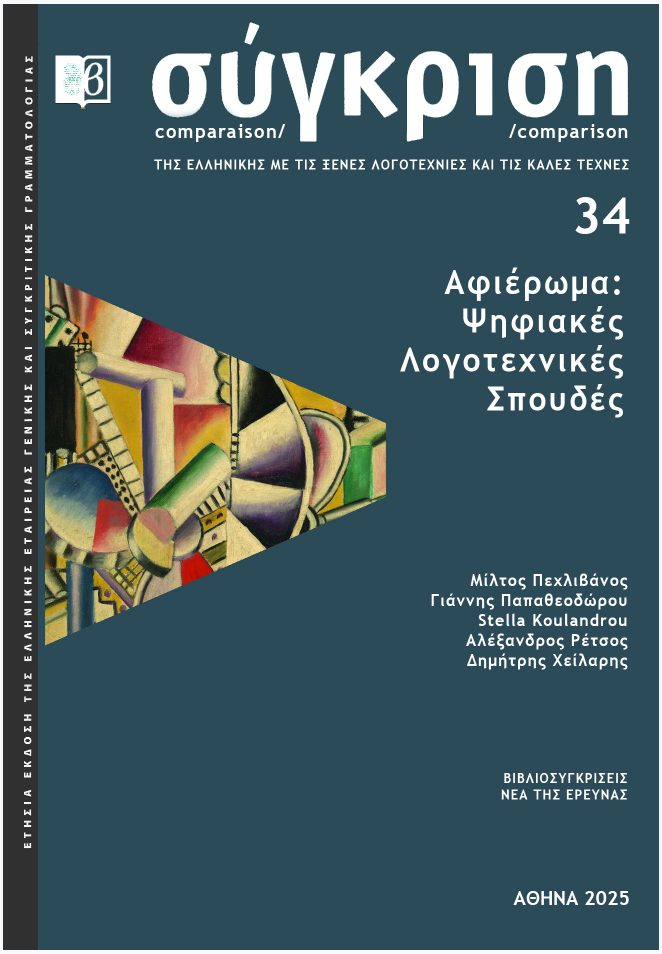
Σύγκριση/Comparaison/Comparison

Αρ. 34 (2025)

Ψηφιακές Λογοτεχνικές Σπουδές



ΣΝΕΛ: Ένας νέος γλωσσικός πόρος για τη μελέτη της λογοτεχνίας στα ελληνικά

Διονύσης Γούτσος, Χριστιάνα Νίκα, Κωνσταντίνος Περήφανος, Γεωργία Φραγκάκη

Βιβλιογραφική αναφορά:

Γούτσος Δ., Νίκα Χ., Περήφανος Κ., & Φραγκάκη Γ. (2026). ΣΝΕΛ: Ένας νέος γλωσσικός πόρος για τη μελέτη της λογοτεχνίας στα ελληνικά. *Σύγκριση/Comparaison/Comparison*, (34), 60–86. ανακτήθηκε από https://ejournals.epublishing.ekt.gr/index.php/sygkrisi/article/view/43858

ΔΙΟΝΥΣΗΣ ΓΟΥΤΣΟΣ,[1] **ΧΡΙΣΤΙΑΝΑ ΝΙΚΑ,**[2] **ΚΩΝΣΤΑΝΤΙΝΟΣ ΠΕΡΗΦΑΝΟΣ,**[1] **ΓΕΩΡΓΙΑ ΦΡΑΓΚΑΚΗ**[3]
[1] Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, [2] Jagiellonian University, [3] Πανεπιστήμιο Πελοποννήσου

ΣΝΕΛ: Ένας νέος γλωσσικός πόρος για τη μελέτη της λογοτεχνίας στα ελληνικά

1. Εισαγωγή

Το άρθρο αυτό παρουσιάζει τις αρχές συγκρότησης και τη διαδικασία δημιουργίας ενός νέου γλωσσικού πόρου για τα ελληνικά, του *Σώματος κειμένων Νεοελληνικής Λογοτεχνίας* (ΣΝΕΛ), που έχει σχεδιαστεί με σκοπό τη συστηματική διαχρονική μελέτη της λογοτεχνίας του 20ού αιώνα στα ελληνικά. Το ΣΝΕΛ αναπτύχθηκε με αφορμή μια ερευνητική συνεργασία στο ευρύτερο πεδίο των Ψηφιακών Ανθρωπιστικών Επιστημών (ΨΑΕ), που περιλαμβάνει την ιστορική οικονομολογία και έχει ως στόχο να αναπτύξει μεθόδους για τον αυστηρό έλεγχο ερευνητικών υποθέσεων που αφορούν τη διαχρονική σχέση του πολιτισμού, της ψυχολογίας και της κοινωνικής οικονομίας (Martins & Baumard, 2022· Atari & Henrich, 2023· Jackson & Atari, 2025).[1] Στόχος της συνεργασίας είναι η χρήση υπολογιστικών εργαλείων για τη διερεύνηση των αξιών, των προτιμήσεων και των γνωσιακών αναπαραστάσεων του παρελθόντος έτσι ώστε να γίνει δυνατή η σύνδεση του πολιτισμικού περιβάλλοντος με τις κυρίαρχες κοινωνικο-οικονομικές τάσεις και τα ιστορικά γεγονότα (βλ. Δημητρούλια κ.ά., 2024, σσ. 198 κ.εξ. για παρόμοιες έρευνες).

Είναι χαρακτηριστικό ότι σε αυτό το διευρυμένο ερευνητικό πλαίσιο οι ερευνητικές υποθέσεις και θεωρίες που αναπτύσσονται απαιτούν πρόσβαση σε εκτεταμένους γλωσσικούς πόρους, και μάλιστα εξειδικευμένων κειμενικών ειδών όπως η λογοτεχνία, για την οποία, πέρα από την καλλιτεχνική και αισθητική της αξία, γίνεται η παραδοχή ότι αποτυπώνει με ευαισθησία και ακρίβεια τις εξελίξεις στο κοινωνικο-οικονομικό πεδίο των ιστορικών της συμφραζομένων. Τέτοιου είδους γλωσσικοί πόροι, σε αντίθεση με άλλες γλώσσες, όπως θα διαπιστωθεί στην επόμενη ενότητα, λείπουν από τα ελληνικά με αποτέλεσμα να είναι δυσχερής η ανάπτυξη μακροσκοπικών υποθέσεων για τη διαχρονία της ελληνικής τόσο για τη γλώσσα όσο και για τη λογοτεχνία ως γλωσσικό και πολιτισμικό πεδίο. Η έλλειψη συστηματικά οργανωμένων δεδομένων που να εκτείνονται σε βάθος χρόνου ερμηνεύει και την απουσία ερευνών στα ελληνικά με την οπτική της εξ αποστάσεως ανάγνωσης (distant reading) της λογοτεχνίας (ή κειμενικών δεδομένων ευρύτερα), σύμφωνα με τον όρο που εισήγαγε ο Franco Moretti, ή με όρους της πολιτισμικής ανάλυσης του Manovich (βλ. Δημητρούλια κ.ά., 2024, 196-198 για μια εισαγωγή). Το *Σώμα κειμένων Νεοελληνικής Λογοτεχνίας* έχει σχεδιαστεί για να αποτελέσει έναν ολοκληρωμένο γλωσσικό πόρο για τα ελληνικά, που μπορεί να χρησιμοποιηθεί σε όλο το φάσμα των σχετικών πεδίων της ΨΑΕ, που περιλαμβάνουν την υφολογία σωμάτων κειμένων (corpus stylistics, βλ. Φραγκάκη, 2024, 2025), την εξ αποστάσεως ανάγνωση της λογοτεχνίας (Δημητρούλια 2022) και

[1] Η ερευνητική συνεργασία μεταξύ της Σχολής Ψυχολογίας του Πανεπιστημίου της Βιέννης και του Ινστιτούτου Ψυχολογίας του Πανεπιστημίου Jagiellonian της Πολωνίας διεξάγεται στο πλαίσιο του προγράμματος IP-PAD (MSCA Doctoral Network), που χρηματοδοτείται από την Ευρωπαϊκή Ένωση (Horizon Europe: HORIZON-MSCA-2021-DN-01, Grant Agreement No. 101072992).

την πολιτισμική και κοινωνικο-οικονομική ανάλυση, στα οποία αναφερθήκαμε πιο πάνω (βλ. και Green, 2017).

Στη συνέχεια του άρθρου θα αναφερθούμε στα σώματα κειμένων λογοτεχνίας που είναι διαθέσιμα σε διάφορες γλώσσες και στα ελληνικά και στις αρχές συγκρότησης και τη δημιουργία του Σώματος κειμένων Νεοελληνικής Λογοτεχνίας, με αναφορά σε ξεχωριστή ενότητα στις ιδιαιτερότητες της ψηφιοποίησης πολυτονικών κειμένων, και θα παρουσιάσουμε ενδεικτικά παραδείγματα ανάλυσης των δεδομένων του ΣΝΕΛ, ενώ στο τέλος θα συνοψίσουμε τα συμπεράσματα και τις προοπτικές από τη δημιουργία του.

2. Σώματα κειμένων λογοτεχνίας

Τα σώματα κειμένων λογοτεχνίας ανήκουν στα λεγόμενα ειδικά σώματα κειμένων, εκείνα δηλαδή που περιλαμβάνουν δεδομένα από ένα μόνο κειμενικό είδος ή μια γλωσσική ποικιλία, σε αντίθεση με τα γενικά σώματα κειμένων, που συλλέγουν δεδομένα από όλα τα κειμενικά είδη και τις ποικιλίες, επιχειρώντας να ανασυστήσουν μια αντιπροσωπευτική εικόνα ολόκληρης της γλώσσας (Γούτσος & Φραγκάκη, 2015, σ. 32).

Η πρακτική δημιουργίας λογοτεχνικών σωμάτων κειμένων διαφέρει από γλώσσα σε γλώσσα, αντανακλώντας κατά κάποιο τρόπο την αντίληψη για τη λογοτεχνία στις αντίστοιχες κουλτούρες. Για παράδειγμα, είναι χαρακτηριστικό ότι στα γαλλικά η λογοτεχνία αποτελεί το 90% των δεδομένων του πιο γνωστού και παλαιότερου γενικού σώματος κειμένων, του Frantext, που αποτελεί εξέλιξη του *Trésor de la Langue Française.*[2] Τα λογοτεχνικά κείμενα που περιλαμβάνονται ανήκουν σε όλα τα κειμενικά υπο-είδη της λογοτεχνίας (μυθιστορήματα, απομνημονεύματα, αυτοβιογραφίες, προσωπικά ημερολόγια, θέατρο, ποίηση, δοκίμια) και εκτείνονται από τον 12ο αι. έως σήμερα.

Δύο ειδικά σώματα κειμένων λογοτεχνίας στα γαλλικά είναι το *Théâtre Classique*,[3] που περιλαμβάνει 2.100 θεατρικά έργα από τον 12ο έως τον 20ό αι. και το *Corpus Malherbe*[4] με ποιητικά και θεατρικά έργα σε στίχους από τον 17ο έως τον 20ό αι. (πάνω από 1 εκατ. στίχοι). Ανάλογο είναι το *Corpus Chapitres*, που περιλαμβάνει 2.000 μυθιστορήματα στα γαλλικά από τον 19ο και 20ό αι. (Leblond 2022). Υπάρχουν ακόμη συλλογές που περιλαμβάνουν μόνο ένα εκτεταμένο λογοτεχνικό έργο, ιδίως παλαιότερων περιόδων.[5] Τέλος, υπάρχουν και μεγάλα αποθετήρια όπως η ψηφιακή βιβλιοθήκη της Εθνικής Βιβλιοθήκης της Γαλλίας *Gallica* (Langlois, 2021),[6] απ' όπου αντλούν δεδομένα διάφορα ερευνητικά προγράμματα.

