Χαράλαμπος Χ. Σπυρίδης
Σταύρος Ν. Δημητριάδης

ΠΡΟΤΑΣΗ ΓΙΑ ΕΝΑ ΜΑΘΗΜΑΤΙΚΟ ΠΡΟΣΔΙΟΡΙΣΜΟ ΤΟΥ ΛΕΞΙΛΟΓΙΚΟΥ ΠΛΟΥΤΟΥ ΚΕΙΜΕΝΩΝ ΤΗΣ ΝΕΟΕΛΛΗΝΙΚΗΣ ΓΛΩΣΣΑΣ (ΕΦΑΡΜΟΓΗ ΣΤΟ ΠΟΙΗΤΙΚΟ ΕΡΓΟ ΤΟΥ Γ. ΣΕΦΕΡΗ)

1. Εισαγωγή

Η σχέση του αριθμού των διαφορετικών λέξεων προς τον αριθμό των ολικών λέξεων, που υπάρχουν σε κάποιο κείμενο, είναι φανερό ότι μπορεί να αποτελέσει δείκτη μέτρησης του λεξιλογικού πλούτου του συγγραφέα.

Στην εργασία που ακολουθεί εφαρμόζεται στατιστική ανάλυση σε ποιήματα του Γ. Σεφέρη και παρουσιάζονται οι εξισώσεις (συσχέτιση - γραμμική παλινδρόμηση) που περιγράφουν τη λεξιλογική ποικιλία στα κείμενα αυτά.

2. Βασικές Έννοιες: Ολόνια και Μορφοποιητικά Πεδία στη Γλώσσα

Ένας δημιουργός κειμένου (συγγραφέας, ποιητής, γραμματέας κλπ.) αποτελεί μια πηγή κειμένου που κατά τη λειτουργία της «εκπέμπει» γλωσσικά ολόνια. Η οργάνωση των ολονίων καθορίζεται από την ισχύ των διαφόρων μορφοποιητικών πεδίων, που διέπουν τη λειτουργία της πηγής.

Ο όρος «ολόνιο» εισάγεται από τον Koestler[1] για τον χαρακτηρισμό μιας ολοκληρωμένης, σχετικά αυτόνομης δομής, που από τη μια αποτελείται από μονάδες κατωτέρων επιπέδων και από την άλλη δομείται μαζί με ισότιμες μονάδες με αποτέλεσμα τη δημιουργία δομών ανωτέρου επιπέδου. Τα γλωσσικά ολόνια είναι γνωστά με διαφορετικά ονόματα, ανάλογα με το επίπεδο που ανήκουν: γράμματα ή φωνήματα, μορφήματα, λέξεις, προτάσεις, παράγραφοι.

Με τον όρο «μορφοποιητικό πεδίο» αναφερόμαστε στο σύνολο των παραγόντων των οποίων η συνδυασμένη δράση εκφράζεται από τους κανόνες και δεσμούς που διαμορφώνουν τη δομή του γλωσσικού υλικού σε κάθε επίπεδο της ιεραρχίας των γλωσσικών ολονίων.

Έτσι θεωρούμε πως τα ολόνια του πρώτου επιπέδου του γραπτού κειμένου (τα γράμματα) υπόκεινται στη δράση του «μορφοποιητικού πεδί-

ου Α». Στο πεδίο αυτό περιλαμβάνονται όλοι εκείνοι οι παράγοντες (ιστορικοί, γεωμορφολογικοί κλπ.) που καταλήγουν τελικά στο να επιτρέπουν την ύπαρξη στη γραπτή ελληνική γλώσσα αλληλουχιών όπως πχ. «θαλ» ή «ορμ» ενώ ταυτόχρονα αποτρέπουν την ύπαρξη άλλων όπως πχ. «μτλ» ή «ζγκα».

