

**ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΓΝΩΡΙΣΗ
ΕΙΔΟΥΣ ΚΕΙΜΕΝΟΥ ΚΑΙ ΣΥΓΓΡΑΦΕΑ
ΣΕ ΝΕΟΕΛΛΗΝΙΚΑ ΚΕΙΜΕΝΑ
ΧΩΡΙΣ ΠΕΡΙΟΡΙΣΜΟΥΣ**

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΕΥΣΤΑΘΙΟΥ Σ. ΣΤΑΜΑΤΑΤΟΥ

ΔΙΠΛ. ΗΛΕΚΤΡΟΛΟΓΟΥ ΜΗΧΑΝΙΚΟΥ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΡΙΘΜΟΣ ΔΙΑΤΡΙΒΗΣ 87

ΑΠΡΙΛΙΟΣ 2000

Περιεχόμενα

<i>Κατάλογος Σχημάτων</i>	<i>v</i>
<i>Κατάλογος Πινάκων</i>	<i>vii</i>
<i>Κατάλογος Συντομογραφιών</i>	<i>ix</i>
<i>Περίληψη</i>	<i>xi</i>
<i>Summary</i>	<i>xiii</i>
1 Εισαγωγή	1
1.1 <i>Τάσεις στη Σύγχρονη Υπολογιστική Γλωσσολογία</i>	<i>1</i>
1.2 <i>Γλώσσα και Ύφος</i>	<i>3</i>
1.2.1 <i>Το ύφος ως απόκλιση</i>	<i>5</i>
1.2.2 <i>Το ύφος ως επιλογή</i>	<i>6</i>
1.2.3 <i>Στατιστική υφολογία</i>	<i>6</i>
1.3 <i>Το Ύφος στην Επεξεργασία Φυσικής Γλώσσας</i>	<i>7</i>
1.4 <i>Προδιαγραφές - Καινοτομίες της Διατριβής</i>	<i>9</i>
1.5 <i>Διάρθρωση της Διατριβής</i>	<i>12</i>
2 Ανίχνευση Ορίων Περιόδων	15
2.1 <i>Εισαγωγή</i>	<i>15</i>

2.2 Σχετική Έρευνα	17
2.3 Μεθοδολογία	19
2.3.1 Απλές παράμετροι.....	20
2.3.2 Εκμάθηση Βάσει Μετασχηματισμών	23
2.3.3 Διαδικασία αποσαφήνισης.....	25
2.3.4 Αυτόματη εξαγωγή κανόνων	26
2.4 Αξιολόγηση.....	28
2.4.1 Σώματα κειμένων.....	28
2.4.2 Απόδοση	29
2.4.3 Σύγκριση με τη θεωρία EBM	31
2.5 Περίληψη - Συμπεράσματα.....	32
3 Ανίχνευση Ορίων Φράσεων	35
3.1 Εισαγωγή.....	35
3.2 Σχετική Έρευνα	38
3.3 Εκτίμηση Μορφολογικής Πληροφορίας	40
3.4 Ανάλυση Πολλαπλού Περάσματος.....	44
3.5 Αξιολόγηση.....	47
3.6 Προσέγγιση Βασιζόμενη σε Λεξικό.....	50
3.7 Περίληψη - Συμπεράσματα.....	52
4 Εξαγωγή Υφολογικών Δεικτών	55
4.1 Τάσεις στην Σύγχρονη Υφομετρία.....	55
4.2 Προηγούμενες Υφομετρικές Προτάσεις	58
4.2.1 Επίπεδο δείγματος	58
4.2.2 Συντακτικός σχολιασμός	59
4.2.3 Πλούτος λεξιλογίου	60
4.2.4 Συχνότητα λέξεων.....	61
4.3 Η Πρότασή μας	62
4.4 Περίληψη - Συμπεράσματα.....	66
5 Αναγνώριση Είδους Κειμένου	69
5.1 Εισαγωγή.....	69