Αντίθετα, στην παράδοση της αγγλικής η διαχρονική λογοτεχνία δεν αποτελεί το επίκεντρο γενικών σωμάτων κειμένων, ούτε όμως δημιουργούνται εξ υπαρχής σώματα κειμένων λογοτεχνίας, αλλά από μεγάλα αποθετήρια αντλούνται κείμενα με τα οποία δημιουργούνται ad hoc σώματα κειμένων πεζογραφίας ή ποίησης, όπως και στην περίπτωση της γαλλικής *Gallica*. Βασικές πηγές δεδομένων

[2] https://www.frantext.fr
[3] https://www.theatre-classique.fr/index.html
[4] https://www.ortolang.fr/market/corpora/malherbe/v3.7.5
[5] Βλ. για παράδειγμα: https://www.chansondaspremont.eu/index.html - https://hdl.handle.net/11403/digulleville/v1
[6] https://gallica.bnf.fr/selections/fr/html/litterature-de-fiction-en-vers-et-en-prose

είναι η Ψηφιακή Βιβλιοθήκη HathiTrust,[7] ένα συλλογικό αποθετήριο με 17,5 εκατομμύρια τίτλους, το Oxford Text Archive,[8] με πλήθος κειμένων, αλλά και σωμάτων κειμένων στα αγγλικά και σε άλλες γλώσσες, και, φυσικά, το γνωστό Project Gutenberg,[9] που περιλαμβάνει 60.000 περίπου ψηφιακά βιβλία χωρίς πνευματικά δικαιώματα. Οι έρευνες του Stanford Digital Humanities La[10] βασίζονται σε τέτοιου είδους πηγές, ενώ χαρακτηριστικά παραδείγματα είναι άρθρα όπως των Elson et al. (2010) και Bamman et al. (2014) για την πεζογραφία, ή των Underwood & Sellers (2015) και του Jacobs (2018), των Kim et al. (2020) κ.ά., για την ποίηση, αντίστοιχα. Είναι προφανές ότι το μεγάλο μέγεθος των αποθετηρίων αυτών επιτρέπει τη χρήση δεδομένων τους για μεγάλου εύρους έρευνες λ.χ. με τη χρήση τεχνικών βαθείας μάθησης (λ.χ. Yang, 2025).[11]

Σημαντικές είναι επίσης οι έρευνες στα αγγλικά που βασίζονται σε επιλεγμένα δεδομένα πεζογραφίας ή ποίησης, τα οποία μελετούν εις βάθος όπως η έρευνα της Mahlberg για τον Dickens (λ.χ. Mahlberg, 2013· March et al., 2023), για την λογοτεχνία του 19ου αι. (Mahlberg et al., 2019) και ειδικότερα την παιδική λογοτεχνία της περιόδου (Čermáková & Mahlberg, 2022) με βάση τα σώματα κειμένων CLiC.[12] Παρόμοιες είναι οι συμβολές των Semino & Short (2004), της Fischer-Starcke (λ.χ. 2009, 2010) για το έργο της Jane Austen, του Busse (2002) για το έργο του Σαίξπηρ, της Busse (2020) για τη λογοτεχνία του 19ου αι. κ.ά.

Τέλος, λογοτεχνικά υποσώματα είναι διαθέσιμα σε μεγάλα γενικά σώματα κειμένων αναφοράς της αγγλικής όπως το COCA ή το COHA[13] ή και σε μικρότερα όπως το SCOTS.[14] Η ίδια πρακτική ακολουθείται για τις περισσότερες ευρωπαϊκές γλώσσες λ.χ. τα αλβανικά,[15] τα ιταλικά,[16] τα ισπανικά[17] κ.ά. Από την άλλη, αυτόνομα σώματα κειμένων λογοτεχνίας υπάρχουν επίσης για τις περισσότερες ευρωπαϊκές γλώσσες, όπως φαίνεται στον Πίνακα 1, που συνοψίζει τα σχετικά στοιχεία:[18]

Πίνακας 1: *Αυτόνομα σώματα κειμένων λογοτεχνίας σε ευρωπαϊκές γλώσσες*

Γλώσσα	**Σώμα κειμένων**	**Μέγεθος (αριθμός δειγμάτων)**
Αγγλικά (παλαιά)	Complete Corpus of Anglo-Saxon Poetry	
	York-Helsinki Parsed Corpus of Old English Poetry	71.490
Γερμανικά	Deutscher Novellenschatz	

[7] http://www.hathitrust.org
[8] https://llds.ling-phil.ox.ac.uk/llds/xmlui
[9] https://www.gutenberg.org
[10] Βλ. https://digitalhumanities.stanford.edu/projects/, https://litlab.stanford.edu/projects
[11] Άλλες μεγάλες συλλογές λογοτεχνικών κειμένων είναι διαθέσιμες στο: https://textual-optics-lab.uchicago.edu/english και στο: https://emed.folger.edu/about
[12] https://github.com/mahlberg-lab/corpora/blob/master/INDEX.pdf
[13] Βλ. https://www.english-corpora.org/coca και https://www.english-corpora.org/coha
[14] https://www.scottishcorpus.ac.uk/corpus-details
[15] https://albanian.web-corpora.net
[16] http://tlio.ovi.cnr.it/TLIO/index2.html
[17] http://www.rae.es
[18] Βλ. σχετικά και: https://www.clarin.eu/resource-families/literary-corpora, καθώς και Γούτσος (υπό προετοιμασία).

Γλώσσα	**Σώμα κειμένων**	**Μέγεθος (αριθμός δειγμά-των)**
Δανικά	Johannes V. Jensen Corpus	1.760.093
Εσθονικά	Collection of older original Estonian-language works of fiction	
	Corpus of Estonian fiction	5.768.504
	Estonian Runic Songs' Database	
Ισπανικά	Banco de Datos de Once Novelas Españolas 1951—1971 (SOL) (2014-10-08)	1.267.391
	Electronic corpus of 15th-century Castilian cancionero manuscripts	
Κροατικά	One-million Corpus of Croatian Literary Language	1.000.000
Λετονικά	Latvian literature classics	
Νορβηγικά (Bokmal)	NorGramBank – Fiction in Norwegian Bokmål	26.903.637
	NorGramBank children's fiction in Norwegian Bokmål	4.111.213
Νορβηγικά (Nynorsk)	NorGrambank children's fiction in Norwegian Nynorsk	1.043.260
	NorGramBank fiction in Norwegian Nynorsk	2.884.376
Πολωνικά	1000 Novels Corpus	
	1000PLUS Novels Corpus (1.0)	17.352.826
	Late 19th- and Early 20th-Century Polish Novels	
	POE: Microcorpus of 20th century Polish poetry	
Πορτογαλικά	LT Corpus	1.781.083
Ρωσικά	Russian Poetry Corpus	14.097.265
Σααμικά (βόρεια)	North Saami Corpus (Literature) (UHLCS)	17.830
Σλοβενικά	Corpus of longer narrative Slovenian prose KDSP	11.000.000
	The corpus of older Slovenian narrative prose PriLit	1.275.209
Σουηδικά	August Strindberg's novels	4.309.037
	Bonnier novels I (1976/77)	6.578.675
	Bonnier novels II (1980/81)	4.304.271
Τσεχικά	Corpus of Contemporary Czech Poetry (C3P)	17.500.000
	Corpora of texts by Karel Čapek	2.300.000
	Cep corpus	420.000
Φιλανδικά	Classics of Finnish Literature, Kielipankki Version	1.500.000
	Corpus of Early Literary Finnish	
	Corpus of Finnish Literary Classics	1.456.658

Γλώσσα	Σώμα κειμένων	Μέγεθος (αριθμός δειγμά-των)
	Corpus of Old Literary Finnish	3.428.618
	Finnish Corpus (Literature) (UHLCS)	68.425
	The Finnish Gutenberg Corpus	34.487.420
	The Morpho-Syntactic Database of Mikael Agricola's Works	428.314

Τέλος, ιδιαίτερη μνεία πρέπει να γίνει σε σώματα κειμένων που περιλαμβάνουν λογοτεχνικά κείμενα από περισσότερες από μία γλώσσες όπως αυτά που δημιουργήθηκαν στο πλαίσιο ευρωπαϊκών προγραμμάτων, από τα οποία χαρακτηριστικότερα είναι το *DraCor* (Drama Corpora, βλ. Fischer et al., 2019)[19] με θεατρικά έργα από την ελληνική αρχαιότητα και άλλες 22 γλώσσες και το *ELTeC* (European Literary Text Collection, βλ. Burnard et al., 2021) με 100 μυθιστορήματα της περιόδου 1840 με 1920 από 12 γλώσσες, αλλά και το *Corpus of the Canon of Western Literature* (Green, 2017) με 805 έργα της δυτικής λογοτεχνίας στα αγγλικά από το Project Gutenberg.[20]

Για τα ελληνικά δεν διαθέτουμε τα μεγάλα αποθετήρια έργων της αγγλικής, αν και το Project Gutenberg περιλαμβάνει κάποια λογοτεχνικά έργα στα ελληνικά χωρίς πνευματικά δικαιώματα, απ' όπου έχουν αντλήσει δεδομένα μεμονωμένα ερευνητικά προγράμματα. Τα γενικά σώματα κειμένων *Σώμα Ελληνικών Κειμένων* (ΣΕΚ, βλ. Goutsos, 2010), που περιλαμβάνει κείμενα από το 1990 έως το 2010, και *Διαχρονικό Σώμα Ελληνικών Κειμένων του 20ού αιώνα* (ΣΕΚ20, βλ. Goutsos et al., 2017), που περιλαμβάνει κείμενα από το 1900 έως το 1989, διαθέτουν λογοτεχνικά υποσώματα μεγέθους 2.595.542 και 1.289.550 λέξεων, αντίστοιχα.[21] Πρόκειται για τα μεγαλύτερα σώματα κειμένων λογοτεχνίας στα ελληνικά έως τώρα, από τα οποία έχουν προκύψει τα πρώτα ερευνητικά πορίσματα για το είδος αυτό (Φραγκάκη, υπό προετοιμασία).

Μια δεύτερη κατηγορία αφορά μεγάλα αποθετήρια ή πλατφόρμες όπως η Ψηφιακή Βιβλιοθήκη Νεοελληνικών Σπουδών *Ανέμη*,[22] ο *Πολιτιστικός Θησαυρός της Ελληνικής Γλώσσας* (ΠΟΘΕΓ),[23] μια συλλογή κειμένων της νεοελληνικής λογοτεχνίας από το 1774 έως το 2000 σε ψηφιακή μορφή, που περιλαμβάνει αποσπάσματα του έργου 108 δημιουργών, εργαλεία για την αναζήτηση ποσοτικών δεδομένων αλλά και φωτογραφικό υλικό, βιογραφικά και εργογραφικά στοιχεία κ.ά., και η *Ανεμόσκαλα*,[24] η μόνη εκτενής συλλογή ποιητικών κειμένων, η οποία διαθέτει σώματα κειμένων και τη δυνατότητα δημιουργίας συμφραστικών πινάκων για 10 νεοέλληνες ποιητές, ενώ για το έργο άλλων 5 ποιητών δίνει μόνο τη δυνατότητα δημιουργίας συμφραστικών πινάκων.

Όσον αφορά τα αυτόνομα σώματα κειμένων, μπορούμε να αναφέρουμε τα *Ελληνικά Μεσαιωνικά Κείμενα* του Πανεπιστημίου Αιγαίου με 3.419.553 λέξεις,

[19] https://dracor.org/#

[20] https://www.dropbox.com/scl/fi/lf7bk6npx8s7s6ehmipca/Corpus-of-the-Canon-of-Western-Literature-1.0.rar?rlkey=jemnkzbr23xf97hp96l0scv9c&e=1&dl=0

[21] Βλ. http://sek.edu.gr και http://greekcorpus20.phil.uoa.gr.

