntactic patterns | λεξικο-συντακτικά πρότυπα |
| linguistic knowledge acquisition bottleneck | δυσκολίας απόκτησης γλωσσολογικής γνώσης |
| machine learning | μηχανική μάθηση |
| matching techniques | διαδικασίες ταιριάσματος |
| maximal posterior probability | μέγιστη μεταγενέστερη πιθανότητα |
| maximum entropy | μεγιστοποίηση της εντροπίας |
| maximum entropy tagger | ταξινομητής μεγιστοποίησης της εντροπίας |
| memory-based learning | μάθηση με αποθήκευση στην μνήμη |
| mentions | αναφορές, εναλλακτικά ονόματα μιας οντότητας |
| minimally supervised approaches | ελάχιστα επιβλεπόμενες προσεγγίσεις |
| minimum description length | ελάχιστο μήκος περιγραφής |
| mutual information | αμοιβαία πληροφορία |
| named entity extraction | αναγνώριση ονομάτων οντοτήτων |

| | |
|---|---|
| named entity recognition | αναγνώριση ονομάτων οντοτήτων |
| named entity recognition and classification (NERC) | αναγνώριση ονομάτων οντοτήτων |
| n-gram taggers | ν-γραμματικοί αναγνωριστές |
| non-terminal symbols | μη τερματικά σύμβολα |
| noun phrase chunker | σύστημα κατάτμησης ονοματικών φράσεων |
| noun phrase chunking | κατάτμηση ονοματικών φράσεων |
| noun phrases | ονοματικές φράσεις |
| np chunker | αναγνωριστής ονοματικών φράσεων |
| occurrence | εμφάνιση |
| over-fitting | απομνημόνευση |
| overgeneralisation | υπεραπλούστευση |
| parse trees | δέντρα ανάλυσης |
| part of speech | μέρος του λόγου |
| part of speech tagger | αναγνωριστής μερών του λόγου |
| patterns | πρότυπα |
| phrasal classes | φραστικές κατηγορίες |
| phrase structured grammars | γραμματικές δομημένων φράσεων |
| precision | ακρίβεια |
| pruning | περικοπή |
| pure context-free languages | «καθαρά» ανεξάρτητες από τα συμφραζόμενα γλώσσες |
| recall | ανάκληση |
| recursion | αναδρομή |
| recursive grammars | αναδρομικές γραμματικές |
| recursive partitioning | αναδρομικός χωρισμός |
| regular expressions | κανονικές εκφράσεις |
| regular grammar | κανονική γραμματική |
| relation extraction | εξαγωγή σχέσεων/συσχετίσεων μεταξύ ονομάτων οντοτήτων |
| relations among entities | συσχετίσεις/σχέσεις μεταξύ οντοτήτων |
| relationship extraction | εξαγωγή σχέσεων/συσχετίσεων μεταξύ ονομάτων οντοτήτων |
| representation change | αλλαγή αναπαράστασης |
| reversible languages | αντιστρέψιμες γλώσσες |
| robust | εύρωστος |
| scalability | επεκτασιμότητα, κλιμακωσιμότητα |
| scenario template | σχεδιότυπο σεναρίου |
| seed rules | ενδεικτικοί κανόνες |
| sentence splitter | αναγνωριστής προτάσεων |
| sentiment analysis | ανάλυση συναισθήματος |
| set of finite state automata | σύνολο από αυτόματα πεπερασμένων καταστάσεων |
| shallow syntactic parser | ρηχός συντακτικός αναλυτής |
| slot-filler | δεδομένα πλήρωσης θέσης |
| slots | πεδία |
| sparseness | διάσπαση |
| standard deviation | τυπική απόκλιση |
| start-state rule | κανόνας αρχικοποίησης |
| stems | θέμα (λέξης) |

| | |
|---|---|
| strictly deterministic automata | αυστηρώς ντετερμινιστικά αυτόματα |
| structured information | δομημένη πληροφορία |
| subsymbolic or numeric | στοχαστικός ή αριθμητικός |
| succession management events | επιτυχή γεγονότα διαδοχής διαχείρισης |
| supervised | επιβλεπόμενος |
| supervised learning | επιβλεπόμενη μάθηση |
| symbolic | συμβολικός |
| tag | ετικέτα |
| tag set | σύνολο ετικετών |
| target-slot | θέση-στόχος |
| template | σχεδιότυπο |
| template element | σχεδιότυπο οντότητας |
| template element filling | πλήρωση σχεδιότυπων |
| template relation | σχεδιότυπο σχέσης |
| terminal symbol | τερματικό σύμβολο |
| text understanding | κατανόηση κειμένου |
| threshold | κατώφλι |
| token | λεκτική μονάδα |
| tokeniser | αναγνωριστής λέξεων |
| transformation-based error-driven learning | μάθηση βασισμένη σε κανόνες μετασχηματισμού καθοδηγούμενη από σφάλματα |
| trigram tagger | τρι-γραμματικοί αναγνωριστές |
| unambiguous | αποσαφηνισμένες |
| undecidability | αναποφασιστικότητα |
| unseen examples | απαρατήρητα παραδείγματα |
| unsupervised learning | μάθηση χωρίς επίβλεψη |
| verb concordances | συνταυτίσεις ρημάτων |
| word forms | λεκτικές μορφές |
| word sense disambiguation | αποσαφήνιση εννοιών λέξεων |

# ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

| | |
|---|---|
| DL | Description Length |
| DDL | Derivations Description Length |
| GDL | Grammar Description Length |
| MDL | Minimum Description Length |
| POS | Part of Speech |
| ΕΚΕΦΕ | Εθνικό Κέντρο Ερευνών Φυσικών Επιστημών |
| ΕΚΠΑ | Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών |
| ΟΠΑ | Οικονομικό Πανεπιστήμιο Αθηνών |

# ΠΑΡΑΡΤΗΜΑ Ι

## Αναγνωριστής Μερών του Λόγου: το πλήρες σύνολο ετικετών

| | |
|---|---|
| | **Άρθρα** |
| DDT | Οριστικό Άρθρο |
| IDT | Αόριστο Άρθρο |
| | **Ουσιαστικά** |
| NNM | Ουσιαστικό, ενικός, αρσενικό |
| NNF | Ουσιαστικό, ενικός, θηλυκό |
| NNN | Ουσιαστικό, ενικός, ουδέτερο |
| NNSM | Ουσιαστικό, πληθυντικός, αρσενικό |
| NNSF | Ουσιαστικό, πληθυντικός, θηλυκό |
| NNSN | Ουσιαστικό, πληθυντικός, ουδέτερο |
| NNPM | Κύριο Όνομα (ουσιαστικό), ενικός, αρσενικό |
| NNPF | Κύριο Όνομα (ουσιαστικό), ενικός, θηλυκό |
| NNPN | Κύριο Όνομα (ουσιαστικό), ενικός, ουδέτερο |
| NNPSM | Κύριο Όνομα (ουσιαστικό), πληθυντικός, αρσενικό |
| NNPSF | Κύριο Όνομα (ουσιαστικό), πληθυντικός, θηλυκό |
| NNPSN | Κύριο Όνομα (ουσιαστικό), πληθυντικός, ουδέτερο |
| | **Επίθετα** |
| JJM | Επίθετο, ενικός, αρσενικό |
| JJF | Επίθετο, ενικός, θηλυκό |
| JJN | Επίθετο, ενικός, ουδέτερο |