Τέλος, το ανώτερο επίπεδο των ολονίων - λέξεων θεωρούμε πως υπόκειται στη δράση του «μορφοποιητικού πεδίου Β». Στο πεδίο Β ανήκουν οι παράγοντες που επηρεάζουν την οργάνωση των λέξεων της γλώσσας και επιτρέπουν πχ. τη δομή «της θάλασσας» και όχι «του θάλασσας» (γραμματική) ή τη δομή «θα έρθω αύριο» και όχι «αύριο έρθω θα» (συντακτικό) ή ακόμη τη δομή «τρεχούμενο νερό» και όχι «τρεχούμενο ψωμί» (λογική νοηματική δομή).

3. Η Μαθηματική Περιγραφή

Για να προχωρήσουμε στη μαθηματική μελέτη, ξεκινάμε από την παραδοχή πως ο ποιητής αποτελεί έναν αδέσμευτο κανόνων γραφής δημιουργό κειμένου (Context Free: CF). Για την ερμηνεία του όρου αυτού - καθώς και δύο ακόμη συγγενικών όρων - παρατηρούμε τα εξής: το μορφοποιητικό πεδίο Β, που δρα κατά τη δόμηση των ολονίων - λέξεων μεταξύ τους, αποτελείται από δύο μεγάλες κατηγορίες ομοειδών παραγόντων: α) τους υποκειμενικούς παράγοντες, που καθορίζουν τις επιλογές του δημιουργού κειμένου (πχ. διανοητικές ικανότητες, ποικιλία ύφους κλπ.) και β) τους αντικειμενικούς παράγοντες της επιλογής, που τους αποδίδουμε με τη γενική ονομασία «κανόνες γραφής». Σαν κανόνες γραφής αναφέρουμε τη μορφή της γλώσσας (γραπτής ή προφορικής), τη μορφή του κειμένου (διάλογος ή μονόλογος) και τη λειτουργία του. Οι πραγματικοί αυτοί παράγοντες (τόσο οι υποκειμενικοί, όσο και οι αντικειμενικοί) είναι πάντοτε παρόντες κατά τη δημιουργία ενός κειμένου. Ποικίλει, όμως, η σχετική ισχύς με την οποία η κάθε κατηγορία παραγόντων επηρεάζει το ύφος του δημιουργού.

Κατά τον Lubomir Dolezel[2], μπορούμε να διακρίνουμε τρεις βασικές κατηγορίες δημιουργών κειμένου: α) τον αδέσμευτο των κανόνων γραφής (CF) που αναφέραμε και παραπάνω. Στην περίπτωση αυτή οι αντικειμενικοί παράγοντες εξαφανίζονται από το προσκήνιο και οι υφολογικές παράμετροι συνδέονται μόνο με τους υποκειμενικούς παράγοντες επιλογής. Εμπειρικό παράδειγμα αυτής της κατηγορίας είναι ο ποιητής, που βάζει τη σφραγίδα του προσωπικού του ύφους σε όποιο κείμενο παράγει, ανεξάρτητα από την αντικειμενική διαφοροποίηση των κειμένων μεταξύ τους (πχ. θεματολογική διαφορά). β) Στην άλλη άκρη του φάσματος βρίσκεται ο δημιουργός ο δεσμευμένος των κανόνων γραφής

(Context Bound: CB). Στα κείμενα που παράγει οι επιλογές καθορίζονται από τους εξωτερικούς κανόνες γραφής και όχι από τις ποιότητες που τον χαρακτηρίζουν σαν υποκείμενο. Παράδειγμα σ' αυτήν την κατηγορία αποτελεί πχ. ένας γραφέας δημόσιας υπηρεσίας που ανεξάρτητα από τον οποιονδήποτε εκφραστικό του πλούτο είναι υποχρεωμένος να συντάσσει κείμενα βάσει ορισμένων παγιωμένων μορφών έκφρασης. γ) Ανάμεσα στους δύο προηγούμενους βρίσκεται ο δημιουργός ο ευαίσθητος προς τους κανόνες γραφής (Context Sensitive: CS). Αυτός αφ' ενός μεν συμμορφώνεται προς τις απαιτήσεις κάποιων εξωτερικών κανόνων γραφής, αφ' ετέρου δε διατηρεί τα χαρακτηριστικά εκείνα στοιχεία της ατομικότητάς του, που τον διαχωρίζουν από άλλους δημιουργούς.