5.2 Μέθοδοι Κατηγοριοποίησης.....	72
5.2.1 Πολλαπλή παλινδρόμηση	73
5.2.2 Διαχωριστική ανάλυση	76
5.3 Λεξιλογικές Υφομετρικές Προσεγγίσεις	78
5.4 Σώμα Κειμένων ανά Είδος.....	79
5.5 Αξιολόγηση.....	81
5.5.1 Μέγεθος σώματος εκπαίδευσης.....	85
5.5.2 Σημαντικότητα των υφολογικών δεικτών.....	86
5.5.3 Ελαττωματική ανάλυση	88
5.6 Περίληψη - Συμπεράσματα.....	89
6 Προσδιορισμός Συγγραφέα	93
6.1 Εισαγωγή.....	93
6.2 Σώμα Κειμένων ανά Συγγραφέα	95
6.3 Αναγνώριση Συγγραφέα.....	97
6.3.1 Μέγεθος σώματος εκπαίδευσης.....	101
6.3.2 Συνδυασμός με λεξιλογική προσέγγιση.....	103
6.3.3 Σημαντικότητα υφολογικών δεικτών.....	107
6.4 Επιβεβαίωση Συγγραφέα.....	109
6.5 Περίληψη - Συμπεράσματα.....	113
7 Συμπεράσματα - Προοπτικές.....	117
7.1 Συμβολή της Διατριβής.....	117
7.2 Προοπτικές.....	122
Παράρτημα Α: Κανόνες Ανίχνευσης Φράσεων.....	125
Παράρτημα Β: Παράδειγμα Ανάλυσης Κειμένου.....	131
Βιβλιογραφία	135
Δημοσιεύσεις.....	145

Κατάλογος Σχημάτων

Σχήμα 1.1. Η προτεινόμενη λύση.	10
Σχήμα 2.1. Ένα παράδειγμα των χαρακτηριστικών που χρησιμοποιούμε.	22
Σχήμα 2.2. Ένα παράδειγμα των παραμέτρων.	22
Σχήμα 2.3. Η διαδικασία εκμάθησης σύμφωνα με την θεωρία EBM.	24
Σχήμα 2.4. Ένα παράδειγμα περιβάλλοντος.	25
Σχήμα 2.5. Δομή της διαδικασίας αποσαφήνισης.	26
Σχήμα 3.1. Δομή του συστήματος ανίχνευσης ορίων φράσεων.	37
Σχήμα 3.2. Κατανομή των καταλήξεων συναρτήσει του μήκους τους.	42
Σχήμα 3.3. Κατανομή των καταλήξεων συναρτήσει των μορφολογικών περιγραφών που υποδηλώνουν.	43
Σχήμα 3.4. Καταλήξεις μιας μορφολογικής περιγραφής ανά μέρος-του-λόγου.	43
Σχήμα 3.5. Ανάλυση δείγματος κειμένου μέσω πολλαπλού περάσματος.	48
Σχήμα 3.6. Δομή του συστήματος ανίχνευσης ορίων φράσεων βάσει λεξικού.	50
Σχήμα 3.7. Μορφολογική ανάλυση του σώματος ελέγχου από την προσέγγιση με χρήση λεξικού.	51
Σχήμα 4.1. Τα τρία επίπεδα υφολογικών δεικτών.	65
Σχήμα 4.2. Ανάλυση δείγματος κειμένου από τον ανιχνευτή ορίων περιόδων και φράσεων.	66
Σχήμα 5.1. Διαδικασία κατηγοριοποίησης.	73
Σχήμα 5.2. Το σώμα ελέγχου στον χώρο των δύο πρώτων κυρίων συνιστωσών.	82
Σχήμα 5.3. Συγκριτικά αποτελέσματα για την αναγνώριση είδους κειμένου.	82
Σχήμα 5.4. Κατανομή του σώματος ελέγχου σύμφωνα με την ακρίβεια κατηγοριοποίησης και το μήκος κειμένου.	84