| | |
|---|---|
| JJSM | Επίθετο, πληθυντικός, αρσενικό |
| JJSF | Επίθετο, πληθυντικός, θηλυκό |
| JJSN | Επίθετο, πληθυντικός, ουδέτερο |
| CD | Απόλυτα αριθμητικά (ένα, δύο, τρία, ... και νούμερα: 1, 2, 3...) |
| | **Αντωνυμίες** |
| PRP | Προσωπική Αντωνυμία |
| PP | Κτητική Αντωνυμία |
| REP | Αυτοπαθής Αντωνυμία |
| DP | Οριστική Αντωνυμία |
| IP | Δεικτική Αντωνυμία |
| WP | Αναφορική Αντωνυμία |
| QP | Ερωτηματική Αντωνυμία |
| INP | Αόριστη Αντωνυμία |
| | **Ρήματα** |
| VB | Ρήμα παροντικού χρόνου |
| VBD | Ρήμα παρελθοντικού χρόνου |
| VBF | Ρήμα μελλοντικού χρόνου |
| VBS | Ρήμα παροντικού χρόνου, πληθυντικός |
| VBDS | Ρήμα παρελθοντικού χρόνου, πληθυντικός |
| VBFS | Ρήμα μελλοντικού χρόνου, πληθυντικός |
| MD | Βοηθητικό ρήμα (έχω, είμαι) |
| | **Μετοχές** |

| | |
|---|---|
| VBG | Μετοχή ενεργητικού χρόνου (-οντας, -ώντας) |
| VBP | Μετοχή παροντικού χρόνου |
| VBPD | Μετοχή παρελθοντικού χρόνου |
| VBPF | Μετοχή μελλοντικού χρόνου |
| | **Άκλιτα Μέρη του Λόγου** |
| RB | Επίρρημα |
| IN | Πρόθεση<br>(με, σε, για, ως, προς, κατά, μετά, παρά, αντί, από, δίχως, χωρίς, ίσαμε, δια, εκ, εξ, εν, επί, περί, προ, υπέρ, υπό, συν, μείον, πλην) |
| CC | Σύνδεσμος |
| RP | Μόριο (ας, θα, να, μα, για) |
| UH | Επιφώνημα |
| FW | Ξένη λέξη |
| | **Διάφορα Σύμβολα** |
| DATE | Ημερομηνία |
| TIME | Ώρα |
| AB | Σύντμηση |
| SYM | Σύμβολο |
| . | Τελεία (.) |
| , | Κόμμα (,) |
| : | Άνω-κάτω τελεία (:) |
| ; | Ερωτηματικό (; , ?) |
| ! | Θαυμαστικό (!) |

| | |
|---|---|
| ( | Αριστερή παρένθεση "(" |
| ) | Δεξιά παρένθεση ")" |
| " | Αριστερό εισαγωγικό ( ", «) και Δεξιό εισαγωγικό (", ») |

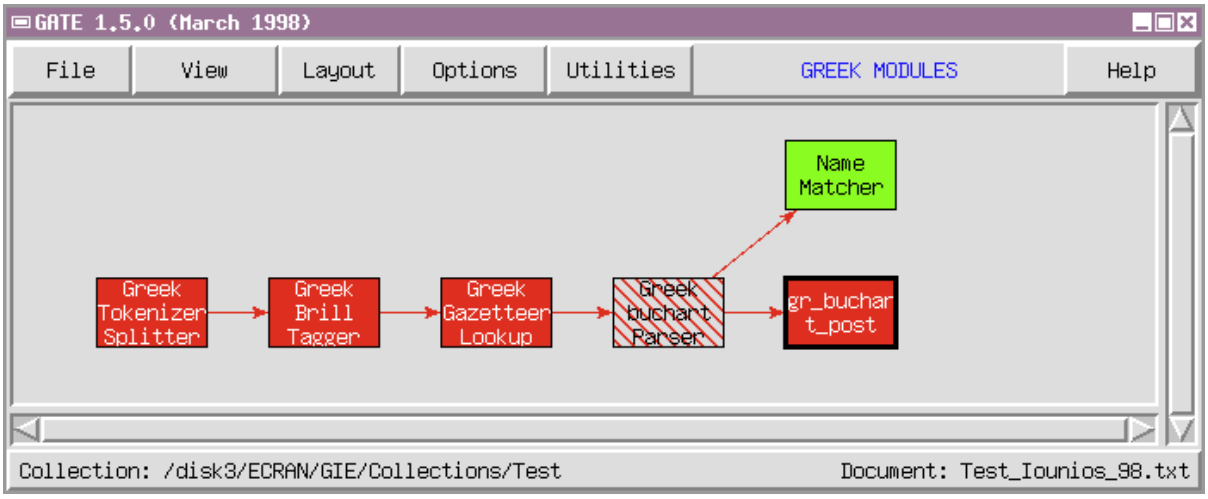
**Πίνακας 48: Το πλήρες σύνολο των ετικετών του αναγνωριστή μερών του λόγου για την Ελληνική γλώσσα.**

# ΠΑΡΑΡΤΗΜΑ ΙΙ

## Αναγνωριστής Ονομάτων Οντοτήτων: χειρωνακτικά κατασκευασμένο σύστημα

Στο παράρτημα αυτό παρουσιάζεται ένα συμβατικό σύστημα αναγνώρισης ονομάτων οντοτήτων, το οποίο κατασκευάστηκε στα πλαίσια αυτής της διατριβής, το οποίο είναι βασισμένο σε κανόνες (γραμματική) χειρωνακτικά κατασκευασμένους. Αν και ένα τέτοιο σύστημα είναι εκτός των στόχων της διατριβής, η ανάπτυξη αυτού του συστήματος βοήθησε στην απόκτηση τεχνογνωσίας κατασκευής ενός αναγνωριστή ονομάτων οντοτήτων για την Ελληνική γλώσσα, στον εντοπισμό των προβλημάτων που σχετίζονται με την χειρωνακτική κατασκευή συμβατικών συστημάτων, ενώ ταυτόχρονα βοήθησε και στην αξιολόγηση ενός συμβατικού συστήματος για την Ελληνική γλώσσα. Το σύστημα που περιγράφεται σε αυτό το παράρτημα αποτέλεσε, αν όχι το πρώτο, ένα από τα πρώτα συστήματα αναγνώρισης ονομάτων οντοτήτων για την Ελληνική γλώσσα.

Το σύστημα αυτό βασίστηκε στο σύστημα VIE του Πανεπιστημίου του Sheffield [84], το οποίο περιλαμβάνει αναγνωριστή λέξεων (*tokeniser*), αναγνωριστή προτάσεων (*sentence splitter*), αναγνωριστή μερών του λόγου (*part of speech tagger*), λεξικό (*gazetteer list*), και ρηχό συντακτικό αναλυτή, με κατάλληλη γραμματική για την αναγνώριση ονομάτων οντοτήτων για την Αγγλική γλώσσα. Όλα τα υποσυστήματα προσαρμόστηκαν (ή αναπτύχθηκαν νέα), έτσι ώστε να μπορούν να επεξεργάζονται κείμενα στα Ελληνικά, ενώ το σύστημα που προέκυψε εμφανίζεται στην Εικόνα 44.