Με την εισαγωγή των διαχωρισμών αυτών δεν υπονοείται ότι ο δημιουργός παράγει το κείμενο οδηγούμενος κάθε φορά αποκλειστικά από μια συγκεκριμένη κατηγορία παραγόντων. Απλά τονίζεται η διαφορετική κατά περίσταση θέση του κέντρου βάρους της παραγωγής αυτής. Εξ άλλου, είναι φανερό ότι το ίδιο φυσικό πρόσωπο μπορεί να εμφανίζεται σαν δημιουργός κειμένου διαφορετικής κατηγορίας, ανάλογα με τους εξωτερικούς και εσωτερικούς (σε σχέση με το υποκείμενο) παράγοντες που ρυθμίζουν τη συγκεκριμένη περίπτωση.

Κατά τη μελέτη της στατιστικής της γλώσσας του Γ. Σεφέρη θελήσαμε να απαντήσουμε στο ερώτημα: **«ποιά είναι η σχέση ανάμεσα στις ολικές λέξεις που παραθέτει ο ποιητής και τις διαφορετικές λέξεις που εμφανίζονται στα κείμενά του; Πώς εξελίσσεται η σχέση αυτή»;**

Για να δώσουμε μια απάντηση στα ερωτήματα αυτά μετρήσαμε τα μεγέθη:

N=ολικές λέξεις στο κείμενο και
V=διαφορετικές λέξεις στο κείμενο

για ένα διαφορετικό κάθε φορά πλήθος κειμένων.

Για την επιλογή του προς μελέτην υλικού από το έργο του ποιητή χωρίσαμε το σύνολο των ποιημάτων σε στοιχειώδη κείμενα μήκους περίπου 350 συμβόλων[3].

Στον αριθμό αυτό καταλήξαμε έχοντας τις εξής απαιτήσεις:
α) τα στοιχειώδη αυτά κείμενα να είναι αρκετά μεγάλα, ώστε η στατιστική τους να μην επηρεάζεται αποφασιστικά από σποραδικές ιδιομορφίες του λόγου (πχ. επανάληψη για έμφαση της ίδιας πρότασης) και
β) να βρίσκονται σε αναλογία μικρών ακέραιων αριθμών με το συνολικό ποίημα. Η αναλογία αυτή κυμάνθηκε από 1:1 (κείμενο=ολόκληρο ποίημα) μέχρι 1:10 (κείμενο=το ένα δέκατο ολόκληρου ποιήματος).

Για να δημιουργήσουμε τμήματα μεταβλητού μεγέθους, πήραμε κάθε φορά για μελέτη και ένα διαφορετικό αριθμό στοιχειωδών κειμένων. Η συρραφή αυτή των κειμένων και η από κοινού μελέτη τους προϋποθέτουν την παραδοχή πως ο ποιητής ανήκει στην κατηγορία των CF δημι-

ουργών κειμένου (αδέσμευτος κανόνων γραφής).

Δεχόμαστε πως ο τρόπος με τον οποίο αναπτύσσονται οι λεξιλογικές ικανότητες του ποιητή δεν καθορίζονται από τα συγκεκριμένα ποιήματα απ' όπου προέρχονται τα προς μελέτη κείμενα.