[22] https://anemi.lib.uoc.gr/?lang=el

[23] http://www.potheg.gr/intro.aspx

[24] https://www.greek-language.gr/digitalResources/literature/tools/concordance/index.html

που περιλαμβάνει όμως και μη λογοτεχνικά κείμενα από τον 4ο αιώνα έως τον 16ο αιώνα.[25] Κατά πολύ μικρότερα είναι η «Συλλογή ελληνικών διηγημάτων από τα τέλη του 19ου και τις αρχές του 20ου αιώνα δημοσιευμένων στον ελληνικό περιοδικό τύπο» του Πανεπιστημίου Κρήτης,[26] που περιλαμβάνει 40 διηγήματα (με απλή σάρωση, ωστόσο, χωρίς περαιτέρω επεξεργασία), η ελληνική συλλογή του ELTeC με μόλις 17 κείμενα του 19ου και των αρχών του 20ου αιώνα (98.607 λέξεις), που δεν συμπεριλαμβάνεται λόγω μεγέθους στην οριστική έκδοση του προγράμματος[27] και ad hoc συλλογές όπως των Stamou et al. (2020) με 5 μυθιστορήματα (380.000 λέξεις περίπου). Τέλος, σώματα κειμένων για το έργο μεμονωμένων δημιουργών (π.χ. Παπαδιαμάντη, Καρυωτάκη, Καβάφη, Βιζυηνού, Σουρή, Δημουλά) διατίθενται και από τα αποθετήρια ελληνικών Πανεπιστημίων στο πλαίσιο της εθνικής υποδομής γλωσσικών πόρων και τεχνολογιών Clarin:el.[28]

Είναι σαφές ότι υπάρχει ανάγκη όχι μόνο για περισσότερους γλωσσικούς πόρους για τη νεοελληνική λογοτεχνία, αλλά και για πιο συστηματικές συλλογές κειμένων με τη μορφή αυτόνομων σωμάτων κειμένων που δεν θα περιορίζονται σε μεμονωμένους λογοτέχνες, εποχές ή κειμενικά υπο-είδη, αλλά θα είναι συγκροτημένα με σκοπό τη μελέτη των ιδιαιτεροτήτων του κειμενικού είδους της λογοτεχνίας στα ελληνικά. Αυτή την ανάγκη έχει σχεδιαστεί να καλύψει το ΣΝΕΛ, επιδιώκοντας να αποτελέσει έναν νέο γλωσσικό πόρο για τη μελέτη της νεοελληνικής λογοτεχνίας στο σύνολό της.

3. Αρχές δημιουργίας και συγκρότηση του ΣΝΕΛ

Η διαδικασία συγκρότησης ενός σώματος κειμένων, σύμφωνα με το Γούτσος & Φραγκάκη (2015, σσ. 42 κ.εξ.· βλ. και Percillier, 2017 ειδικά για τα λογοτεχνικά σώματα κειμένων), περιλαμβάνει τα εξής στάδια:

- καθορισμός αρχών δημιουργίας του σώματος κειμένων
- διερεύνηση πηγών (επιλογή δεδομένων)
- εισαγωγή των δεδομένων με αντιγραφή, πληκτρολόγηση, σάρωση, οπτική αναγνώριση χαρακτήρων
- καθαρισμός (διόρθωση, επιμέλεια) και αποθήκευση
- τυποποίηση
- κωδικοποίηση
- επισημείωση.

Σε κάθε περίπτωση, η δημιουργία ενός σώματος κειμένων μιμείται τον ερμηνευτικό κύκλο, έτσι ώστε οι αρχές συγκρότησής του να βελτιώνονται από τις πραγματικές συνθήκες συλλογής των δεδομένων (Γούτσος & Φραγκάκη, 2015, σ. 42).

Σύμφωνα με τις αρχές του ερευνητικού προγράμματος που αποτέλεσε την αφετηρία της δημιουργίας του, το *Σώμα κειμένων Νεοελληνικής Λογοτεχνίας* θα έπρεπε να περιλαμβάνει, για λόγους δειγματοληψίας, δύο ολόκληρα λογοτεχνικά έργα δημοσιευμένα σε κάθε έτος της χρονικής περιόδου 1927-1999 (73 έτη). Τα έργα αυτά μπορεί να είναι μυθιστορήματα, συλλογές διηγημάτων, ποιητικές συλλογές ή θεατρικά έργα, αλλά σε κάθε περίπτωση θα πρέπει η πρώτη δημοσίευσή τους να έχει γίνει στο συγκεκριμένο έτος αναφοράς και να μην αποτελούν λ.χ.

[25] https://inventory.clarin.gr/corpus/890?lang=el
[26] https://inventory.clarin.gr/corpus/994
[27] https://distantreading.github.io/ELTeC/gre/index.html
[28] https://inventory.clarin.gr

συλλογές έργων που έχουν δημοσιευθεί παλαιότερα ή επανεκδόσεις. Ως έτος αναφοράς θεωρείται το έτος δημοσίευσης και κυκλοφορίας του έργου και όχι το έτος συγγραφής του. Είναι αυτονόητο ότι κάθε έργο θα πρέπει να είναι αρκετά εκτεταμένο ώστε να δικαιολογεί τη συμπερίληψή του στο σώμα κειμένων.

Με βάση αυτές τις αρχές ένα βασικό μέλημα κατά τη δημιουργία του ΣΝΕΛ υπήρξε η χρονολόγηση των υπό συμπερίληψη δημοσιευμένων έργων, κάτι που δεν ήταν πάντοτε τόσο εύκολο όσο φαίνεται εκ πρώτης όψεως και μπορεί να προϋποθέτει εκτεταμένη φιλολογική έρευνα. Συγκεκριμένα, είναι ιδιαίτερα δύσκολο να εντοπιστούν λογοτεχνικά έργα στα ελληνικά για το 1941 λόγω των συνθηκών της Κατοχής, που, μεταξύ άλλων, περιλάμβαναν την έλλειψη τυπογραφικού χαρτιού και μελάνης (βλ. και Καστρινάκη 2015, σσ. 12-13). Οι πρώτες εκδόσεις της συγκεκριμένης χρονιάς είναι δυσεύρετες και η ψηφιοποίησή τους θα προϋπέθετε δυσανάλογο κόπο και προσπάθεια, ενώ πολλά έργα που αποδίδονται σε ορισμένα έτη έχουν στην πραγματικότητα δημοσιευτεί άλλοτε ή είναι διαθέσιμα μόνο μέσω μεταγενέστερων εκδόσεων.

Η τελευταία αυτή παρατήρηση επισημαίνει την συγκυριακή φύση της συγκρότησης σωμάτων κειμένων, καθώς η διαδικασία εξεύρεσης δεδομένων καθορίζεται από παράγοντες όπως η διαθεσιμότητα, η ευκολία της επεξεργασίας τους κ.λπ. Στο ΣΝΕΛ έγινε προσπάθεια αξιοποίησης ήδη ψηφιοποιημένων κειμένων σε μορφή επεξεργάσιμη ή σε μορφή εικόνας ή pdf από διαθέσιμα αποθετήρια (π.χ. *Ανέμη*), από πηγές με ελεύθερα διαθέσιμα κείμενα στο διαδίκτυο ή προσωπικές ψηφιακές συλλογές,[29] ενώ για τα υπόλοιπα έργα έγινε σάρωση. Για τα κείμενα σε μορφή εικόνας ή pdf ακολούθησε ψηφιοποίηση μέσω οπτικής αναγνώρισης χαρακτήρων. Η διαδικασία αυτή επίσης ήταν ιδιαίτερα προβληματική, καθώς η οπτική αναγνώριση χαρακτήρων στο πολυτονικό σύστημα δεν ήταν ικανοποιητική. Για τον σκοπό αυτό αναπτύχθηκε ειδική πλατφόρμα οπτικής αναγνώρισης χαρακτήρων πολυτονικών κειμένων, στην οποία γίνεται λεπτομερής αναφορά στην επόμενη ενότητα. Αξίζει, ωστόσο, εδώ να σημειώσουμε ότι, εξαιτίας της πολύπλοκης ιστορίας της ελληνικής τυπογραφίας, ορισμένα κείμενα στο ΣΝΕΛ είναι ψηφιοποιημένα με μονοτονικό ορθογραφικό σύστημα και άλλα με πολυτονικό, και από αυτά τα τελευταία ορισμένα με την εκδοχή του πολυτονικού που χρησιμοποιεί βαρεία («τριτονικό») και άλλα με την εκδοχή που δεν χρησιμοποιεί βαρεία («διτονικό»). Επιπρόσθετα, στα παλιότερα κείμενα γίνεται εκτεταμένη ή μικρή χρήση της υπογεγραμμένης.

Η επόμενη διαδικασία που ακολουθήθηκε για τη συγκρότηση του ΣΝΕΛ περιλαμβάνει τον καθαρισμό και τον έλεγχο με το πρωτότυπο κείμενο, καθώς και την τυποποίηση των δεδομένων με επισημείωση τίτλων και υπότιτλων, παραγράφων και στίχων για τα ποιητικά έργα. Η διαδικασία αυτή βρίσκεται σε πρόοδο και σε επόμενη φάση θα περιλάβει την επισημείωση αποσπασμάτων με πλάγια ή έντονα στοιχεία και άλλες τυπογραφικές ιδιαιτερότητες. Τα κείμενα κωδικοποιούνται σε μορφή XML και αποθηκεύονται με το έτος δημοσίευσης και τον τίτλο στο όνομα του αρχείου σε μορφή UTF-8.

Ο πλήρης κατάλογος των τίτλων που περιλαμβάνονται στο ΣΝΕΛ παρατίθεται στο Παράρτημα, ενώ στον Πίνακα 2 παρουσιάζονται λεπτομερή αριθμητικά στοιχεία.

---

[29] Θα θέλαμε να ευχαριστήσουμε την Τιτίκα Δημητρούλια, τον Νίκο Μαθιουδάκη και την Μαρία Ακριτίδου για την βοήθειά τους στον εντοπισμό πηγών.

Πίνακας 2: *Σύσταση του ΣΝΕΛ*

Τυπογραφικό σύστημα	**Αριθμός έργων**	**Αριθμός λέξεων**
Πολυτονικό	61	2.148.444
μυθιστορήματα	*44*	
διηγήματα	*7*	
ποιητική συλλογή	*8*	
θεατρικό έργο	*2*	
Μονοτονικό	85	4.298.430
μυθιστορήματα	*69*	
διηγήματα	*1*	
ποιητική συλλογή	*14*	
θεατρικό έργο	*1*	
ΣΥΝΟΛΟ	146	6.446.874

Όπως μπορεί κανείς να διαπιστώσει από το Παράρτημα, τα έργα που περιλαμβάνονται στο ΣΝΕΛ δεν περιορίζονται στον λογοτεχνικό κανόνα, αλλά είναι διαφόρων ειδών, καθώς περιλαμβάνουν αστυνομική, παιδική λογοτεχνία κ.λπ. και πλήθος διαφορετικών συγγραφέων, ενώ ταυτόχρονα ορισμένοι συγγραφείς εκπροσωπούνται με περισσότερα του ενός έργα. Η ποικιλία αυτή, αν και δεν μπορεί να υποστηριχθεί ότι είναι αντιπροσωπευτική του λογοτεχνικού είδους στο σύνολό του, εξυπηρετεί τους σκοπούς του ερευνητικού προγράμματος για το οποίο δημιουργήθηκε το ΣΝΕΛ και ταυτόχρονα προσφέρει τη βάση για περαιτέρω ανάπτυξη του σώματος κειμένων λ.χ. σε ένα αμιγές σώμα κειμένων μυθιστορημάτων ή ένα πιο ισορροπημένο σώμα κειμένων με διαφορετικά λογοτεχνικά υποείδη.

4. Ψηφιοποίηση πολυτονικών κειμένων

Όπως αναφέρθηκε ήδη, ένα από τα βασικά προβλήματα στη δημιουργία ικανοποιητικών σωμάτων κειμένων στα ελληνικά με διαχρονικά δεδομένα είναι η έλλειψη μιας αξιόπιστης εφαρμογής για την οπτική αναγνώριση των πολυτονικών χαρακτήρων στους οποίους έχει τυπωθεί το μεγαλύτερο μέρος των κειμένων στα ελληνικά. Σε αυτά περιλαμβάνονται όλα τα κείμενα του 19ου αιώνα και του 20ού αιώνα, τουλάχιστον έως το 1982, αλλά σε ορισμένες περιπτώσεις και πολύ αργότερα (βλ. Παράρτημα). Για τον λόγο αυτό, απαραίτητη προϋπόθεση για τη δημιουργία ενός διαχρονικού σώματος κειμένων στα ελληνικά είναι η ανάπτυξη μιας διαδικτυακής εφαρμογής για την αναγνώριση πολυτονικών χαρακτήρων που θα διαθέτει αποτελεσματικότητα (ακριβή αποτελέσματα με μικρό ποσοστό λαθών), εμβέλεια (το μεγαλύτερο δυνατό εύρος πολυτονικών γραμματοσειρών που έχουν χρησιμοποιηθεί στην ελληνική τυπογραφία), διασυστηματικότητα (ικανότητα να προσαρμόζεται σε μια ποικιλία κειμενικών πηγών και σε διαφορετικά λειτουργικά συστήματα), διαθεσιμότητα (ελεύθερα διαθέσιμο εργαλείο γλωσσικής τεχνολογίας ανοικτού κώδικα) και επεκτασιμότητα.