**Εικόνα 44: Ο αναγνωριστής ονομάτων οντοτήτων για την Ελληνική γλώσσα, όπως εμφανίζεται στην πλατφόρμα επεξεργασίας φυσικής γλώσσας GATE. [144]**

Όσον αφορά το στάδιο της προ-επεξεργασίας, η αναγνώριση λέξεων και προτάσεων συγχωνεύτηκαν, σε ένα άρθρωμα (*component*) βασισμένο σε κανονικές γραμματικές (*regular grammars*). Για το υποσύστημα της αναγνώρισης μερών του λόγου, επιστρατεύθηκε μηχανική μάθηση, και συγκεκριμένα μάθηση στηριζόμενη σε κανόνες μετασχηματισμού καθοδηγούμενη από σφάλματα, η οποία περιγράφηκε και αξιολογήθηκε εκτενώς στο κεφάλαιο 3. Όσον αφορά το λεξικό και τις λίστες γνωστών ονομάτων οντοτήτων, αυτές δημιουργήθηκαν από την αρχή με χειρονακτικό τρόπο. Συνολικά συγκεντρώθηκαν 842 ονόματα προσώπων, 475 ονόματα εταιριών, 159 ονόματα τοποθεσιών, 107 ονόματα θέσεων εργασίας, 34 προσδιοριστές ημερομηνιών και 19 προσδιοριστές εταιριών. (Παραδείγματα ονομάτων παρουσιάζονται στον πίνακα:

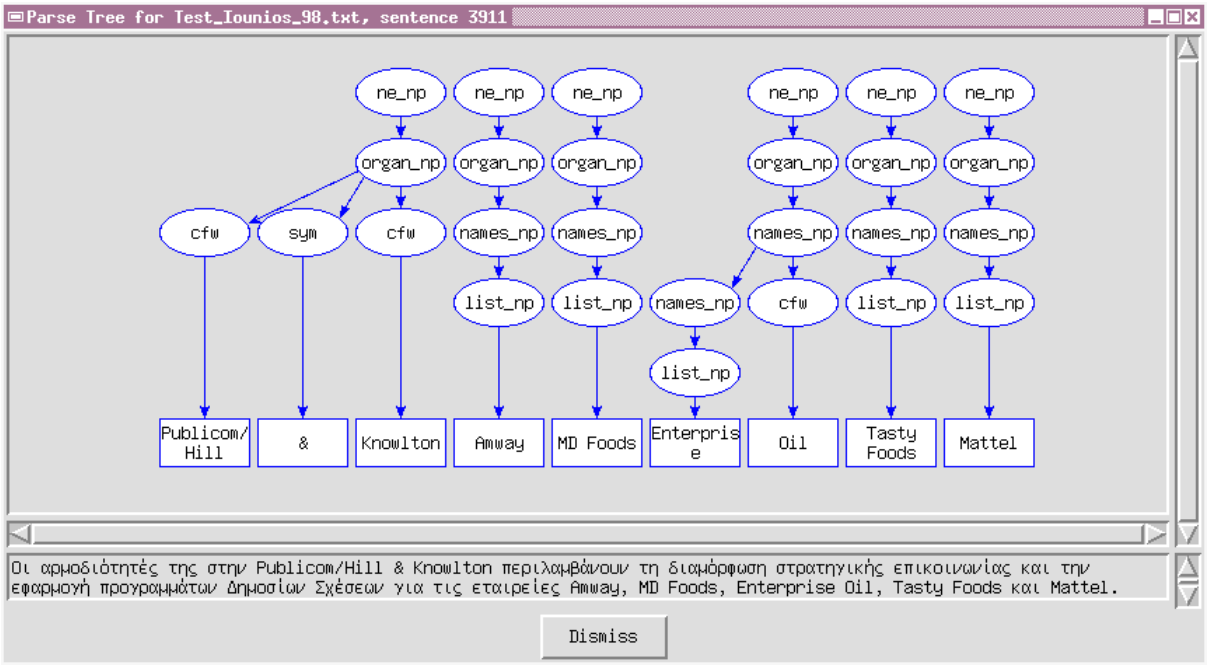
Πίνακας 49) Αξίζει να σημειωθεί ότι όλοι αυτοί οι γλωσσικοί πόροι που δημιουργήθηκαν για τις ανάγκες του αναγνωριστή ονομάτων οντοτήτων για την Ελληνική γλώσσα, διατίθενται ελεύθερα για κάθε χρήση με άδεια χρήσης ανοικτού λογισμικού (LGPL), σαν αρθρώματα της ανοικτού κώδικα πλατφόρμας επεξεργασίας φυσικής γλώσσας «Έλλογον» [53].

| Πρόσωπα | Οργανισμοί | Τοποθεσίες |
|---|---|---|
| Φραγκίσκος | Σκάι 100,4 FM | Ουκρανίας |
| Τράγκας | ΣΚΑΪ 100,4 FM | Ουγγαρίας |
| Τόνια | ΣΚΑΙ | Ουγγαρία |
| Σόφη | Ποπ-Κορν | Ολλανδίας |
| Σόνια | Ποπ&Ροκ | Ολλανδία |
| Σωτήρης | Ποπ Κορν | Νοτίου Ελλάδος |
| Σίσσυ | Μελωδία FM 100 | |
| Σίμος | Μελωδία | |
| Σίλια | ΜΕΛΩΔΙΑ FM 100 | **Θέσεις εργασίας** |
| Σέργιος | Ι.Γ. Δραγούνης & Υιοί ΑΕ | Σύμβουλος Media |
| Λία | Ι.Γ. Δραγούνης & Υιοί | Σύμβουλος Marketing |
| Λένα | Ι.Γ. Δραγούνης | Πρόεδρος Διοικητικού Συμβουλίου |
| Ιφιγένεια | Ι. Γ. Δραγούνης & Υιοί ΑΕ | Διεύθυνση Στρατηγικού Σχεδιασμού |
| Ισίδωρος | Ι. Γ. Δραγούνης | Διεύθυνση Επικοινωνίας |
| Ελισάβετ | Flash 96.1 | Διεύθυνση Διαφήμισης |
| Ελεονόρα | Flash 96,1 FM | Διεύθυνση Marketing |
| Ελεάνας | Flash 96,1 | Senior Product Manager |
| Ελεάνα | Flash 9,61 FM | Senior Media Planner |
| | Flash 9,61 | |