Στο σχ. 1 παρουσιάζεται σε log-log διάγραμμα η πειραματική καμπύλη που συνδέει τα μεγέθη N και V. Η γραμμική εξάρτηση των λογαρίθμων οδηγεί σε μια σχέση για τα N και V της μορφής:

$$V = AN^B \qquad (3.1)$$

Οι σταθερές A και B βρέθηκαν ίσες με:

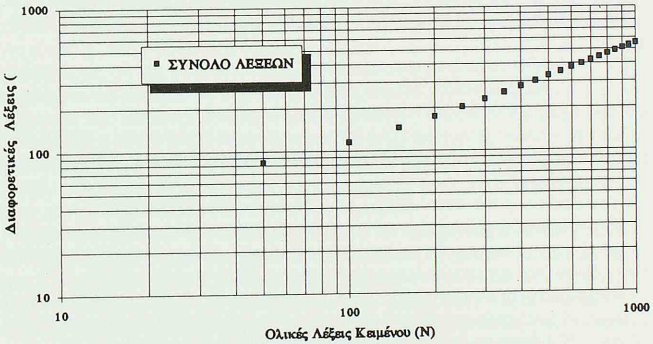
A = 2,000

B = 0,808

Άρα: $V = 2.\ N^{0,808}$

Η σχέση (3.1) δίνει τον αριθμό V των διαφορετικών λέξεων ενός κειμένου σε συνάρτηση με τον αριθμό N των ολικών λέξεων του κειμένου.

Σχήμα 1



Αν δεχτούμε σαν δείκτη λεξιλογικού πλούτου το κλάσμα:

$$\Lambda = \frac{V}{N} = \frac{2.N^{0,808}}{N} = 2.N^{-0,192}$$

είναι φανερό πως για σταθερό όγκο κειμένου (N) ο πλούτος του λεξιλογίου (διαφορετικές λέξεις V) είναι τόσο μεγαλύτερος, όσο το Λ τείνει στην τιμή 1. Επειδή προφανώς πρέπει:

$$\Lambda \leq 1$$

μπορούμε να προσδιορίσουμε ένα κατώτατο όριο μεγέθους κειμένου για το οποίο ισχύει η σχέση (3.1). Πρέπει:

$$\Lambda \leq 1$$

$$0{,}192 \ln N \geq \ln 2$$

$$\ln N \geq \frac{\ln 2}{0{,}192} = 3{,}61$$

$$N \geq 36{,}97$$

Η εξίσωση, επομένως, (3.1) ισχύει για συνολικό αριθμό λέξεων μεγαλύτερο ή ίσο με 37. Για N μικρότερο αυτής της τιμής η σχέση μεταξύ V και N μπορεί να είναι οποιαδήποτε με όριο την περίπτωση όπου V=N. Η τιμή N=37 μας δίνει ένα στατιστικά προσδιοριζόμενο όριο σχετικά με την ικανότητα του ποιητή να παραθέτει έναν αριθμό διαδοχικών λέξεων μέσα στο κείμενο, οι οποίες να είναι διαφορετικές μεταξύ τους.

Εκτός από τις συνολικές και διαφορετικές λέξεις στο κείμενο υπολογίσαμε τις εξισώσεις τις αντίστοιχες της (3.1) για ιδιαίτερες ομάδες λέξεων, που σαν κοινό χαρακτηριστικό έχουν την ταυτότητα του αρχικού τους γράμματος. Έτσι αν N_A είναι οι ολικές λέξεις που αρχίζουν από το γράμμα Α σ' ένα συγκεκριμένου μεγέθους κείμενο και V_A οι διαφορετικές λέξεις ανάμεσα σ' αυτές τα μεγέθη V_A και N_A συνδέονται με τη σχέση:

$$V_A = 1{,}929\, N_A^{0{,}808}$$

Δίνουμε στη συνέχεια τις αντίστοιχες εξισώσεις για τα επτά πρώτα γράμματα από άποψη συχνότητας εμφάνισής τους ως αρκτικών γραμμάτων στις λέξεις των ποιημάτων του Γ. Σεφέρη.