Οι τρέχουσες προσεγγίσεις που χρησιμοποιούν υφιστάμενα προγράμματα οπτικής αναγνώρισης χαρακτήρων ή έχουν αποπειραθεί να αναπτύξουν συναφή εξειδικευμένα προγράμματα έχουν αποτύχει να δώσουν μια αποτελεσματική απάντηση στο πρόβλημα – για παράδειγμα οι Gatos et al. (2011), που βασίζονται στην πρωτότυπη εργασία του Σταματάτου (2011), αναφέρουν επιτυχή αναγνώριση λέξεων μόλις 63%, ενώ τα σχετικά εργαλεία δεν διαθέτουν ανοικτή

πρόσβαση στην κοινότητα. Η χρήση προγραμμάτων οπτικής αναγνώρισης χαρακτήρων γενικού σκοπού απαιτεί επομένως εκτεταμένη εκπαίδευση ή επεξεργασία εκ των υστέρων χωρίς εγγυημένα αποτελέσματα. Η αιχμή της έρευνας βρίσκεται στις προσπάθειες κλασικών φιλολόγων να αναπτύξουν πολυτονικά συστήματα αναγνώρισης χαρακτήρων για αρχαία ελληνικά κείμενα, καθώς κάτι τέτοιο αποτελεί άμεση αναγκαιότητα γι' αυτούς. Ειδικότερα, οι Boschetti et al. (2009) και Robertson (2013) αναφέρονται σε ερευνητικές απόπειρες με αυτόν το στόχο, οι οποίες ωστόσο απέχουν πολύ από το να παραγάγουν ένα συνολικά ικανοποιητικό αποτέλεσμα, απαιτώντας εκτεταμένη περαιτέρω εκπαίδευση.

Για τους σκοπούς της δημιουργίας του ΣΝΕΛ χρησιμοποιήθηκε η βιβλιοθήκη «Κάλχας»[30] που αναπτύχθηκε στο πλαίσιο του ερευνητικού προγράμματος «Διαχρονικό σώμα ελληνικών κειμένων του 19ου αιώνα» στο πλαίσιο της δράσης με τίτλο «Έρευνα στις μετακλασικές σπουδές στο πλαίσιο του καταπιστεύματος Κωνσταντίνου Τσαγκαδά» του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών (2022-2023). Η μεθοδολογία του έργου ακολουθεί την τεχνική της μηχανικής μάθησης (machine learning), η οποία στηρίζεται στη δυνατότητα δημιουργίας υπολογιστικών μοντέλων από ένα σύνολο δεδομένων (βλ. Mitchell, 1997· Bengio, 2009). Η σημασία της μηχανικής μάθησης έγκειται ακριβώς στη δυνατότητα να επεκτείνεται το υπολογιστικό μοντέλο που έχει χρησιμοποιηθεί ώστε να «μαθαίνει», δηλαδή να αλλάζει τη συμπεριφορά του κατά τέτοιο τρόπο ώστε να αποδίδει καλύτερα στο μέλλον (Witten & Frank, 2000). Συμπεριλαμβάνει ένα ευρύ φάσμα μεθόδων που βασίζονται σε συγκεκριμένους αλγορίθμους, οι οποίοι εφαρμόζονται σε αναπαραστάσεις δεδομένων (επιβλεπόμενης, ημι-επιβλεπόμενης ή μη επιβλεπόμενης) εκμάθησης.

Οι σχετικοί αλγόριθμοι που χρησιμοποιήθηκαν στο πρόγραμμα «Κάλχας» ανήκουν στην κατηγορία της βαθείας εκμάθησης (deep learning, βλ. Bengio et al., 2013· Schmidhuber, 2015). Αφετηρία του προγράμματος ήταν προηγούμενες έρευνες στην οπτική αναγνώριση χαρακτήρων ιστορικών εγγράφων όπως των Simistira et al. (2015), Katsouros et al. (2016), Robertson & Boschetti (2017) και Sichani et al. (2019), αξιοποιώντας, ωστόσο, σε αντίθεση με τις έρευνες αυτές, τις προόδους στην αρχιτεκτονική πολύπλοκων νευρωνικών δικτύων που επέφερε η τεχνική της βαθείας μάθησης. Ουσιαστικά επεκτείνουμε την εργασία των Simistira et al. (2015), αντικαθιστώντας την αρχιτεκτονική τους και καταγράφοντας έτσι σημαντική βελτίωση στα αποτελέσματα αξιολόγησης, που μειώνουν ουσιαστικά το ποσοστό σφάλματος χαρακτήρων (Character Error Rate) στο περίπου 1,18% κατά μέσο όρο στο δημοσίως διαθέσιμο σύνολο δεδομένων τους.[31]

Στο Perifanos & Goutsos (2025) περιγράφουμε αναλυτικά την πλατφόρμα *Λόγιος*, που αναπτύχθηκε ως αποτέλεσμα αυτής της έρευνας και η οποία βελτιώνει σημαντικά την απόδοση του τελικού μοντέλου.[32] Για τη διευκόλυνση της εκπαίδευσης και της αξιολόγησης, αναπτύξαμε μια ειδική εφαρμογή για τη μεταφόρτωση εγγράφων, την κατάτμηση (segmentation), την οπτική αναγνώριση χαρακτήρων (OCR) και τη συλλογή επισημειώσεων (labels). Παρότι η εφαρμογή αρχικά σχεδιάστηκε με στόχο την υποστήριξη της δημιουργίας πολυτονικών σωμάτων κειμένων, σύντομα συνειδητοποιήσαμε ότι ένα τέτοιο εργαλείο θα είχε

[30] https://github.com/kperi/Kalchas

[31] Μια συγκρίσιμη προσπάθεια γίνεται στο Tzogka et al. (2021) με αναφερόμενο χαμηλότερο ποσοστό ακρίβειας.

[32] Η πλατφόρμα βρίσκεται σε δοκιμαστική φάση και θα είναι σύντομα διαθέσιμη στην ιστοσελίδα: https://logios.phil.uoa.gr

μεγάλη αξία για την ακαδημαϊκή κοινότητα. Για τον λόγο αυτό, αποφασίσαμε να επεκτείνουμε και να βελτιώσουμε περαιτέρω την εφαρμογή και να τη διαθέσουμε δωρεάν στο κοινό.

Η πλατφόρμα *Λόγιος* έχει υλοποιηθεί με χρήση της βιβλιοθήκης Kalchas OCR και της βιβλιοθήκης Streamlit της Python ως μηχανής διεπαφής χρήστη (UI engine). Χρησιμοποιούμε OpenCV για επεξεργασία εικόνας, Kraken για αποθορυβοποίηση εγγράφων και τμηματοποίηση γραμμών, καθώς και DocLayout-YOLO για ανάλυση διάταξης (layout analysis). Το σύστημα υποστηρίζει την εξαγωγή σελίδων από αρχεία PDF και παρέχει τόσο χειρωνακτική όσο και αυτόματη ανίχνευση διάταξης και οπτική αναγνώριση χαρακτήρων. Το παραγόμενο κείμενο είναι επεξεργάσιμο, γεγονός που επιτρέπει τη χρήση του εργαλείου για επισημείωση εγγράφων. Επιπλέον, παρέχουμε τη δυνατότητα εκπαίδευσης ή προσαρμογής (fine-tuning) των υπαρχόντων μοντέλων μας με δεδομένα που αποκλίνουν από αυτά που χρησιμοποιήθηκαν κατά την αρχική εκπαίδευση του «Κάλχα». Επιπλέον, διαθέτουμε ένα νέο σύνολο δεδομένων αποτελούμενο από 6.796 δείγματα εκπαίδευσης και παρέχουμε ανοικτή πρόσβαση (open-source) στα βέλτιστα μοντέλα και τα δεδομένα μας, ενσωματωμένα σε μια εύχρηστη βιβλιοθήκη Python.

Προσδοκούμε ότι η περαιτέρω βελτίωση της πλατφόρμας *Λόγιος* θα προσφέρει μια οριστική και αποτελεσματική απάντηση στο πρόβλημα της οπτικής αναγνώρισης χαρακτήρων για ελληνικές πολυτονικές γραμματοσειρές, που έχει ταλανίσει επί καιρό την ελληνική ερευνητική κοινότητα, και θα δώσει έτσι τη δυνατότητα για εύκολη και μαζική ψηφιοποίηση παλαιότερων ελληνικών κειμένων. Πρόκειται για ένα απαραίτητο έργο υποδομής, που καλύπτει μια ανάγκη ύψιστης προτεραιότητας και αναμένεται να προσφέρει σημαντική ώθηση στις ελληνικές ψηφιακές ανθρωπιστικές επιστήμες.

5. Προκαταρκτικά ευρήματα

Όπως αναφέρθηκε πιο πάνω, η συγκρότηση του ΣΝΕΛ είναι συγκυριακή με αποτέλεσμα να έχουν συμπεριληφθεί τόσο μονοτονικά όσο και πολυτονικά κείμενα σύμφωνα με τη διαθεσιμότητά τους. Αυτό δυσχεραίνει την επεξεργασία του με έτοιμα λογισμικά όπως το AntConc ή το Voyant Tools (βλ. Δημητρούλια κ.ά., 2024, σσ. 201 κ.εξ.· Φραγκάκη 2024, 2025) και απαιτεί την ομογενοποίησή του σε επόμενη φάση. Ωστόσο, μπορούμε εδώ να αναφερθούμε ενδεικτικά σε ορισμένα προκαταρκτικά ευρήματα που προκύπτουν από την ανάλυση του ΣΝΕΛ.

Καταρχάς, η ανάλυση του καταλόγου συχνότητας του συνόλου του ΣΝΕΛ στην παρούσα φάση επιβεβαιώνει τα ευρήματα της Φραγκάκη (υπό προετοιμασία) για τα χαρακτηριστικά του λογοτεχνικού κειμενικού είδους στα ελληνικά. Πιο συγκεκριμένα, το συννεφόλεξο στην Εικόνα 1 παρουσιάζει τις 50 συχνότερες λέξεις στο ΣΝΕΛ ύστερα από την αφαίρεση των πρώτων 200 γραμματικών λέξεων. Όπως μπορεί να διαπιστώσει κανείς, στους πιο συχνούς λεξικούς τύπους συμπεριλαμβάνονται μέλη του σώματος (*μάτια, χέρι, κεφάλι, καρδιά, πόδια*), αναφορά στο *σπίτι*, λέξεις που δηλώνουν χρόνο (*στιγμή, χρόνια, ώρα, μέρες, φορά*), κοινά ρήματα με μεγάλη συχνότητα (*είχα, θέλω, πει, πάρει, γίνει, κάνει, ξέρω, ρώτησε*), όπως και ουσιαστικά όπως *φωνή, ζωή, δρόμο*. Οπωσδήποτε, τα ευρήματα αυτά είναι μόνο ενδεικτικά καθώς ο ίδιος τύπος μπορεί να εμφανίζεται στον κατάλογο λ.χ. με οξεία, βαρεία ή χωρίς καθόλου τόνο, καθιστώντας πολύπλοκο τον ακριβή υπολογισμό της συχνότητάς του. Η γενική εικόνα όμως που προκύπτει ενισχύει το προφίλ της λογοτεχνίας ως κειμενικού είδους που ασχολείται με τον

άνθρωπο ως σωματική παρουσία και τη χρονική αλλά και τοπική διάσταση της δράσης του. Ιδιαίτερο ενδιαφέρον έχει και η εμφάνιση των λέξεων *γυναίκα* και *κυρία*. Ίσως η εμφάνιση αυτών (και όχι λ.χ. των τύπων *άντρας* και *κύριος*) με τόσο υψηλή συχνότητα να οφείλεται στο ότι οι θηλυκές οντότητες είναι χαρακτηρισμένες στα ελληνικά, κάτι που επιβάλλει τη ρητή αναφορά στο φύλο, όπως έχει υποστηριχθεί για άλλα κειμενικά είδη στα ελληνικά (Goutsos & Fragaki, 2009, σ. 321). Μια τέτοια υπόθεση όμως χρειάζεται σαφώς περισσότερη ποιοτική έρευνα για να υποστηριχθεί.