**Πίνακας 49: Ενδεικτικά παραδείγματα από το λεξικό του χειρονακτικού αναγνωριστή ονομάτων οντοτήτων.**

Για τις ανάγκες του χειρονακτικού αναγνωριστή ονομάτων οντοτήτων, χρησιμοποιήθηκε ένας ρηχός συντακτικός αναλυτής σε Prolog των Gazdar και Mellish [145]. Ακολουθώντας μια από κάτω προς τα άνω προσέγγιση (*bottom-up chart parser*) και με την κατάλληλη γραμματική, ο αναλυτής είναι σε θέση να εντοπίσει ονόματα οντοτήτων. Οι πληροφορίες που περιέχονται στην γραμματική περιλαμβάνουν τα μέρη του λόγου και την κατηγοριοποίηση από τις *λίστες γνωστών ονομάτων οντοτήτων* (*gazetteer tags*). Στόχος του αναλυτή δεν είναι η πλήρης συντακτική ανάλυση κάθε πρότασης, αλλά ο εντοπισμός τμημάτων της πρότασης που περιγράφουν οντότητες, μαζί με ένα περιβάλλον λέξεων που περιγράφεται από τους κανόνες. Σαν βάση για την κατασκευή των κανόνων χρησιμοποιήθηκαν οι αντίστοιχοι Αγγλικοί (οι οποίοι στο σύνολό τους ήταν 189), από τους οποίους όμως αρκετοί αφαιρέθηκαν, λόγω διαφορών των δύο γλωσσών. Για την συγγραφή νέων κανόνων, χρησιμοποιήθηκε ένα σώμα κειμένων με θεματική περιοχή τα «επιτυχή γεγονότα διαδοχής διαχείρισης» (*management succession events*), και περιείχε άρθρα ειδήσεων από την Ελληνική εφημερίδα «Διαφημιστική Εβδομάδα» [59] (πρόκειται για το ίδιο σώμα κειμένων που περιγράφηκε στην ενότητα 3.8). Ωστόσο παρατηρήθηκε ότι η συγγραφή κατάλληλων κανόνων είναι σημαντικά δυσκολότερη για την Ελληνική γλώσσα σε σχέση με την Αγγλική, καθώς προσδιοριστές ονομάτων και προσώπων (π.χ. τα Ελληνικά αντίστοιχα των “Mr.”, “Mrs.”, “Ltd.”, “Co.”, κλπ) δεν χρησιμοποιούνται τόσο συχνά στα Ελληνικά, για αυτή την θεματική περιοχή. Η ανυπαρξία προσδιοριστών σε συνδυασμό με την ανυπαρξία αξιόπιστων προτύπων συμφραζόμενων για τα ονόματα οντοτήτων, οδήγησαν στην

αύξηση της σημασίας του λεξικού για την αναγνώριση ονομάτων οντοτήτων, με την γραμματική να διαδραματίζει ρόλο επικύρωσης και συγχώνευσης λέξεων σε ονόματα οντοτήτων. Από το παράδειγμα στην Εικόνα 45, όπου εμφανίζεται η ανάλυση μιας πρότασης, είναι εμφανές ότι όλες οι οντότητες περιέχουν λέξεις οι οποίες αναγνωρίστηκαν από το λεξικό (και συμβολίζονται στην γραμματική σαν “list_np”). Η ισχυρή εξάρτηση από το λεξικό όμως δεν είναι επιθυμητή, αφού η διατήρηση ενός ενημερωμένου λεξικού δεν είναι απλή υπόθεση. Λύση σε αυτή την εξάρτηση δόθηκε με την έμμεση χρήση μηχανικής μάθησης: από την στιγμή που χρησιμοποιείται μηχανική μάθηση στο επίπεδο της αναγνώρισης μερών του λόγου, επεκτάθηκε το σύνολο των κατηγοριών του αναγνωριστή (ΠΑΡΑΡΤΗΜΑ Ι) με πληροφορία σχετική με το κατά πόσο μια λέξης είναι κύριο όνομα, επιτρέποντας τουλάχιστον την αναγνώριση ονομάτων προσώπων, ακόμα και χωρίς να περιέχονται στο λεξικό. Όμως αυτή η μεταφορά κατέδειξε την δυσκολία της χειρωνακτικής κατασκευής συστημάτων αναγνώρισης ονομάτων οντοτήτων, η οποία εμφανίζεται να είναι δυσκολότερη στην Ελληνική από την Αγγλική γλώσσα, καθιστώντας την χρήση μηχανικής μάθησης για την αντιμετώπιση του προβλήματος επιτακτικότερη. Περισσότερες λεπτομέρειες για τον χειρωνακτικά κατασκευασμένο αναγνωριστή ονομάτων οντοτήτων μπορούν να βρεθούν στις εργασίες [58] και [146].




**Εικόνα 45: Ενδεικτικό αποτέλεσμα της εφαρμογής της γραμματικής για τα Ελληνικά σε μια πρόταση.**

### *Πειραματική αξιολόγηση και αποτελέσματα*

Τόσο ο αρχικός αναγνωριστής (για την Αγγλική γλώσσα), όσο και ο προσαρμοσμένος αναγνωριστής (για την Ελληνική γλώσσα), αξιολογήθηκαν σε σώματα κειμένων της ίδιας θεματικής περιοχής. Για την αξιολόγηση του συστήματος VIE [84] για τα Αγγλικά, χρησιμοποιήθηκε ένα σώμα κειμένων από το συνέδριο MUC-6 [8], το οποίο αφορούσε «επιτυχή γεγονότα διαδοχής διαχείρισης» (*management succession events*) ή πιο απλά μετακινήσεις στελεχών επιχειρήσεων. Αντίστοιχα, για την αποτίμηση του συστήματος για τα Ελληνικά, χρησιμοποιήθηκε το σώμα κειμένων από την Ελληνική εφημερίδα «Διαφημιστική Εβδομάδα» [59], το οποίο περιείχε κείμενα της ίδιας θεματικής περιοχής. Το σώμα κειμένων του MUC-6 περιέχει 461 ονόματα οργανισμών και 373 ονόματα

προσώπων, ενώ το αντίστοιχο Ελληνικό περιέχει 425 ονόματα οργανισμών και 262 ονόματα προσώπων. Τα αποτελέσματα της αξιολόγησης φαίνονται στον ακόλουθο πίνακα (Πίνακας 50):

| | Ακρίβεια | Ανάκληση | F-measure |
|---|---|---|---|
| **Αγγλικά - Πρόσωπα** | 92.50 % | 84.97 % | 88.57 % |
| **Αγγλικά - Οργανισμοί** | 83.42 % | 69.25 % | 75.68 % |
| **Ελληνικά - Πρόσωπα** | 88.80 % | 77.00 % | 82.48 % |
| **Ελληνικά - Οργανισμοί** | 57.30 % | 40.40 % | 47.39 % |

**Πίνακας 50: Αποτελέσματα αξιολόγησης του συστήματος VIE [84] και του χειρωνακτικού αναγνωριστή ονομάτων για τα Ελληνικά.**

Τα αποτελέσματα του συστήματος VIE εμφανίζονται σημαντικά χαμηλότερα από τα συνολικά αποτελέσματα που παρουσιάζονται για τα διάφορα συστήματα που συμμετέχουν στα συνέδρια MUC-6 και MUC-7 (Πίνακας 11). Αυτό οφείλεται στη δυσκολία του προσδιορισμού των ονομάτων προσώπων και οργανισμών, όπου σε αυτή την αξιολόγηση τα βλέπουμε απομονωμένα, και όχι συνολικά με ευκολότερους τύπους οντοτήτων, όπως οι τοποθεσίες και οι ημερομηνίες. Και στις δύο γλώσσες, τα αποτελέσματα είναι καλύτερα για τα πρόσωπα από ότι για τους οργανισμούς. Σαφώς η συμβολή του λεξικού είναι καθοριστική για την αναγνώρισή τους, αφού τουλάχιστον το μικρό όνομα ενός προσώπου θα είναι γνωστό, ενώ ταυτόχρονα έχουν και πιο περιορισμένο μήκος σε αριθμό λέξεων (συνήθως δύο λέξεις). Αυτές οι ιδιότητες καθιστούν τον προσδιορισμό τους ευκολότερο από ότι για τα ονόματα οργανισμών, τα οποία ποικίλουν σε μήκος και απαρτίζονται από λέξεις που ανήκουν σε διάφορες κατηγορίες μερών του λόγου. Επίσης, τα αποτελέσματα του Αγγλικού συστήματος VIE είναι υψηλότερα από τα αντίστοιχα του συστήματος για τα Ελληνικά. Αυτό οφείλεται κυρίως στο περιορισμένο μέγεθος του λεξικού (ειδικά για την περίπτωση των οργανισμών), καθώς και στην ύπαρξη διάφορων αγγλικών ονομάτων στα ελληνικά κείμενα. Σημαντικός είναι επίσης και ο ρόλος του πιο περιορισμένου συνόλου γραμματικών κανόνων, το οποίο απαιτεί σημαντική επένδυση χρόνου για να επεκταθεί περεταίρω.