Σχήμα 5.5. Το σφάλμα αναγνώρισης συναρτήσει του μεγέθους του σώματος εκπαίδευσης.....	86
Σχήμα 6.1. Αναπαράσταση του σώματος ελέγχου της ομάδας A στο χώρο των δύο πρώτων κυρίων συνιστωσών.	98
Σχήμα 6.2. Αναπαράσταση του σώματος ελέγχου της ομάδας B στο χώρο των δύο πρώτων κυρίων συνιστωσών.	98
Σχήμα 6.3. Συγκριτικά αποτελέσματα για την αναγνώριση συγγραφέα στην ομάδα A.	99
Σχήμα 6.4. Συγκριτικά αποτελέσματα για την αναγνώριση συγγραφέα στην ομάδα B.	99
Σχήμα 6.5. Κατανομή του σώματος ελέγχου και των δύο ομάδων συναρτήσει του μήκους τους και της ακρίβειας ταξινόμησης.	101
Σχήμα 6.6. Η ακρίβεια ταξινόμησης συναρτήσει του μεγέθους σώματος εκπαίδευσης της ομάδας B.....	102
Σχήμα 6.7. Ακρίβεια ταξινόμησης της λεξιλογικής προσέγγισης συναρτήσει του αριθμού των πιο συχνά εμφανιζόμενων λέξεων.	105
Σχήμα 6.8. Συγκριτικά αποτελέσματα ακρίβειας ταξινόμησης συναρτήσει του μεγέθους σώματος εκπαίδευσης.	107
Σχήμα 6.9. Ακρίβεια ταξινόμησης για κάθε επίπεδο ξεχωριστά.....	109
Σχήμα 6.10. Μέσος όρος σφάλματος απόρριψης, σφάλματος αποδοχής και μέσου σφάλματος συναρτήσει του κατωφλιού για την ομάδα A.	112
Σχήμα 6.11. Μέσος όρος σφάλματος απόρριψης, σφάλματος αποδοχής και μέσου σφάλματος συναρτήσει του κατωφλιού για την ομάδα B.....	112

Κατάλογος Πινάκων

Πίνακας 2.1. Οι τύποι των χαρακτήρων.....	21
Πίνακας 2.2. Τα σώματα κειμένων.....	29
Πίνακας 2.3. Η απόδοση του ανιχνευτή ορίων περιόδων.	30
Πίνακας 2.4. Αναλυτικά αποτελέσματα για το σώμα κειμένων του <i>Βήματος</i>	30
Πίνακας 2.5. Ανάλυση της χρησιμότητας των προτεινόμενων παραμέτρων.....	30
Πίνακας 2.6. Σύγκριση των δύο αλγορίθμων εκμάθησης.	32
Πίνακας 3.1. Η απόδοση του ανιχνευτή ορίων φράσεων.....	49
Πίνακας 3.2. Η απόδοση του ανιχνευτή ορίων φράσεων με χρήση λεξικού.	51
Πίνακας 4.1. Οι τιμές των υφολογικών δεικτών για το δείγμα κειμένου του σχήματος 4.2.....	66
Πίνακας 5.1. Το σώμα κειμένων ανά είδος.	80
Πίνακας 5.2. Τα αποτελέσματα αναγνώρισης είδους κειμένου.	83
Πίνακας 5.3. Κατηγοριοποίηση του σώματος ελέγχου με βάση τη διαχωριστική ανάλυση.....	83
Πίνακας 5.4. Μέσες τιμές t των συντελεστών παλινδρόμησης.....	87
Πίνακας 5.5. Συγκριτικές τιμές του μέσου όρου του απόλυτου t για κανονική και ελαττωματική ανάλυση.	88
Πίνακας 5.6. Συγκριτικά αποτελέσματα συστημάτων αναγνώρισης είδους κειμένου.....	89
Πίνακας 6.1. Η δομή της εφημερίδας <i>Το Βήμα</i>	95
Πίνακας 6.2. Το σώμα κειμένων ανά συγγραφέα.....	96
Πίνακας 6.3. Αποτελέσματα αυτόματης αναγνώρισης συγγραφέα.....	100
Πίνακας 6.4. Πίνακας σύγχυσης της ομάδας B με βάση 20 κείμενα από κάθε συγγραφέα ως σώμα εκπαίδευσης.	102

Πίνακας 6.5. Οι 50 πιο συχνά εμφανιζόμενες λέξεις του σώματος εκπαίδευσης της ομάδας B.	103
Πίνακας 6.6. Πίνακας σύγκρισης της ομάδας B βάσει της λεξιλογικής προσέγγισης.	104
Πίνακας 6.7. Πίνακας σύγκρισης της ομάδας B βάσει του συνδυασμού της δικής μας μεθόδου και της λεξιλογικής προσέγγισης.	106
Πίνακας 6.8. Μέσες τιμές t των συντελεστών παλινδρόμησης συνολικά για τις ομάδες A και B.	108
Πίνακας 6.9. Αποτελέσματα επιβεβαίωσης συγγραφέα για τις ομάδες A και B (κατώφλι= $R/2$).	111