Εικόνα 1: *Κατάλογος συχνότητας του ΣΝΕΛ (συννεφόλεξο)*

Αν προχωρήσουμε σε μια λεπτομερέστερη ανάλυση των καταλόγων συχνότητας, μπορούμε να διαπιστώσουμε ότι τα μέλη του σώματος κυριαρχούν και στους καταλόγους λεξικών συμπλεγμάτων τόσο στο μονοτονικό όσο και στο πολυτονικό τμήμα του ΣΝΕΛ. Τα λεξικά συμπλέγματα είναι συνεχόμενες ακολουθίες τριών ή περισσότερων λεξικών μονάδων που επαναλαμβάνονται με στατιστικά σημαντικό τρόπο (Γούτσος & Φραγκάκη, 2015, σσ. 86 κ.εξ.) και παρουσιάζουν ιδιαίτερο ενδιαφέρον γιατί δίνουν έμφαση στη συχνή φρασεολογία που συνθέτει τα δομικά συστατικά ενός κειμενικού είδους.

Οι Πίνακες 3, 4 και 5 παρουσιάζουν τα πρώτα σε συχνότητα λεξικά συμπλέγματα 3 λέξεων στο μονοτονικό και στο πολυτονικό τμήμα του ΣΝΕΛ και συμπλέγματα 4 λέξεων στο μονοτονικό τμήμα του ΣΝΕΛ.

Πίνακας 3: *Συχνότερα λεξικά συμπλέγματα 3 λέξεων στο μονοτονικό τμήμα του ΣΝΕΛ*

Λεξικά συμπλέγματα	*Συχνότητα*	*Κάλυψη*
τα μάτια του	1053	57
τα μάτια της	599	56
το χέρι του	592	56
το κεφάλι του	555	56
τα χέρια του	512	58
η φωνή του	474	52

την ώρα που	461	59
δεν μπορούσε να	420	50
για να μην	384	60
να σου πω	344	51
κι άρχισε να	319	40
δεν μπορώ να	317	53
το πρόσωπό του	314	44
δεν μπορεί να	310	53
κάτω από το	306	43
για πρώτη φορά	305	50
τα μάτια μου	304	40
θα μπορούσε να	293	51
κι οι δυο	287	34
την άλλη μέρα	287	48
το χέρι της	279	47
η φωνή της	278	47
στο σπίτι του	276	45
το στόμα του	264	45
η γυναίκα του	260	47

Πίνακας 4: *Συχνότερα λεξικά συμπλέγματα 3 λέξεων στο πολυτονικό τμήμα του ΣΝΕΛ*

Λεξικά συμπλέγματα	*Συχνότητα*	*Κάλυψη*
τὰ μάτια του	292	33
τὰ μάτια της	212	31
τὴν ἄλλη μέρα	196	29
τὸ κεφάλι του	188	25
ἡ μητέρα μου	169	19
τὴν ὥρα ποὺ	150	26
ὁ πατέρας μου	148	25
τὰ χέρια του	144	29
γιὰ νὰ μὴν	138	29
γιὰ νὰ μὴ	136	28
τὸ χέρι του	132	33
δὲν μποροῦσε νὰ	131	24
τὸν ἑαυτό του	129	22
θὰ μποροῦσε νὰ	128	26
ἡ μάνα μου	122	18
νὰ σοῦ πῶ	119	22
οἱ δυό τους	110	21
ὁ πατέρας της	108	24
μὲ τὰ μάτια	103	29
μὲ τὰ χέρια	98	29

Πίνακας 5: *Συχνότερα λεξικά συμπλέγματα 4 λέξεων στο μονοτονικό τμήμα του ΣΝΕΛ*

Λεξικά συμπλέγματα	*Συχνότητα*	*Κάλυψη*
και τα μάτια του	111	32

προς το μέρος του	108	22
δόξα σοι ο θεός	97	15
τι να σου πω	69	20
η πρώτη φορά που	66	32
από το κεφάλι του	61	18
από την άλλη μεριά	60	27
προς το μέρος της	60	24
ήταν η πρώτη φορά	59	29
με τη γυναίκα του	57	27
από δω και πέρα	55	24
για όνομα του θεού	55	15
ο ένας τον άλλο	55	29
από τα χέρια του	52	19
κι οι δυο τους	51	18
από το σπίτι του	50	27
εδώ που τα λέμε	50	19

Εκτός από τα μέλη του σώματος και τις χρονικές φράσεις (*για πρώτη φορά, την ώρα που, την άλλη μέρα*), στα λεξικά συμπλέγματα της λογοτεχνίας εμφανίζονται φράσεις τροπικότητας (*δεν/θα μπορούσε/μπορώ να*), μέλη της οικογένειας (*ο πατέρας μου/της, η μητέρα/μάνα μου*), αλλά και χαρακτηριστικές φράσεις από την αναπαράσταση του προφορικού λόγου (*να σου πω, δόξα σοι ο θεός, για όνομα του θεού, εδώ που τα λέμε*). Και σε αυτή την περίπτωση είναι απαραίτητη ενδελεχέστερη έρευνα για να διαπιστωθεί το εύρος των λεξικών συμπλεγμάτων και η λειτουργία τους στη λογοτεχνία.

Ιδιαίτερο ενδιαφέρον παρουσιάζουν τα ποσοτικά δεδομένα για τα έργα που συμπεριλαμβάνονται στο ΣΝΕΛ στο σύνολό τους. Για παράδειγμα, στον Πίνακα 6 παρουσιάζεται η λεξιλογική πυκνότητα στα λογοτεχνικά κείμενα του ΣΝΕΛ, που εκφράζει την αναλογία διαφορετικών τύπων στο σύνολο των δειγμάτων που περιλαμβάνει κάθε έργο.

Πίνακας 6: *Λεξιλογική πυκνότητα στα έργα του ΣΝΕΛ*

Μέγιστη	**Ελάχιστη**
Μαραμπού (0.636)	Ο τελευταίος πειρασμός (0.097)
Ερήμην (0.615)	Στα μυστικά του βάλτου (0.099)
Μπολιβάρ (0.522)	Ο Καπετάν Μιχάλης (0.103)
Η μητέρα μου στην εκκλησία (0.452)	Ο Χριστός ξανασταυρώνεται (0.113)
Πούσι (0.515)	Αιολική γη (0.117)
Στόχος (0.506)	Έγκλημα στο Κολωνάκι (0.124)
Τα ελεγεία της οξώπετρας (0.429)	Νυχτερινό δελτίο (0.129)
Το ξύλινο παλτό (0.421)	Η αρραβωνιαστικιά του Αχιλλέα (0.131)
Παραμυθένια πολιτεία (0.417)	Τρελαντώνης (0.132)
Επιτάφιος (0.408)	Γαλήνη (0.135)

Όπως ίσως θα ανέμενε κανείς, τα ποιητικά έργα του ΣΝΕΛ έχουν τη μεγαλύτερη πυκνότητα, με τη μέγιστη να εμφανίζεται στη συλλογή *Μαραμπού* του Νίκου

Καββαδία. Με τον τρόπο αυτό διαπιστώνουμε εμπειρικά την εντύπωση της συμπύκνωσης και της έλλειψης πλεονασμού στην ποίηση. Στην αντίθετη στήλη με την ελάχιστη πυκνότητα, δηλαδή τον μεγαλύτερο βαθμό επανάληψης του λεξιλογίου, βρίσκονται τρία μυθιστορήματα του Ν. Καζαντζάκη, δύο παιδικά μυθιστορήματα της Π. Δέλτα, δύο έργα του Η. Βενέζη και ένα της Α. Ζέη, καθώς και δύο αστυνομικά μυθιστορήματα. Η ενδιαφέρουσα αυτή συνύπαρξη μας επιτρέπει να διατυπώσουμε ερευνητικές υποθέσεις που συνδέονται με την ευκολία πρόσβασης και τη δημοφιλία των έργων αυτών.

Εξειδικεύοντας τις υποθέσεις αυτές, μπορούμε να διερευνήσουμε ποσοτικά στοιχεία όπως τον μέσο όρο λέξεων ανά πρόταση, όπως φαίνονται στον Πίνακα 7.

Πίνακας 7: *Μέσος όρος λέξεων ανά πρόταση στα έργα του ΣΝΕΛ*

Υψηλός	**Χαμηλός**
Βιολί για μονόχειρα (1452.0)	Το νούμερο 31328 (8.2)
Παραλογαίς (251.0)	Ο Γκιούλιβερ στη χώρα των γιγάντων (8.2)
Εγχειρίδιο ευθανασίας (77.4)	Βάρδια (8.4)
Ο τυφλός με τον λύχνο (73.0)	Ο όρκος του πεθαμένου (8.9)
Μικρά Αγγλία (53.6)	Ιστορία ενός αιχμαλώτου (9.0)
Τα ρέστα (39.7)	Ανθρωποφύλακες (9.2)
Η γραμμή του ορίζοντος (34.6)	Εικοστός αιώνας (9.7)
Ραμπαστέν (34.1)	Συναξάρι Ανδρέα Κορδοπάτη (9.9)
Ο τελευταίος πειρασμός (33.8)	Ζιγκ ζαγκ στις νεραντζιές (10.2)
Πεισίστρατος (32.9)	Τα πλοία δεν άραξαν (10.3)

Από τα στοιχεία του Πίνακα 7 προκύπτει ότι οι μεγαλύτερες σε αριθμό λέξεων προτάσεις εντοπίζονται στα ποιητικά έργα του Τ. Λειβαδίτη με πρώτη την ποιητική συλλογή *Βιολί για μονόχειρα*, ο δείκτης για την οποία είναι σχεδόν 6 φορές μεγαλύτερος από το επόμενο έργο. Οπωσδήποτε, ο ρόλος της στίξης στον οποίο βασίζεται ο δείκτης αυτός είναι διαφορετικός στα ποιητικά από τα πεζά έργα και τα όρια του ποιητικού στίχου ορίζουν μια επιπλέον μονάδα που θα πρέπει να υπολογιστεί. Τα πεζά έργα με τον υψηλότερο μέσο όρο λέξεων ανά πρόταση είναι ετερόκλιτα: πεζογραφήματα ευρείας κυκλοφορίας όπως η *Μικρά Αγγλία* και *Τα ρέστα* εμφανίζονται μαζί με πιο πειραματικά ή «δύσκολα» έργα όπως *Η γραμμή του ορίζοντος* ή ο *Πεισίστρατος*. Από την άλλη πλευρά, μικρότερο αριθμό λέξεων ανά πρόταση έχουν επίσης τόσο μοντερνιστικά έργα όπως ο *Εικοστός αιώνας* της Μ. Αξιώτη, μυθιστορήματα που βασίζονται στην πρωτοπρόσωπη αφήγηση όπως η *Ιστορία ενός αιχμαλώτου* και το *Συναξάρι Ανδρέα Κορδοπάτη*, ποιητικά και θεατρικά έργα (*Ο όρκος του πεθαμένου* του Ζ. Παπαντωνίου) και στη χαμηλότερη θέση του δείκτη ένα μυθιστόρημα του Η. Βενέζη.

Μια πιο πολύπλοκη ένδειξη για την αναγνωσιμότητα των έργων παρέχεται από το πρόγραμμα Voyant Tools και στηρίζεται στον δείκτη Coleman–Liau, που υπολογίζει το μέσο όρο των γραμμάτων ανά 100 λέξεις σε συνάρτηση με το μέσο όρο των προτάσεων ανά 100 λέξεις, στοχεύοντας ακριβώς στη μέτρηση της ευκολίας της ανάγνωσης. Ο Πίνακας 8 παρουσιάζει τα 10 έργα με τον υψηλότερο δείκτη, δηλαδή τη μεγαλύτερη δυσκολία ανάγνωσης και τα 10 με τον χαμηλότερο δείκτη αναγνωσιμότητας.