# ΑΝΑΦΟΡΕΣ


[1] P. S. Jacobs and L. F. Rau, "SCISOR: extracting information from on-line news," *Communications of the ACM,* vol. 33, no. 11, pp. 88--97, November 1990.

[2] Y. Wilks, J. Pustejovsky and J. Cowie, "Diderot: TIPSTER program, automatic data extraction from text utilizing semantic analysis," in *Proceedings of the workshop on Human Language Technology (HLT '94)*, Plainsboro, NJ, 1994.

[3] J. Cowie, T. Wakao, L. Guthrie, W. Jin, J. Pustejovsky and S. Waterman, "Diderot Information Extraction System," in *Proceedings of the First Conference of the Pacific Association for Computational Linguistics, (PACLING 93)*, 1993.

[4] F. Vichot, F. Wolinski, J. Tomeh, S. Guennou, B. Dillet and S. Aydjian, "High Precision Hypertext Navigation Based on NLP Automatic Extractions," in *Proceedings of Hypertext, Information Retrieval, Multimedia (HIM'97)*, Dortmund, Germany, 1997.

[5] P. M. Andersen, P. J. Hayes, A. K. Huettner, L. M. Schmandt, I. B. Nirenburg and S. P. Weinstein, "Automatic extraction of facts from press releases to generate news stories," in *Proceedings of the third conference on Applied Natural Language Processing*, Trento, Italy, 1992.

[6] M. Stevenson, R. Basili, G. D. Rossi, P. Velardi and O. Ansaldi, *ECRAN: Extraction of Content: Research at Near-Market,*

[7] "MUC5 '93: Proceedings of the 5th conference on Message understanding," Baltimore, Maryland, USA, 1993.

[8] MUC6 '95: Proceedings of the Sixth Message Understanding Conference, San Francisco, CA: Morgan Kaufmann, 1995.

[9] M. P. Marcus, "Overview of the fifth DARPA speech and natural language workshop (HLT '91)," in *Proceedings of the workshop on Speech and Natural Language*, Harriman, New York, 1992.

[10] *AVENTINUS: advanced information system for multinational drug enforcement,* 2000.

[11] R. Evans and A. F. Hartley, "The traffic information collator," *Expert Systems: The International Journal of Knowledge Engineering,* vol. 7, no. 4, pp. 209--214, 1990.

[12] R. Evans, R. Gaizauskas, L. J. Cahill, J. Walker, J. Richardson and A. Dixon, "POETIC: A System for Gathering and Disseminating Traffic Information," *Journal of Natural Language Engineering,* vol. 1, no. 4, pp. 363-387, 1995.

[13] P. Langley and S. Stromsten, "Learning Context-Free Grammars with a Simplicity Bias," in *Proceedings of the 11th European Conference on Machine Learning (ECML '00)*, 2000.

[14] B. Starkie, F. Coste and M. v. Zaanen, "The Omphalos Context-Free Grammar Learning Competition," in *Grammatical Inference: Algorithms and Applications, 7th International Colloquium (ICGI 2004)*, Athens, Greece, 2004.

[15] E. D. Brill, "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging," *Computational Linguistics,* vol. 21, no. 4, pp. 543--566, 1995.

[16] G. Petasis, S. Petridis, G. Paliouras, V. Karkaletsis, S. J. Perantonis and C. D. Spyropoulos, "Symbolic and neural learning for named-entity recognition," in *Proceedings of the Symposium on Computational Intelligence and Learning (COIL 2000)*, Chios, Greece, June 19-23, 2000.

[17] R. C. Schank and R. P. Abelson, Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures, Lawrence Erlbaum Associates, 1977.

[18] R. C. Schank, J. L. Kolodner and G. DeJong, "Conceptual information retrieval," in *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval (SIGIR '80)*, Cambridge, England, 1980.

[19] MUC7 '98: Proceedings of the Seventh Message Understanding Conference (MUC-7), Morgan Kaufmann, 1998.

[20] R. Grishman, "Information Extraction," in *The Oxford Handbook of Computational Linguistics (Oxford Handbooks in Linguistics S.)*, R. Mitkov, Ed., Oxford University Press, 2003.

[21] D. E. Appelt, J. R. Hobbs, J. Bear, D. J. Israel and M. Tyson, "FASTUS: A Finite-state Processor for Information Extraction from Real-world Text," in *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI 1993)*, Chambéry, France, 1993.

[22] A. G. Valarakos, "Αυξητική Πληθυσμιακή Ενημέρωση Οντολογίας στα πλαίσια της Συντήρησης Οντολογιών," 2009.

[23] T. M. Mitchell, Machine Learning, McGraw-Hill Education (ISE Editions), 1997.

[24] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning,* vol. 1, no. 1, pp. 81--106, March 1986.

[25] J. R. Quinlan, C4.5: programs for machine learning, Morgan Kaufmann Publishers Inc., 1993.

[26] E. D. Brill, "A simple rule-based part-of-speech tagger," in *Proceedings of the third Conference on Applied Natural Language Processing (ANLP'92)*, Trento, Italy, 1992.

[27] E. D. Brill, "A Corpus-based Approach to Language Learning," Philadelphia, 1993.

[28] E. D. Brill, "Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging," in *Proceedings of the 3rd Workshop on Natural Language Processing Using Very Large Corpora*, Massachusetts, USA, 1995.

[29] G. H. John and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers," in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 1995.

[30] D. W. Aha, D. Kibler and M. K. Albert, "Instance-based learning algorithms," *Machine Learning,* vol. 6, pp. 37--66, 1 January 1991.

[31] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Magazine,* pp. 4-15, January 1986.

[32] D. H. Fisher, "Knowledge Acquisition Via Incremental Conceptual Clustering," *Machine Learning,* vol. 2, no. 2, pp. 139-172, 1987.

[33] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," in *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)*, 1994.

[34] P. Cheeseman and J. Stutz, "Bayesian classification (AutoClass): theory and results," in *Advances in knowledge discovery and data mining*, American Association for Artificial Intelligence, 1996, pp. 153--180.

[35] B. B. Greene and G. M. Rubin, "Automatic Grammatical Tagging of English," Providence, Rhode Island,

[36] A. Voutilainen, "A syntax-based part-of-speech analyser," in *EACL '95: Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, Dublin, Ireland, 1995.

[37] R. Weischedel, R. Schwartz, J. Palmucci, M. Meteer and L. Ramshaw, "Coping with ambiguity and unknown words through probabilistic models," *Comput. Linguist.,* vol. 19, no. 2, pp. 361-382, 1993.