Κατάλογος Συντομογραφιών

ΕΒΜ:	Εκμάθηση Βάσει Μετασχηματισμών
ΕΛ:	Επόμενη Λέξη
ΕΣΣ:	Επόμενα Σημεία Στίξης
ΕΦ:	Επιρρηματικές Φράσεις
ΟΦ:	Ονοματικές Φράσεις
ΠΛ:	Προηγούμενη Λέξη
ΠΣΣ:	Προηγούμενα Σημεία Στίξης
ΠΦ:	Προθετικές Φράσεις
ΡΦ:	Ρηματικές Φράσεις
ΣΥΟΠ:	Σύνολο Υποψηφίων Ορίων Περιόδων
ΣΦ:	Συνδετικές Φράσεις

Περίληψη

Αντικείμενο αυτής της διατριβής είναι η στατιστική ανάλυση του ύφους Νεοελληνικών κειμένων χωρίς περιορισμούς, με στόχο την αυτόματη ταξινόμησή τους τόσο ως προς το είδος τους όσο και ως προς τον συγγραφέα τους. Η υφολογική πληροφορία εξάγεται μέσω της ανάλυσης του κειμένου από ένα υπολογιστικό εργαλείο ικανό να ανιχνεύει τα όρια των περιόδων και των φράσεων σε οποιοδήποτε κείμενο. Ο ανιχνευτής ορίων περιόδων και φράσεων επιτυγχάνει πολύ ικανοποιητικά αποτελέσματα ανάλυσης, αν και βασίζεται σε ελάχιστους πόρους. Πιο συγκεκριμένα, ο ανιχνευτής περιόδων βασίζεται σε πολύ απλή πληροφορία (όπως το μήκος των λέξεων) και σε κανόνες που εξάγονται αυτόματα από ένα σώμα εκπαίδευσης, σύμφωνα με μία νέα προσέγγιση μηχανικής εκμάθησης. Η ακρίβεια που επιτυγχάνει είναι της τάξης του 99,4% για ένα σώμα κειμένων 200.000 λέξεων. Απ' την άλλη, ο ανιχνευτής ορίων φράσεων βασίζεται σε ένα μικρό σύνολο λέξεων-κλειδιών και στις καταλήξεις των λέξεων για να εκτιμήσει την πιο πιθανή μορφολογική περιγραφή της κάθε λέξης. Η διαδικασία της ανάλυσης γίνεται μέσω της τεχνικής πολλαπλού περάσματος και στο σώμα που αναφέρθηκε παραπάνω επιτυγχάνει ανάκληση και ακρίβεια της τάξης του 90% και 95% αντίστοιχα. Έτσι, το εργαλείο αυτό προσφέρει γρήγορη και αξιόπιστη ανάλυση μεγάλων όγκων κειμένου, με ελάχιστο υπολογιστικό κόστος. Για την αναπαράσταση του ύφους χρησιμοποιούνται 22 υφολογικοί δείκτες που διακρίνονται σε τρία υφομετρικά επίπεδα: (i) *επίπεδο δείγματος*, που σχετίζεται με την έξοδο του ανιχνευτή ορίων περιόδων, (ii) *επίπεδο φράσης*, που σχετίζεται με την έξοδο του ανιχνευτή ορίων φράσεων και (iii) *επίπεδο ανάλυσης*, που αφορά στον τρόπο με τον οποίο έγινε η ανάλυση του κειμένου από τον ανιχνευτή ορίων φράσεων. Το τελευταίο επίπεδο αποτελεί ένα εναλλακτικό τρόπο σύλληψης της υφολογικής πληροφορίας και είναι η πρώτη φορά που χρησιμοποιείται. Επιπλέον, δεν