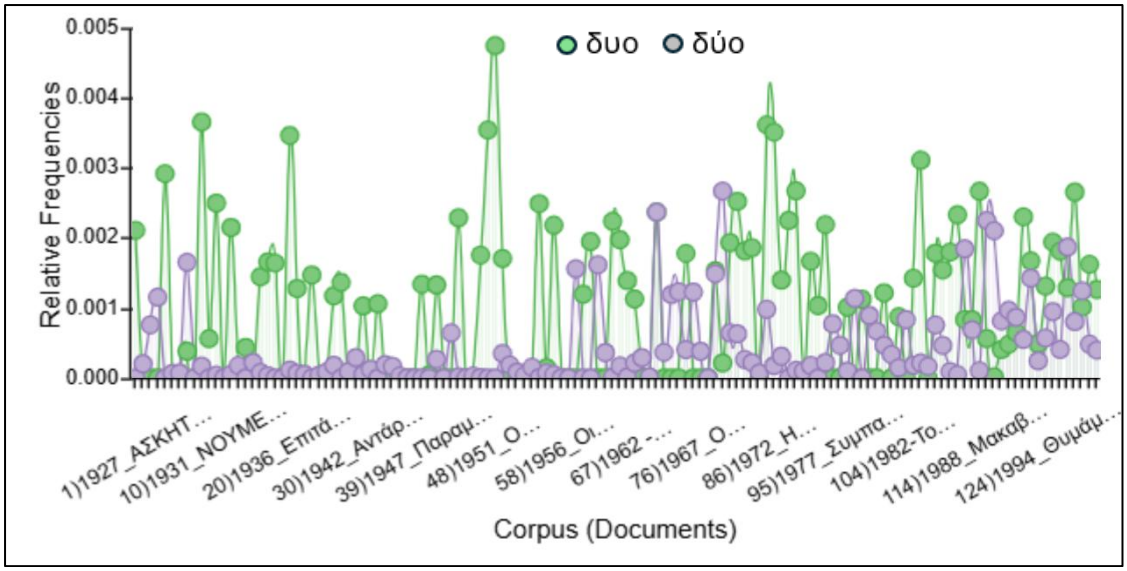
Πίνακας 8: *Δείκτης αναγνωσιμότητας στα έργα του ΣΝΕΛ*

Υψηλή	**Χαμηλή**
Αναφορά περιπτώσεων (13.516)	Επιτάφιος (3.196)
Στόχος (13.423)	Το μεγάλο μας τσίρκο (5.620)
Μικρά Αγγλία (13.318)	Ωκεανός (6.259)
Η γυναίκα κουρσάρος (12.537)	Τρακτέρ (6.653)
Η γραμμή του ορίζοντος (12.419)	Βάρδια (6.649)
Η παρτίδα (12.013)	Γραγκάντα (6.836)
Ραμπαστέν (11.973)	Πούσι (6.872)
Φημισμένοι άντρες και λησμονημένοι (11.819)	Ο άνθρωπος με το γαρύφαλλο (6.941)
Αρμαγεδδών (11.595)	Βαμμένα κόκκινα μαλλιά (7.215)
Πεισίστρατος (11.569)	Αιολική γη (7.306)

Με βάση το δείκτη αναγνωσιμότητας, τα μοντερνιστικά έργα του Α. Σχινά *Αναφορά περιπτώσεων* και *Η παρτίδα*, ο *Πεισίστρατος* του Γ. Χειμωνά και *Η γραμμή του ορίζοντος* του Χ. Βακαλόπουλου, αλλά και η ποιητική συλλογή *Στόχος* του Μ. Αναγνωστάκη δεν αποτελούν έκπληξη, σε αντίθεση με τα μυθιστορήματα της Καρυστιάνη, του Τσιφόρου, του Κόντογλου και νεότερων συγγραφέων, που θα ανέμενε κανείς να είναι πιο ευανάγνωστα. Στον αντίποδα, είναι αναμενόμενο ποιητικές συλλογές και θεατρικά να διαθέτουν χαμηλό δείκτη αναγνωσιμότητας, αλλά έχει ενδιαφέρον ότι τέσσερις συλλογές του Γ. Ρίτσου βρίσκονται στην κατηγορία αυτή, μαζί με δημοφιλή έργα όπως το μυθιστόρημα *Βαμμένα κόκκινα μαλλιά* του Κ. Μουρσελά και δύο έργα του Η. Βενέζη. Είναι προφανές ότι υπάρχει μεγάλο περιθώριο για τη διερεύνηση της υπόθεσης της αναγνωσιμότητας με αφετηρία τα ποσοτικά αυτά στοιχεία.

Τέλος, η συγκρότηση του ΣΝΕΛ στο διαχρονικό άξονα 73 ετών προσφέρεται για εξειδικευμένες παρατηρήσεις σχετικές με τη γλωσσική αλλαγή εντός του χρονικού αυτού διαστήματος, τόσο από γλωσσική πλευρά όσο και από την άποψη του περιεχομένου. Στην Εικόνα 2, για παράδειγμα, παρουσιάζεται η συχνότητα των τύπων *δυο* και *δύο* στον άξονα του χρόνου.

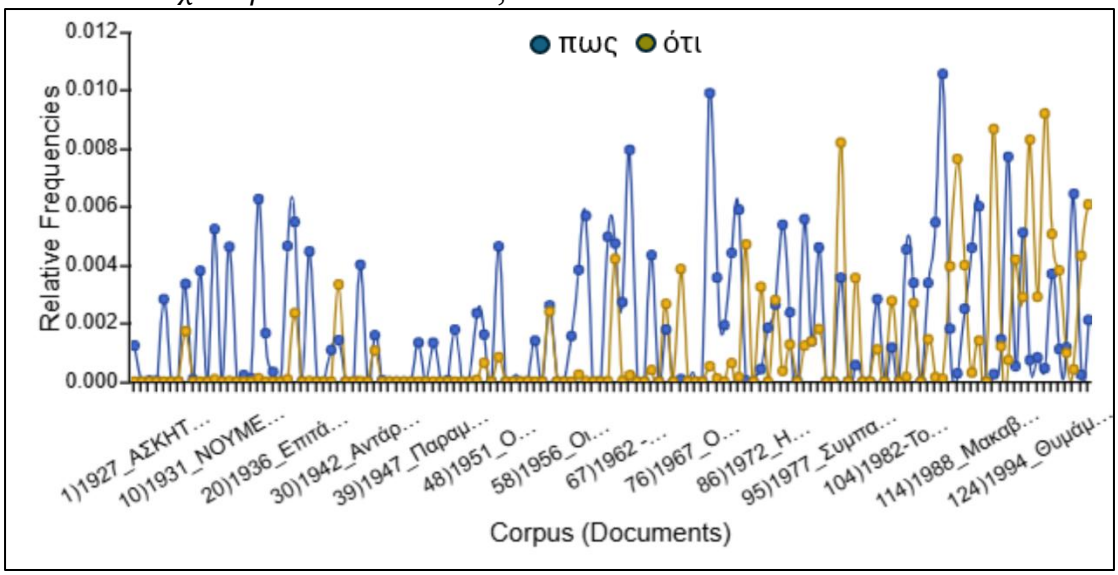
Εικόνα 2: *Συχνότητα των τύπων δυο και δύο στο ΣΝΕΛ*



Σε γενικές γραμμές μπορούμε να διακρίνουμε ότι ο προτιμώμενος τύπος του αριθμητικού για το μεγαλύτερο διάστημα στα λογοτεχνικά έργα του ΣΝΕΛ είναι το *δυο*, ενώ το *δύο* εμφανίζεται με σχεδόν ισότιμη συχνότητα στα έργα της δεκαετίας του 1960, του 1980 και του 1990. Επομένως, φαίνεται ότι ο αμαρκάριστος τύπος για την ελληνική λογοτεχνία του 20ού αιώνα είναι ο *δυο*.

Η περίπτωση του *πως* και *ότι*, που παρουσιάζεται στην Εικόνα 3, είναι παρόμοια, εδώ όμως έχουμε την αντικατάσταση του *πως* από το *ότι* από τα τέλη της δεκαετίας του 1980 και μετά.

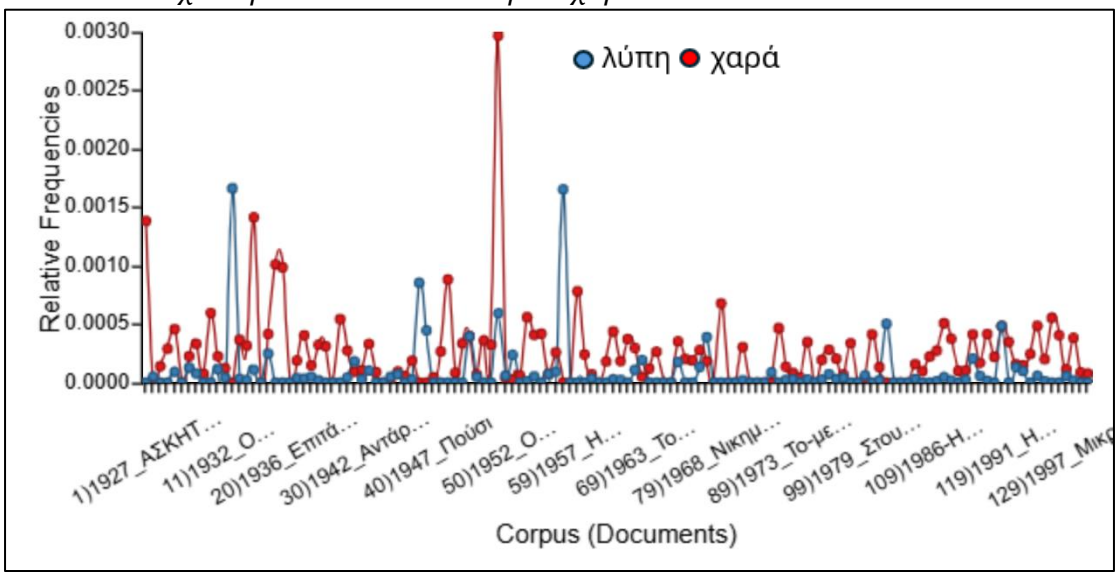
Εικόνα 3: *Συχνότητα των τύπων πως και ότι στο ΣΝΕΛ*



Τα ποσοτικά αυτά δεδομένα επιβεβαιώνουν τις παρατηρήσεις που έχουν γίνει για τους συμπληρωματικούς δείκτες *πως* και *ότι* στη βιβλιογραφία, σύμφωνα με τις οποίες το *πως* αποτελεί την κατεξοχήν γλωσσική επιλογή της λογοτεχνίας (και ειδικότερα της ποίησης), σε αντίθεση με τα υπόλοιπα κειμενικά είδη (Γούτσος 2025, Γούτσος υπό προετοιμασία).

Ανάλογες παρατηρήσεις μπορούμε να κάνουμε για λεξικούς τύπους όπως οι τύποι *λύπη* και *χαρά*, η εξέλιξη της συχνότητας των οποίων παρουσιάζεται στην Εικόνα 4.

Εικόνα 4: *Συχνότητα των τύπων λύπη και χαρά στο ΣΝΕΛ*



Σύμφωνα με τα στοιχεία αυτά, ο τύπος χαρά κυριαρχεί από άποψη συχνότητας σε σχέση με τον τύπο λύπη σε όλη τη χρονική περίοδο των έργων του ΣΝΕΛ με εξαίρεση κάποιες σποραδικές εξάρσεις του δεύτερου τύπου σε μεμονωμένα έργα. Είναι σαφές ότι δεν μπορούμε να εξαγάγουμε ασφαλή συμπεράσματα από δύο μόνο επιλεγμένους τύπους από τα αντίστοιχα σημασιολογικά πεδία για τη διακύμανση των συναισθηματικών αναπαραστάσεων στον 20ό αιώνα, αλλά τέτοια στοιχεία μπορούν να δώσουν ένα αδρό έναυσμα για τη διατύπωση συγκροτημένων ερευνητικών υποθέσεων παρόμοιων με αυτές του ερευνητικού προγράμματος για τη διαχρονική σχέση του πολιτισμού, της ψυχολογίας και της κοινωνικής οικονομίας, που αποτέλεσε την αφορμή για τη δημιουργία του ΣΝΕΛ.