[38] T. Brants, "TnT -- A Statistical Part-of-Speech Tagger," in *Proceedings of the Sixth Conference on Applied Natural Language Processing*, Seattle, Washington, USA, 2000.

[39] E. Dermatas and G. Kokkinakis, "Automatic stochastic tagging of natural language texts," *Comput. Linguist.,* vol. 21, no. 2, pp. 137-163, 1995.

[40] H. Schmid, "Part-of-speech tagging with neural networks," in *COLING '94: Proceedings of the 15th conference on Computational linguistics*, Kyoto, Japan, 1994.

[41] R. Rosenfeld, "A Maximum Entropy Approach to Adaptive Statistical Language Modeling," *Computer, Speech and Language,* vol. 10, pp. 187--228, 1996.

[42] A. Ratnaparkhi, "A Maximum Entropy Model for Part-Of-Speech Tagging," in *Proceedings of the First Conference on Empirical Methods in Natural Language Processing*, Philadelphia, 1996.

[43] K. Toutanova and C. D. Manning, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," in *EMNLP '00: Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, Hong Kong, 2000.

[44] J. R. Curran and S. Clark, "Investigating GIS and smoothing for maximum entropy taggers," in *EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, Budapest, Hungary, 2003.

[45] H. Huang and X. Zhang, "Part-of-speech tagger based on maximum entropy model," *Computer Science and Information Technology, International Conference on,* vol. 0, pp. 26-29, 2009.

[46] W. Daelemans, J. Zavrel, P. Berck and S. Gillis, "MBT: A Memory-Based Part of Speech Tagger-Generator," in *Proceedings of the fourth Workshop on Very Large Corpora, ACL SIGDAT*, Copenhagen, Denmark, 1996.

[47] D. Hindle, "Acquiring disambiguation rules from text," in *ACL '89: Proceedings of the 27th annual meeting on Association for Computational Linguistics*, Vancouver, British Columbia, Canada, 1989.

[48] D. Yarowsky, "Decision lists for lexical ambiguity resolution: application to accent restoration in Spanish and French," in *ACL '94: Proceedings of the 32nd annual meeting on Association for*

*Computational Linguistics*, Las Cruces, New Mexico, 1994.

[49] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees," in *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

[50] G. S. Orphanos and D. N. Christodoulakis, "POS disambiguation and unknown word guessing with decision trees," in *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, Bergen, Norway, 1999.

[51] O. G. Kalles, K. Dimitris, P. Thanasis and C. Dimitris, "Decision Trees and NLP: A Case Study in POS Tagging," in *Proceedings of the ECCAI Advanced Course on Artificial Intelligence ACAI'99*, Chania, Greece,

[52] Π. Μαλακασιώτης, "Αναγνώριση μερών του λόγου σε ελληνικά κείμενα με τεχνικές ενεργητικής μάθησης," Αθήνα, Ελλάδα, 2005.

[53] G. Petasis, V. Karkaletsis, G. Paliouras, I. Androutsopoulos and C. D. Spyropoulos, "Ellogon: A New Text Engineering Platform," in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, 2002.

[54] D. Spiliotopoulos, G. Petasis and G. Kouroupetroglou, "Prosodically Enriched Text Annotation for High Quality Speech Synthesis," in *Proceedings of the 10th International Conference on Speech and Computer (SPECOM-2005)*, Patras, Greece, 2005.

[55] H. Papageorgiou, P. Prokopidis, V. Giouli and S. Piperidis, "A Unified POS Tagging Architecture and its Application to Greek," in *Proceedings of the 2nd Language Resources and Evaluation Conference (LREC 2000)*, Athens, Greece, 2000.

[56] A. Bies, M. Ferguson, K. Katz and R. MacIntyre, "Bracketing Guidelines for Treebank II Style Penn Treebank Project," 1995.

[57] P. Labropoulou, E. Mantzari and M. Gavrilidou, "Lexicon - Morphosyntactic Specifications: Language Specific Instantiation (Greek)," 1996.

[58] V. Karkaletsis, C. D. Spyropoulos and G. Petasis, "Named Entity Recognition from Greek texts: the GIE Project". In "Advances in Intelligent Systems: Concepts, Tools and Applications," in *Advances in intelligent systems: concepts, tools and applications*, Athens, Kluwer Academic Publishers, 1999, p. 131 – 142.

[59] 1998. [Online]. Available: http://www.adweek.gr.

[60] G. Petasis, V. Karkaletsis, D. Farmakiotou, I. Androutsopoulos and C. D. Spyropoulos, "A Greek morphological lexicon and its exploitation by natural language processing applications," in *Advances in Informatics - Post-proceedings of the 8th Panhellenic Conference in Informatics, Lecture Notes on Computer Science (LNCS), vol. 2563*, Springer-Verlag, 2003, pp. 401--419.

[61] G. Petasis, V. Karkaletsis, D. Farmakiotou, G. Samaritakis, I. Androutsopoulos and C. D. Spyropoulos, "A Greek Morphological Lexicon And Its Exploitation By A Greek Controlled Language Checker," in *Proceedings of the 8th Panhellenic Conference on Informatics*, Nicosia, Cyprus, 2001.

[62] I. Androutsopoulos, V. Kokkinaki, A. Dimitromanolaki, J. Calder, J. Oberlander and E. Not, "Generating Multilingual Personalized Descriptions of Museum Exhibits - The M-PIRO Project," in *Proceedings of the International Conference on Computer Applications and Quantitative Methods in Archaeology*, 2001.

[63] D. E. Spiliotopoulos, "ΒΕΛΤΙΩΣΗ ΠΟΙΟΤΗΤΑΣ ΣΥΝΘΕΤΙΚΗΣ ΟΜΙΛΙΑΣ ΜΕΣΩ ΠΡΟΣΩΔΙΑΚΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ," Athens, 2009.

[64] D. Harman, "The DARPA TIPSTER project," *SIGIR Forum,* vol. 26, no. 2, pp. 26--28, October 1992.

[65] N. A. Chinchor, "Overview of MUC-7/MET-2," in *In Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

[66] R. Grishman, "The NYU system for MUC-6 or where's the syntax?," in *Proceedings of the 6th conference on Message understanding (MUC-6 '95)*, Columbia, Maryland, USA, 1995.

[67] A. E. Borthwick, "A maximum entropy approach to named entity recognition," New York University, New York, NY, USA, 1999.

[68] G. R. Krupka and K. Hausman, "IsoQuest: Description of the NetOwl extractor system as used in MUC-7," in *In Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

[69] W. Black, F. Rinaldi and D. Mowatt, "FACILE: Description of the NE System Used for MUC-7," in *In Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

[70] S. Sekine, "NYU: Description of the Japanese NE system used for MET-2," in *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

[71] Y. Tsuruoka, J. Tsujii and S. Ananiadou, "Accelerating the annotation of sparse named entities by

dynamic sentence selection," *BMC Bioinformatics,* vol. 9, p. S8, 2008.

[72] D. M. Bikel, S. Miller, R. Schwartz and R. Weischedel, "Nymble: a high-performance learning name-finder," in *Proceedings of the fifth conference on Applied Natural Language Processing*, 1997.

[73] A. Mikheev, C. Grover and M. Moens, "Description of the LTG system used for MUC-7," in *In Proceedings of 7th Message Understanding Conference (MUC-7)*, 1998.