χρησιμοποιείται καμία λεξιλογική πληροφορία σε αντίθεση με τις προηγούμενες προσεγγίσεις. Αυτό το σύνολο των υφολογικών δεικτών σε συνδυασμό με τις στατιστικές τεχνικές της *πολλαπλής παλινδρόμησης* και της *διαχωριστικής ανάλυσης* οδηγούν στην αυτόματη αναγνώριση του είδους κειμένου (π.χ. επιστημονικά άρθρα, ρεπορτάζ εφημερίδων, μαγειρικές συνταγές, κ.ά.). Αυτές οι τεχνικές της πολυπαραγοντικής στατιστικής διακρίνονται για την εύκολη εκπαίδευσή τους και τη γρήγορη απόκρισή τους καθώς βασίζονται στον υπολογισμό απλών γραμμικών συναρτήσεων. Τα πειράματα ελέγχου της προτεινόμενης μεθοδολογίας πραγματοποιήθηκαν σε ένα σώμα που δημιουργήθηκε από κείμενα που βρέθηκαν σε σελίδες του Διαδικτύου και δεν υπέστησαν καμία χειρονακτική προεπεξεργασία. Τα αποτελέσματα κρίνονται ως πολύ ικανοποιητικά καθώς επιτυγχάνεται ακρίβεια της τάξης του 82-85%, που είναι πολύ υψηλότερη από αντίστοιχα συστήματα της Αγγλικής γλώσσας. Αντίστοιχα πειράματα πραγματοποιήθηκαν για την αυτόματη ταξινόμηση ενός σώματος κειμένων ως προς το συγγραφέα. Η προτεινόμενη μεθοδολογία εφαρμόστηκε σε δύο ομάδες τυχαία επιλεγμένων συγγραφέων μίας εβδομαδιαίας εφημερίδας. Η ακρίβεια που επιτυγχάνεται από το σύστημα αναγνώρισης συγγραφέα (αξιολόγηση κλειστού συνόλου) ήταν στην καλύτερη περίπτωση 81%, σαφώς καλύτερη από την αντίστοιχη απόδοση της πιο σύγχρονης λεξιλογικής μεθόδου (74%). Ωστόσο, τα καλύτερα αποτελέσματα επιτεύχθηκαν από τον συνδυασμό των δύο προσεγγίσεων (87%). Επίσης, παρατηρήθηκε ότι τόσο το μήκος του κειμένου όσο και το μέγεθος του σώματος εκπαίδευσης παίζουν πολύ σημαντικό ρόλο για την αξιόπιστη αναπαράσταση του προσωπικού ύφους ενός συγγραφέα. Εκτός από την αναγνώριση συγγραφέα, πραγματοποιήθηκαν και πειράματα επιβεβαίωσης συγγραφέα καθώς και στατιστικοί έλεγχοι της σημαντικότητας των προτεινόμενων υφολογικών δεικτών.

STATISTICAL IDENTIFICATION OF GENRE AND AUTHOR IN UNRESTRICTED MODERN GREEK TEXT

Summary

This dissertation deals with the statistical analysis of the style of unrestricted Modern Greek text, aiming at the automatic text categorization in terms of genre and author. The stylistic information is extracted via the analysis of the input text by a computational tool able to detect sentence and chunk boundaries in unrestricted text. Although the sentence and chunk boundaries detector is based on minimal resources, it achieves very high accuracy results. In more detail, the sentence boundary detector is based on simple metrics (such as word-length) and disambiguation rules that are extracted automatically from a training corpus, according to a new machine learning methodology. This tool achieves 99.4% accuracy for a 200,000 word corpus. On the other hand, the chunk boundary detector is based on a small lexicon of keywords and common word-suffixes in order to assign the most likely morphological description to every word. Multiple-pass parsing is, then, applied and the recall and precision results on the aforementioned corpus are 90% and 95% respectively. This tool, therefore, provides a reliable solution for the rapid analysis of large volumes of text, requiring minimal computational cost. A set of 22 style markers is used for the representation of the style and they may be distinguished in three stylometric levels: (i) *token-level*, that concerns the output of the sentence boundary detector, (ii) *phrase-level*, that deals with the output of the chunk boundary detector, and (iii) *analysis-level*, that concerns the way in which the text has been analyzed by the chunk boundary detector. The latter may be considered as an alternative way for capturing the stylistic information and is used for first time. Additionally, in contrast to other approaches, lexically-based measures are excluded. This set of style markers in combination with the statistical techniques of *multiple regression* and *discriminant analysis* lead to the automatic detection of text-genre (e.g. academic prose, reportage, recipes, etc.). These

techniques of multivariate statistics are characterized by minimal training and response time cost since they are based on the calculation of simple linear functions. Experiments for testing the proposed methodology were conducted on a corpus that has been constructed by texts downloaded from the Internet, without any manual preprocessing. The classification results are quite satisfying since the proposed approach achieves 82-85% accuracy, which is far higher than the ones that have been referred for other systems (proposed for English). Similar experiments were conducted for the automatic categorization of a corpus in terms of author. The proposed methodology was applied to two groups of randomly-selected authors of a weekly newspaper. The classification accuracy achieved by the author identification system (closed set evaluation) is, in the best case, 81%, which is considerably higher than the corresponding performance of the most modern lexically-based approach (74%). However, the best results are achieved by the combination of these two approaches (87%). Moreover, the text-length and the size of the training corpus play an important role for the reliable representation of the personal style of a particular author. In addition to the author identification, we conducted experiments on author verification as well as statistical tests of importance of the proposed set of style markers.