Ευρύτερα, θα πρέπει να σημειώσουμε ότι η σύσταση και οι ιδιαιτερότητες του ΣΝΕΛ δεν επιτρέπουν να επιτευχθεί εύκολα ο στόχος των εκτεταμένων θεματικών, υφολογικών ή άλλων προσεγγίσεων. Ωστόσο, το σώμα κειμένων προσφέρει μια βασική αφετηρία και έναν πολύτιμο γλωσσικό πόρο που θα επιτρέψει σε άλλες μελλοντικές ερευνητικές προσπάθειες και εφαρμογές μια πιο ολοκληρωμένη ερμηνευτική ανάλυση πτυχών της νεοελληνικής λογοτεχνίας.

6. Συμπεράσματα και προοπτικές

Στο άρθρο αυτό παρουσιάσαμε το πλαίσιο και τις αρχές συγκρότησης ενός νέου γλωσσικού πόρου για τη λογοτεχνία στα ελληνικά, του *Σώματος κειμένων Νεοελληνικής Λογοτεχνίας* (ΣΝΕΛ). Όπως διαπιστώθηκε, ενώ στις περισσότερες ευρωπαϊκές γλώσσες υπάρχει πληθώρα πόρων για τη μελέτη της λογοτεχνίας με μεθόδους των Ψηφιακών Ανθρωπιστικών Επιστημών που δίνουν έμφαση στην εξ αποστάσεων ανάγνωση, κάτι τέτοιο δεν συμβαίνει για τα ελληνικά, για τα οποία είναι διαθέσιμα μόνο τα λογοτεχνικά υπο-σώματα δύο γενικών σωμάτων κειμένων. Το κενό αυτό έρχεται να καλύψει το ΣΝΕΛ, που περιλαμβάνει 146

λογοτεχνικά έργα, δύο για κάθε έτος από το 1927 έως το 1999, με συνολικό αριθμό λέξεων (δειγμάτων) γύρω στα 6,5 εκατομμύρια.

Πρόκειται για έναν μείζονα πόρο για τη διαχρονική μελέτη της λογοτεχνίας στα ελληνικά, ο οποίος αναμένεται να αναπτυχθεί περαιτέρω με τη χρήση της πλατφόρμας *Λόγιος*, που επιτρέπει τη μαζική ψηφιοποίηση έργων σε πολυτονικό τυπογραφικό σύστημα, κάτι που δεν ήταν εφικτό έως σήμερα. Στόχος της ερευνητικής προσπάθειας που οδήγησε στο ΣΝΕΛ είναι η ελεύθερη πρόσβαση στα δεδομένα μέσω μιας πλατφόρμας όπως αυτή του ΣΕΚ ή της *Ανεμόσκαλας*, που επιτρέπει την εύρεση συγκεκριμένων πληροφοριών, όχι όμως την ανάκληση ολόκληρων των κειμένων, τα οποία για προφανείς λόγους πνευματικών δικαιωμάτων μπορεί να μην είναι ελεύθερα διαθέσιμα. Έως τότε η πρόσβαση στο ΣΝΕΛ θα είναι εφικτή σε ερευνητές που αναζητούν συγκεκριμένα αποτελέσματα για ακαδημαϊκούς σκοπούς μέσω επικοινωνίας με τους δημιουργούς και διαχειριστές του σώματος κειμένων.

Όπως τονίστηκε, οι ιδιαιτερότητες του ερευνητικού προγράμματος που αποτέλεσαν την αφορμή για τη δημιουργία του ΣΝΕΛ οδήγησαν σε ένα σώμα κειμένων που περιλαμβάνει ποικίλων ειδών έργα από διάφορα κειμενικά υπο-είδη της λογοτεχνίας (μυθιστόρημα, διήγημα, ποίηση, θεατρικά έργα). Το ΣΝΕΛ μπορεί έτσι να αποτελέσει τη βάση για την ανάπτυξη μεγαλύτερων και περισσότερο ειδικευμένων σωμάτων κειμένων (π.χ. σώμα κειμένων μυθιστορημάτων ή ποίησης στα ελληνικά), αλλά και τη χρονική του επέκταση λ.χ. στις δύο πρώτες δεκαετίες του 20ού ή στον 19ο αιώνα, περιόδους για τις οποίες πολλά έργα δεν καλύπτονται από πνευματικά δικαιώματα. Εξίσου εφικτή είναι η επέκτασή του με έργα όπως μεταφράσεις λογοτεχνικών έργων για τις οποίες είναι πρόσφορη η ανάλυση με σώματα κειμένων (βλ. Dimitroulia & Goutsos, 2017) ή σε είδη στο μεταίχμιο της λογοτεχνίας όπως οι βιογραφίες και αυτοβιογραφίες. Τέλος, η δημιουργία γλωσσικών πόρων όπως το ΣΝΕΛ αναμένεται να οδηγήσει στην ανάπτυξη εργαλείων για την επεξεργασία και ανάλυσή τους. Σε αυτά περιλαμβάνονται ειδικά εργαλεία για μεγάλης κλίμακας μετατροπή πολυτονικών κειμένων σε μονοτονικό, για τον χαρακτηρισμό σε μέρη του λόγου (POS-tagging) ή για τη ενιαία λημματοποίηση ορθογραφικών παραλλαγών, διεπαφές αναζήτησης πληροφορίας στα ελληνικά, εφαρμογές και διδακτικά σενάρια για τη διδασκαλία της λογοτεχνίας μέσω σωμάτων κειμένων κ.ά.

Βιβλιογραφικές αναφορές

Atari, M., & Henrich, J. (2023). Historical Psychology. *Current Directions in Psychological Science*, *32*(2), 176-183.

Bamman, D., Underwood, T., & Smith, N. A. (2014). A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 370-379). Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-10

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2, 1-127.

Bengio, Y., Courville, A. & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35* (8): 1798-1828.

Boschetti, F., Romanello, M., Babeu, A., Bamman, D., Crane, G. (2009). Improving OCR Accuracy for Classical Critical Editions. In Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (Eds.), *Research and Advanced Technology for Digital Libraries* (pp. 156-167). ECDL 2009. Lecture Notes in Computer Science, vol 5714. Springer. https://doi.org/10.1007/978-3-642-04346-8_17.

Burnard, L., Schöch, C. & Odebrecht, C. (2021). In search of comity: TEI for distant reading. *Journal of the Text Encoding Initiative, 14*.

Busse, B. (2020). *Speech, Writing, and Thought Presentation in 19th-Century Narrative Fiction. A Corpus-Assisted Approach*. Oxford University Press.

Busse, U. (2002). *Linguistic Variation in the Shakespeare Corpus. Morpho- Syntactic Variability of Second Person Pronouns*. Benjamins.

Čermáková, A. & Mahlberg, M. (2022). The representation of mothers and the gendered social structure of nineteenth-century children's literature'. *English Text Construction, 14* (2), 119-149.

Dimitroulia, T. & Goutsos, D. (eds). (2017). *inTRAlinea*. *Special Issue: Corpora and Literary Translation.* http://www.intralinea.org/specials/corpora_literary_trans

Elson, D. K., Dames, N. & McKeown, K. R. (2010). Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 138-147). Association for Computational Linguistics.

Fischer, Frank, et al. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. *Proceedings of DH2019: "Complexities"*. https://doi.org/10.5281/zenodo.4284002

Fischer-Starcke, B. (2009). Keywords and frequent phrases of Jane Austen's *Pride and Prejudice*: A corpus-stylistic analysis. *International Journal of Corpus Linguistics, 14* (4), 492-523.

Fischer-Starcke, B. (2010). *Corpus Linguistics in Literary Analysis. Jane Austen and her Contemporaries*. Continuum.

Gatos, B., Louloudis, G. & N. Stamatopoulos (2011). Greek polytonic OCR based on efficient character class number reduction. In *2011 International Conference on Document Analysis and Recognition* (ICDAR '11) (pp. 1155-1159). IEEE Computer Society. https://doi.org/10.1109/ICDAR.2011.233.

Goutsos, D. (2010). The Corpus of Greek Texts: A reference corpus for Modern Greek. *Corpora, 5* (1), 29-44.

Goutsos, D. & Fragaki, G. (2009). Lexical choices of gender identity in Greek genres: The view from corpora. *Pragmatics, 19* (3), 317-340.

Goutsos, D., Fragaki, G., Florou, I., Kakousi, V. & Savvidou, P. (2017). The Diachronic Corpus of Greek of the 20th century: Design and compilation. In Georgakopoulos, T. et al. (Eds.), *Proceedings of the 12th International Conference on Greek Linguistics* (Vol. 1) (pp. 369-381). Edition Romiosini/CeMoG, Freie Universität Berlin.

Green, C. (2017). Introducing the Corpus of the Canon of Western Literature: A corpus for culturomics and stylistics. *Language and Literature, 26*(4), 282-299.

Jackson, J. C. & Atari, M. (2025). Historical psychology: How the events of yesterday shaped the minds of today. *Current Research in Ecological and Social Psychology*, 100247.

Jacobs, A. M. 2018. The Gutenberg English Poetry Corpus: Exemplary quantitative narrative analyses. *Frontiers in Digital Humanities, 5*, 5.

Katsouros, V., Papavassiliou, V., Simistira, F. & Gatos, B. (2016). Recognition of Greek polytonic on historical degraded texts using HMMs. In *12th IAPR Workshop on Document Analysis Systems (DAS)* (pp. 346-351). https://doi.org/10.1109/DAS.2016.60.

Kim, S., Tak, J.-Y., Kwak, E. J., Lim, T. Y. & Lee, S. H. 2020. Implications of vocabulary density for poetry: Reading T. S. Eliot's poetry through computational methods. *Digital Scholarship in the Humanities, 36* (2), 371-382.

Langlais, P.-C. (2021). *Fictions littéraires de Gallica / Literary Fictions of Gallica*. Version 1. Zenodo. https://doi.org/10.5281/zenodo.4751204.

Leblond, A. (2022). *Corpus Chapitres*. Version v1.0.0. https://doi.org/10.5281/ zenodo.7446728.

Mahlberg, M. (2013). *Corpus Stylistics and Dickens's Fiction*. Routledge.

March, E., Moran, D., Houlbrook, M., Jewkes, Y. & Mahlberg, M. (2023). Defining the carceral characteristics of the 'Dickensian prison': A corpus stylistics analysis of Dickens's novels. *Victoriographies, 13* (1), 15-41.

Martins, M. J. D. & Baumard, N. (2022). How to develop reliable instruments to measure the cultural evolution of preferences and feelings in history? *Frontiers in Psychology, 13*, 786229. https://doi.org/10.3389/fpsyg.2022.786229.

Mitchell, T. M. (1997). *Machine Learning.* McGraw-Hill.

Percillier, M. (2017). Creating and analyzing literary corpora. In Hai-Jew, S. (Ed.) *Data Analytics in Digital Humanities. Multimedia Systems and Applications* (pp. 91-118). Springer.

Perifanos, K. & Goutsos, D. (2025). Logios: An open source Greek Polytonic Optical Character Recognition system. https://doi.org/10.48550/arXiv.2506.21474

Robertson, B. & Boschetti, F. (2017). Large-scale optical character recognition of Ancient Greek. *Mouseion, 14*, 341-359.

Robertson, D. (2013). An integrated system for polytonic Greek OCR. Paper presented at the Digital Classicist Seminar, Institute of Classical Studies, London, UK, July 19, 2013.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, 85-117.

Semino E. & Short, M. (2004) *Corpus Stylistics. Speech, Writing and Thought Presentation in a Corpus of English Writing*. Routledge.