[74] M. Vilain and D. Day, "Finite-state phrase parsing by rule sequences," in *Proceedings of the 16th conference on Computational linguistics (COLING '96) - Volume 1*, Copenhagen, Denmark, 1996.

[75] J. Aberdeen, J. Burger, D. Day, L. Hirschman, P. Robinson and M. Vilain, "MITRE: description of the Alembic system used for MUC-6," in *Proceedings of the 6th conference on Message understanding*, Columbia, Maryland, 1995.

[76] S. W. Bennett, C. Aone and C. Lovell, "Learning to Tag Multilingual Texts Through Observation," in *In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, 1997.

[77] J. Cowie, "CRL/NMSU: description of the CRL/NMSU systems used for MUC-6," in *Proceedings of the 6th conference on Message understanding MUC-6*, Columbia, Maryland, 1995.

[78] A. Cucchiarelli and P. Velardi, "Finding a domain-appropriate sense inventory for semantically tagging a corpus," *Natural Language Engineering,* vol. 4, no. 4, pp. 325--344, December 1998.

[79] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Comput. Linguist.,* vol. 16, no. 1, pp. 22--29, March 1990.

[80] P. v. d. Smagt and F. Groen, "Approximation with neural networks: Between local and global approximation," in *Proceedings of the 1995 IEEE International Conference on Neural Networks*, Perth, Western Australia, 1995.

[81] S. J. Perantonis, N. Ampazis and V. Virvilis, "A Learning Framework for Neural Networks Using Constrained Optimization Methods," *Annals of Operations Research,* vol. 99, no. 1, pp. 385-401, 2000.

[82] D. A. Karras and S. J. Perantonis, "An efficient constrained training algorithm for feedforward networks," *IEEE Transactions on Neural Networks,* vol. 6, pp. 1420-1454, 1995.

[83] G. Petasis, S. Petridis, G. Paliouras, V. Karkaletsis, S. J. Perantonis and C. D. Spyropoulos, "Symbolic and Neural Learning of Named-Entity Recognition and Classification Systems in Two Languages," in *Advances in Computational Intelligence and Learning: Methods and Applications*, 2002.

[84] K. Humphreys, R. Gaizauskas, H. Cunningham and S. Azzam, "VIE Technical Specifications," Sheffield, UK, 1997.

[85] L. A. Ramshaw and M. P. Marcus, "Text Chunking using transformation-based learning," in *Proceedings of the Third Annual Workshop on Very Large Corpora, in 33rd Annual Meeting of the Association for Computational Linguistics (ACL 1995)*, Cambridge, Massachusetts, USA, 1995.

[86] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory,* vol. 13, no. 2, pp. 260--269, 1967.

[87] V. Karkaletsis, C. D. Spyropoulos, C. Grover, M. T. Pazienza, J. Coch and D. Souflis, "A Platform for Cross-Lingual, Domain and User Adaptive Web Information Extraction," in *European Conference on Artificial Intelligence - ECAI 2004*, Valencia, Spain, 2004.

[88] V. Karkaletsis and C. D. Spyropoulos, "Information Retrieval and Extraction from the Web: the CROSSMARC approach.," in *Proceedings of the RIAO 2004 Conference "Coupling approaches, coupling media and coupling languages for information retrieval"*, Avignon (Vaucluse), France, 2004.

[89] J. R. Curran and S. Clark, "Language independent NER using a maximum entropy tagger," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4 (CONLL '03)*, Edmonton, Canada, 2003.

[90] E. Charniak, D. Blaheta, N. Ge, K. Hall, J. Hale and M. Johnson, "BLLIP 1987-89 WSJ Corpus Release 1," 1987. [Online]. Available: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2000T43.

[91] F. Vichot, F. Wolinski, H. C. Ferri and D. Urbani, "Using Information Extraction for Knowledge Entering," in *Advances in Intelligent Systems - Concepts, Tools and Applications*, Dordrecht, Kluwer academic publishers, 1999, pp. 191--200.

[92] F. Wolinski, F. Vichot and M. Stricker, "Using Learning-based Filters to Detect Rule-based Filtering Obsolescence," in *Proceedings of the 6th International Conference on Computer-Assisted*

*Information Retrieval (Recherche dInformation et ses Applications) - RIAO 2000*, College de France, France, April 12-14, 2000.

[93] G. J. Wolff, "Grammar Discovery as Data Compression," in *Proceedings of the AISB/GI Conference on Artificial Intelligence*, Hamburg, Germany, 1978.

[94] G. J. Wolff, "Language acquisition, data compression and generalization," *Language and Communication,* vol. 2, pp. 57--89, 1982.

[95] M. E. Gold, "Language Identification in the Limit," *Information and Control,* vol. 10, no. 5, pp. 447--474, 1967.

[96] L. G. Valiant, "A theory of the learnable," *Communications of ACM,* vol. 27, no. 11, pp. 1134--1142, November 1984.

[97] D. Angluin and M. Kharitonov, "When wont membership queries help?," *J. Comput. Syst. Sci.,* vol. 50, no. 2, pp. 336--355, April 1995.

[98] F. Denis, "Learning Regular Languages from Simple Positive Examples," *Machine Learning,* vol. 44, no. 1-2, pp. 37--66, July 2001.

[99] F. Denis, "PAC Learning from Positive Statistical Queries," in *Proceedings of the 9th International Conference on Algorithmic Learning Theory (ALT '98)*, 1998.

[100] D. Angluin, "Inference of Reversible Languages," *J. ACM,* vol. 29, no. 3, pp. 741--765, July 1982.

[101] P. Garcia and E. Vidal, "Inference of k-Testable Languages in the Strict Sense and Application to Syntactic Pattern Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 12, pp. 920-925, 1990.

[102] J. D. Emerald, K. G. Subramanian and D. G. Thomas, "Learning code regular and code linear languages," in *Proceedings of the 3rd International Colloquium on Grammatical Inference: Learning Syntax from Sentences*, 1996.

[103] T. Koshiba, E. Mäkinen and Y. Takada, "Inferring pure context-free languages from positive data," *Acta Cybern.,* vol. 14, no. 3, pp. 469--477, June 2000.

[104] N. Tanida and T. Yokomori, "Inductive Inference of Monogenic Pure Context-free Languages," in *Proceedings of the 4th International Workshop on Analogical and Inductive Inference: Algorithmic Learning Theory (AII '94)*, 1994.

[105] T. Yokomori, "On Polynomial-Time Learnability in the Limit of Strictly Deterministic Automata," *Machine Learning,* vol. 19, pp. 153-179, 1995.

[106] H. Rulot and E. Vidal, "Modelling (sub)string-length based constraints through a grammatical inference method," in *Proceedings of the NATO Advanced Study Institute on Pattern recognition theory and applications*, Spa-Balmoral, Belgium, 1987.

[107] A. Stolcke, "Bayesian learning of probabilistic language models," University of California at Berkeley, Berkeley, CA, USA, 1994.

[108] A. Stolcke and S. M. Omohundro, "Inducing Probabilistic Grammars by Bayesian Model Merging," in *Proceedings of the Second International Colloquium on Grammatical Inference and Applications (ICGI '94)*, 1994.

[109] Y. Sakakibara, "Efficient learning of context-free grammars from positive structural examples," *Information and Computation,* vol. 97, no. 1, pp. 23--60, March 1992.