1. Αρχικό Γράμμα Τ (με συχνότητα εμφάνισης 16,18%): $V_T = 0{,}478 N_T^{0{,}808}$

2. Αρχικό Γράμμα Π (με συχνότητα εμφάνισης 10,53%): $V_\Pi = 1{,}642 N_\Pi^{0{,}972}$

3. Αρχικό Γράμμα Σ (με συχνότητα εμφάνισης 10,35%): $V_{\Sigma}=0,934N_{\Sigma}^{0,829}$

4. Αρχικό Γράμμα Κ (με συχνότητα εμφάνισης 9,72%): $V_{K}=1,116N_{K}^{0,827}$

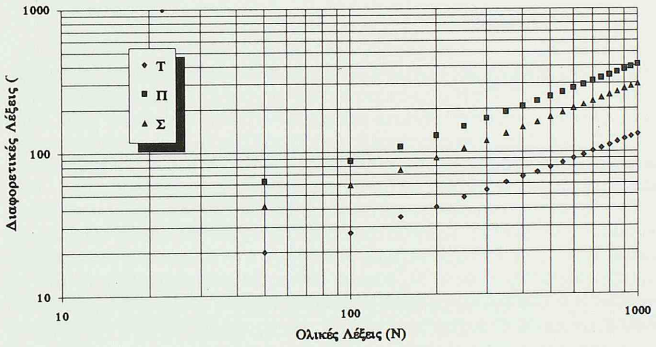
5. Αρχικό Γράμμα Α (με συχνότητα εμφάνισης 8,86%): $V_{A}=1,929N_{A}^{0,809}$

6. Αρχικό Γράμμα Μ (με συχνότητα εμφάνισης 8,77%): $V_{M}=1,634N_{M}^{0,729}$

7. Αρχικό Γράμμα Ε (με συχνότητα εμφάνισης 5,02%): $V_{E}=1,959N_{E}^{0,733}$

Στα σχ. 2 και σχ. 3 παραθέτουμε τα διαγράμματα (N - V) που αφορούν στα παραπάνω γράμματα για διευκόλυνση της συγκριτικής τους μελέτης.

Σχήμα 2



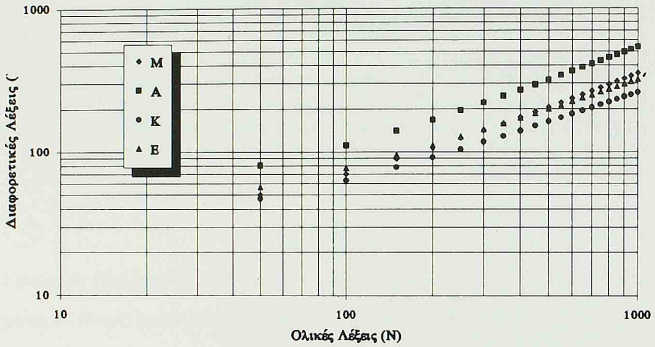
Η σχέση $\mathbf{V = AN^{B}}$ γράφεται ισοδύναμα σαν $\mathbf{\ln V = B\ln N + \ln A}$ απ' όπου φαίνεται πως ο εκθέτης Β είναι η κλίση των διαγραμμάτων στα σχ. 1, 2 και 3. Παρατηρεί κανείς πως υπάρχουν ομάδες γραμμάτων με περίπου ίδια κλίση του διαγράμματός τους (N - V). Αυτές είναι:

α) ομάδα «Τ, Π, Α»: $B_{T} = 0,808$, $B_{\Pi} = 0,792$, $B_{A} = 0,809$

β) ομάδα «Σ, Κ»: $B_{\Sigma} = 0,829$, $B_{K} = 0,827$

γ) ομάδα «Μ, Ε»: B_M =0,729, B_E = 0,733

Σχήμα 3



Για τις ομάδες με την ίδια κλίση ο παράγοντας, ο οποίος καθορίζει την λεξιλογική ποικιλία που αντιπροσωπεύει το κάθε αρχικό γράμμα, είναι η τιμή του Α. Υψηλή τιμή του Α σημαίνει μεγαλύτερη ποικιλία λέξεων, ενώ χαμηλή τιμή του Α αντιπροσωπεύει μικρότερο αριθμό επαναλαμβανόμενων λέξεων.