[110] J. E. Hopcroft, R. Motwani and J. D. Ullman, Introduction to Automata Theory, Languages, and Computation, 3rd Edition ed., Addison-Wesley Longman Publishing Co., Inc., 2006.

[111] E. Mäkinen, "Remarks on the structural grammatical inference problem for context-free grammars," *Information Processing Letters,* vol. 44, no. 3, pp. 125--127, November 1992.

[112] Y. Sakakibara and H. Muramatsu, "Learning Context-Free Grammars from Partially Structured Examples," in *Proceedings of the 5th International Colloquium on Grammatical Inference: Algorithms and Applications (ICGI 2000)*, 2000.

[113] K. Nakamura and T. Ishiwata, "Synthesizing Context Free Grammars from Sample Strings Based on Inductive CYK Algorithm," in *Proceedings of the 5th International Colloquium on Grammatical Inference: Algorithms and Applications (ICGI 2000)*, 2000.

[114] F. Nevado, J.-A. Sánchez and J.-M. Benedí, "Combination of Estimation Algorithms and Grammatical Inference Techniques to Learn Stochastic Context-Free Grammars," in *Proceedings of the 5th International Colloquium on Grammatical Inference: Algorithms and Applications (ICGI 2000)*, 2000.

[115] M. P. Marcus, M. A. Marcinkiewicz and B. Santorini, "Building a large annotated corpus of English: the penn treebank," *Computational Linguistics - Special issue on using large corpora: II,* vol. 19, no.

2, pp. 313--330, June 1993.

[116] D. Freitag, "Using Grammatical Inference to Improve Precision in Information Extraction," in *In ICML-97 Workshop on Automation Induction, Grammatical Inference, and Language Acquisition*, 1997.

[117] H. Ahonen, H. Mannila and E. Nikunen, "Forming Grammars for Structured Documents: an Application of Grammatical Inference," in *Proceedings of the Second International Colloquium on Grammatical Inference and Applications (ICGI '94)*, 1994.

[118] T. Goan, N. Benson and O. Etzioni, "A Grammar Inference Algorithm for the World Wide Web," in *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*, 1996.

[119] M. van Zaanen, "ABL: alignment-based learning," in *Proceedings of the 18th conference on Computational linguistics (COLING-2000) - Volume 2*, 2000.

[120] M. v. Zaanen and P. Adriaans, "Comparing Two Unsupervised Grammar Induction Systems: Alignment-Based Learning vs. EMILE," 2001.

[121] A. Clark, "Unsupervised induction of stochastic context-free grammars using distributional clustering," in *Proceedings of the 2001 workshop on Computational Natural Language Learning (ConLL '01) - Volume 7*, Toulouse, France, 2001.

[122] K. Vanlehn and W. Ball, "A Version Space Approach to Learning Context-free Grammars," *Machine Learning,* vol. 2, no. 1, pp. 39--74, March 1987.

[123] J. Rissanen, Stochastic Complexity in Statistical Inquiry, World Scientific Publishing Co., Inc., 1989.

[124] R. Parekh and V. Honavar, "Grammar Inference, Automata Induction, and Language Acquisition," in *Handbook of Natural Language Processing*, Marcel Dekker, 2000, pp. 727--764.

[125] G. A. Miller, M. Chodorow, S. Landes, C. Leacock and R. G. Thomas, "Using a semantic concordance for sense identification," in *Proceedings of the workshop on Human Language Technology (HLT '94)*, Plainsboro, NJ, 1994.

[126] G. Paliouras and Y. Sakakibara, Grammatical inference: algorithms and applications : 7th international colloquium, ICGI 2004, Athens, Greece, October 11-13, 2004, Springer, 2004.

[127] A. Clark, "Learning Deterministic Context Free Grammars: the Omphalos Competition," *Machine Learning,* vol. 66, pp. 93-110, January 2007.

[128] S. Katrenko and P. Adriaans, "Learning Relations from Biomedical Corpora Using Dependency Tree Levels," in *Proceedings of the Fifteenth Dutch-Belgian Conference on Machine Learning (Benelearn)*, Ghent, Belgium, 2006.

[129] M. Surdeanu, R. Johansson, A. Meyers, L. Màrquez and J. Nivre, "The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies," in *Proceedings of the Twelfth Conference on Natural Language Learning (CoNLL 2008)*, 2008.

[130] X. Carreras and L. Màrquez, "Introduction to the CoNLL-2005 shared task: semantic role labeling," in *Proceedings of the Ninth Conference on Computational Natural Language Learning (CONLL '05)*, Ann Arbor, Michigan, 2005.

[131] X. Carreras and L. Màrques, "Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling," in *Proceedings of CoNLL-2004*, Boston, MA, USA, 2004.

[132] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proceedings of the 14th Conference on Computational Linguistics (ACL 1992) - Volume 2*, Nantes, France, 1992.

[133] D. Davidov, A. Rappoport and M. Koppel, "Fully unsupervised discovery of concept-specific relationships by web mining," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL '07)*, Prague, Czech Republic, 2007.

[134] S. Brody, "Clustering Clauses for High-Level Relation Detection: An Information-theoretic Approach," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, 2007.

[135] R. C. Bunescu, "Learning to extract relations from the web using minimal supervision," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, Prague, Czech Republic, 2007.

[136] F. Xu, H. Uszkoreit and H. Li, "A seed-driven bottom-up machine learning framework for extracting relations of various complexity," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, Prague, Czech Republic, 2007.

[137] B. Rosenfeld and R. Feldman, "Using Corpus Statistics on Entities to Improve Semi-supervised Relation Extraction from the Web," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL '07)*, Prague, Czech Republic, 2007.

[138] J. Pustejovsky, J. Castafio, J. Zhang, M. Kotecki and B. Cochran, "Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations," in *In Proceedings of the Pacific Symposium on Biocomputing*, 2002.

[139] G. Leroy and H. Chen, "Genescene: An ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical texts: Research Articles," *J. Am. Soc. Inf. Sci. Technol.,* vol. 56, no. 5, pp. 457--468, March 2005.

[140] F. Ciravegna, "Adaptive information extraction from text by rule induction and generalisation," in *Proceedings of the 17th international joint conference on Artificial intelligence (IJCAI 2001) - Volume 2*, Seattle, WA, USA, 2001.

[141] F. Ciravegna and A. Lavelli, "LearningPinocchio: adaptive information extraction for real world applications," *Natural Language Engineering,* vol. 10, no. 2, pp. 145--165, June 2004.

[142] E. Niebler, *Boost.Xpressive template library for C++,* 2007.

[143] J. D. Lafferty, A. McCallum and F. C. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*, 2001.

[144] H. Cunningham, Y. Wilks and R. J. Gaizauskas, "GATE - a General Architecture for Text Engineering," in *Proceedings of 16th Conference on Computational Linguistics (COLING '96)*,

[145] G. Gazdar and C. Mellish, Natural Language Processing in PROLOG, Addison-Wesley, 1989.

[146] V. Karkaletsis, G. Paliouras, G. Petasis, N. Manousopoulou and C. D. Spyropoulos, "Named-Entity Recognition from Greek and English Texts," *Journal of Intelligent and Robotic Systems,* vol. 26, no. 2, pp. 123--135, October 1999.