Η πολύ χαμηλή τιμή του συντελεστή Α για το γράμμα «Τ» σημαίνει πως στα κείμενα υπάρχει ένας μικρός αριθμός διαφορετικών λέξεων με αρχικό γράμμα το «Τ» που επαναλαμβάνονται συνεχώς. Κάτι τέτοιο προέρχεται από τη δομή της ελληνικής γλώσσας, που χρησιμοποιεί έναν μικρό αριθμό συχνότατα εμφανιζόμενων λέξεων με αρχικό γράμμα το «Τ» δηλ. τα άρθρα το, τα, την, τους κλπ.

Κάτι το ανάλογο αλλά σε μικρότερο βαθμό, συμβαίνει και με το γράμμα «Σ» όπου $A_Σ$ = 0,934. Πράγματι, η συγχώνευση της πρόθεσης «εις» ή «σε» με τα άρθρα, που προαναφέραμε, δίνει ύπαρξη σε μια κατηγορία μικρού πλήθους και υψηλής συχνότητας εμφάνισης λέξεων, όπως στο, στα, στην, κλπ.

Αντίθετα υψηλή τιμή του συντελεστή Α, όπως στα γράμματα «Α» και «Ε» όπου A_A = 1,929 και A_E = 1,959 υποδηλώνει τη χρήση ενός μεγαλύτερου αριθμού διαφορετικών μεταξύ τους λέξεων, που έχουν σαν αρχικά γράμματα τα «Α» και «Ε», αντίστοιχα.

Για να ερμηνεύσουμε τα αποτελέσματά μας αναφερθήκαμε στη «δομή της ελληνικής γλώσσας», έναν παράγοντα εξωτερικό οπωσδήποτε του υποκειμένου - δημιουργού. Η ερμηνεία αυτή δεν έρχεται σε αντίθεση με την υπόθεσή μας πως ο ποιητής είναι δημιουργός αδέσμευτος κανόνων γραφής (C.F). Αυτό που συμβαίνει είναι πως ένας δημιουργός κειμένου βρίσκεται πάντοτε λιγότερο ή περισσότερο δεσμευμένος από εξωτερικά επιβαλλόμενους κανόνες γραφής. Στην περίπτωση της μελέτης μας η δέσμευση αυτή εκφράζεται στο επίπεδο της γραμματικής και του συντακτικού (χρήση συγκεκριμένων λέξεων - άρθρων και συγκεκριμένης σύνταξης). Η υπόθεσή μας για C.F. δημιουργό κειμένου παραμένει, όσον αφορά στην επιλογή των λέξεων εκείνων, που αποδίδουν τα ουσιαστικότερα νοήματα του λόγου (ουσιαστικά, επίθετα, ρηματικές μορφές).

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. A. Koestler, *Η πράξη της Δημιουργίας*, (Μετάφραση: Ι. Χατζηνικολή), Εκδόσεις Ι. Χατζηνικολή, Αθήνα, 1976.
2. L. Dolezel, and R.W. Bailey (editors), *Statistics and Style*, American Elsevier Publishing Company Inc., New York, 1969.
3. Σ. Ν. Δημητριάδη, *Μαθηματική και Πληροφορική μελέτη της γραπτής Ελληνικής γλώσσας* (Εφαρμογή στο ποιητικό έργο του Γ. Σεφέρη), Διπλωματική Διατριβή υπό την επίβλεψη του Χ. Χ. Σπυρίδη που υποβλήθηκε στο Τμήμα Φυσικής του Α.Π.Θ, Θεσσαλονίκη, 1986.