Sichani, A., Kaddas, P., Mikros, G. & Gatos, B. (2019). OCR for Greek polytonic (multi accent) historical printed documents: Development, optimization and quality control. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage* (pp. 9-13). DATeCH2019, Association for Computing Machinery.

Simistira, F., Ul-Hassan, A., Papavassiliou, V., Gatos, B., Katsouros, V. & Liwicki, M. (2015). Recognition of historical Greek polytonic scripts using LSTM networks. *13th International Conference on Document Analysis and Recognition (ICDAR*), pp. 766-770. https://doi.org/10.1109/ICDAR. 2015. 7333865.

Stamou, V., Malli, M., Takorou P., Xylogianni, A. & Markantonatou, S. (2020). Evaluation of Verb Multiword Expressions discovery measurements in literature corpora of Modern Greek. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (Eds.), *EURALEX Proceedings, Lexicography for inclusion* (Vol. 1) (pp. 295-301). Democritus University of Thrace.

Tzogka, C., Koidaki, F., Doropoulos, S., Papastergiou, I., Agrafiotis, E., Tiktopoulou, K. & Vologiannidis, S. (2021). OCR workflow: Facing printed texts of Ancient, Medieval and Modern Greek literature. In *Proceedings of the Conference on Digital Curation Technologies (Qurator 2021). Berlin, Germany, February 8th to 12th, 2021. CEUR Workshop Proceedings*. http://ceur-ws.org/Vol-2836/#qurator2021_paper_8.

Underwood, T. & Sellers, J. (2015). How quickly do literary standards change? https://tedunderwood.com/2015/05/18/how-quickly-do-literary-standards-change.

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Yang, D. Automatic theme and motif identification in large-scale English literary corpora using deep learning approaches. *Journal of Computational Methods in Sciences and Engineering, 25*(5), 4762-4773.

Γούτσος, Δ. (2025). *ότι* και *πως*: Στοιχεία για την κατανομή τους στη συγχρονία και τη διαχρονία της ελληνικής με βάση σώματα κειμένων. Στο D. Delli, G. Fragaki & F. Guérin (Eds.), *Actes du XLIVe Colloque international de linguistique fonctionnelle, Kalamata, Grèce, 17-20 octobre 2023* (pp.133-138). Éditions Modulaires Européennes (E.M.E.) & InterCommunications s.p.r.l.

Γούτσος, Δ. (υπό προετοιμασία). Διαχρονική αλλαγή στη γλώσσα των ποιημάτων του Γιάννη Ρίτσου. Στο Φραγκάκη, Γ. & Ν. Μαθιουδάκης (επιμ.). *Ψηφιακές εφαρμογές στη νεοελληνική λογοτεχνία*. Gutenberg.

Γούτσος, Δ. & Φραγκάκη, Γ. (2015). *Εισαγωγή στη γλωσσολογία σωμάτων κειμένων.* Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. http://hdl.handle.net/11419/1932.

Δημητρούλια Τ. (2022). Τα ηλεκτρονικά σώματα κειμένων και η ανάλυσή τους. Ερευνητικές προοπτικές για τη μελέτη της λογοτεχνίας. *Σύγκριση*, *31*, 160-184.

Δημητρούλια, Τ., Γούτσος, Δ. & Φραγκάκη, Γ. (2024). *Εισαγωγή στις Ψηφιακές Ανθρωπιστικές Επιστήμες: Ένας πρακτικός οδηγός* (2η έκδ., εμπλουτισμένη). Καρδαμίτσα.

Καστρινάκη, Α. (2015). *Η λογοτεχνία στη δεκαετία 1940-1950*. Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. https://repository.kallipos.gr/handle/11419/167.

Σταματάτος, Ν. (2011). Οπτική επεξεργασία και ανάλυση ιστορικών εγγράφων. Διδακτορική διατριβή, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών.

Φραγκάκη, Γ. (2024). Διδάσκοντας λογοτεχνία με σώματα κειμένων. *Μέντορας,* 22 (1), 89-117.

Φραγκάκη, Γ. (2025). Ανάλυση του λόγου της λογοτεχνίας μέσω της υφολογίας σωμάτων κειμένων: Θεματικές και διαχρονική αλλαγή. *Σύγκριση, 33*, 352-377.

Φραγκάκη, Γ. (υπό προετοιμασία). Γλωσσικά χαρακτηριστικά του λογοτεχνικού λόγου στα ελληνικά: Μια προκαταρκτική προσέγγιση με βάση σώματα κειμένων. Στο Φραγκάκη, Γ. & Ν. Μαθιουδάκης (επιμ.). *Ψηφιακές εφαρμογές στη νεοελληνική λογοτεχνία*. Gutenberg.

https://www.frantext.fr

https://www.theatre-classique.fr/index.html

https://www.ortolang.fr/market/corpora/malherbe/v3.7.5

https://www.chansondaspremont.eu/index.html -

https://hdl.handle.net/11403/digulleville/v1

https://gallica.bnf.fr/selections/fr/html/litterature-de-fiction-en-vers-et-en-prose

http://www.hathitrust.org

https://llds.ling-phil.ox.ac.uk/llds/xmlui

https://www.gutenberg.org

https://digitalhumanities.stanford.edu/projects

https://litlab.stanford.edu/projects

https://textual-optics-lab.uchicago.edu/english

https://emed.folger.edu/about

https://github.com/mahlberg-lab/corpora/blob/master/INDEX.pdf

https://www.english-corpora.org/coca

https://www.english-corpora.org/coha

https://www.scottishcorpus.ac.uk/corpus-details

https://albanian.web-corpora.net

http://tlio.ovi.cnr.it/TLIO/index2.html

http://www.rae.es

https://www.clarin.eu/resource-families/literary-corpora

https://logios.phil.uoa.gr

https://dracor.org/#

https://www.dropbox.com/scl/fi/lf7bk6npx8s7s6ehmipca/Corpus-of-the-Canon-of-Western-Literature-1.0.rar?rlkey=jemnkzbr23xf97hp96l0scv9c&e=1&dl=0

http://sek.edu.gr και http://greekcorpus20.phil.uoa.gr

https://anemi.lib.uoc.gr/ ?lang=el

http://www.potheg.gr/intro.aspx

https://www.greek-language.gr/digitalResources/literature/tools/concordance/index.html

https://inventory.clarin.gr/corpus/890?lang=el

https://inventory.clarin.gr/corpus/994

https://distantreading.github.io/ELTeC/gre/index.html

https://inventory.clarin.gr

https://github.com/kperi/Kalchas

Παράρτημα: Κατάλογος έργων του ΣΝΕΛ

Έτος	**Συγγραφέας**	**Τίτλος**	**Πολυτονικό/ Μονοτονικό**
1927	Καζαντζάκης Νίκος	Ασκητική	Μ
1927	Παπαντωνίου Ζαχαρίας	Διηγήματα	Π
1928	Νιρβάνας Παύλος	Το έγκλημα του Ψυχικού	Π
1928	(Συλλογικός τόμος)	Στρατιωτικά διηγήματα	Π
1929	Δούκας Στρατής	Ιστορία ενός αιχμαλώτου	Μ
1929	Ταγκόπουλος Πέτρος	Η ζωή που πέρασε	Π
1930	Ξενόπουλος Γρηγόριος	Η τρίμορφη γυναίκα	Π
1930	Πολίτης Κοσμάς	Λεμονόδασος	Μ
1931	Βενέζης Ηλίας	Το νούμερο 31328	Μ
1931	Βάρναλης Κώστας	Η αληθινή απολογία του Σωκράτη	Π
1932	Παπαντωνίου Ζαχαρίας	Ο όρκος του πεθαμένου	Π
1932	Δέλτα Πηνελόπη	Τρελαντώνης	Μ
1933	Καββαδίας Νίκος	Μαραμπού	Π
1933	Μυριβήλης Στρατής	Η δασκάλα με τα χρυσά μάτια	Μ
1934	Ρίτσος Γιάννης	Τρακτέρ	Μ
1934	Αλεξίου Έλλη	Γ΄ Χριστιανικόν Παρθεναγωγείον	Π
1935	Πεντζίκης Νίκος	Αντρέας Δημακούδης	Π
1935	Δέλτα Πηνελόπη	Μάγκας	Μ
1936	Ρίτσος Γιάννης	Επιτάφιος	Μ
1936	Καζαντζάκης Νίκος	Ο βραχόκηπος	Μ
1937	Δέλτα Πηνελόπη	Στα μυστικά του βάλτου	Μ
1937	Παπαντωνίου Ζαχαρίας	Η θυσία	Π
1938	Λουντέμης Μενέλαος	Τα πλοία δεν άραξαν	Π
1938	Θεοτοκάς Γιώργος	Το δαιμόνιο	Μ
1939	Βενέζης Ηλίας	Γαλήνη	Μ
1939	Αξιώτη Μέλπω	Δύσκολες νύχτες	Π
1940	Αξιώτη Μέλπω	Θέλετε να χορέψομε, Μαρία	Π
1940	Ψαθάς Δημήτρης	Μαντάμ Σουσού	Μ
1941	Πρεβελάκης Παντελής	Η πιο γυμνή ποίηση	Π
1941	Κορνάρος Θέμος	Καλοί και κακοί	Π
1942	Κόντογλου Φώτης	Φημισμένοι άντρες και λησμονημένοι	Π
1942	Θεοτοκάς Γιώργος	Αντάρα στ' Ανάπλι	Π
1943	Καραγάτσης Μ.	Το χαμένο νησί	Π
1943	Βενέζης Ηλίας	Αιολική γη	Μ
1944	Εγγονόπουλος Νίκος	Μπολιβάρ	Μ

1944	Καραγάτσης Μ.	Ο κοτζάμπασης του Καστρό-πυργου	Π
1945	Λουντέμης Μενέλαος	Ο Μεγάλος Δεκέμβρης	Π
1945	Ψαθάς Δημήτρης	Χειμώνας του '41	Π
1946	Χατζής Δημήτρης	Φωτιά	Π
1946	Αξιώτη Μέλπω	Εικοστός αιώνας	Π
1947	Καββαδίας Νίκος	Πούσι	Μ
1947	Βρεττάκος Νικηφόρος	Παραμυθένια πολιτεία	Π
1948	Σαχτούρης Μίλτος	Παραλογαίς	Μ
1948	Αμαριώτου Μαρία	Ιστορίες της μανούλας μου	Π
1949	Βρεττάκος Νικηφόρος	Το βιβλίο της Μαργαρίτας	Π
1949	Ζαλοκώστας Χρήστος	Το χρονικό της σκλαβιάς	Π
1950	Καζαντζάκης Νίκος	Καπετάν Μιχάλης	Μ
1950	Βρεττάκος Νικηφόρος	Τα θολά ποτάμια	Π
1951	Καζαντζάκης Νίκος	Ο Χριστός ξανασταυρώνεται	Μ
1951	Τερζάκης Άγγελος	Δίχως Θεό	Π
1952	Ρίτσος Γιάννης	Ο άνθρωπος με το γαρύφαλλο	Μ
1952	Λειβαδίτης Τάσος	Μάχη στην άκρη της νύχτας	Μ
1953	Μαρής Γιάννης	Έγκλημα στο Κολωνάκι	Μ
1953	Χατζής Δημήτρης	Το τέλος της μικρής μας πόλης	Π
1954	Καββαδίας Νίκος	Βάρδια	Π
1954	Καραγάτσης Μ.	Η μεγάλη χίμαιρα	Π
1955	Καζαντζάκης Νίκος	Ο τελευταίος πειρασμός	Μ
1955	Νάκου Λιλίκα	Η κυρία Ντορεμί	Π
1956	Λουντέμης Μενέλαος	Οι κερασιές θ' ανθίσουν και φέτος	Μ
1956	Βενέζης Ηλίας	Ωκεανός	Π
1957	Τσίρκας Στρατής	Νουρεντίν Μπόμπα	Π
1957	Βρεττάκος Νικηφόρος	Η μητέρα μου στην εκκλησία	Π
1958	Δημουλά Κική	Ερήμην	Π
1958	Λειβαδίτης Τάσος	Οι γυναίκες με τα αλογίσια μά-τια	Μ
1959	Πρεβελάκης Παντελής	Ο ήλιος του θανάτου	Μ
1959	Κάσδαγλης Νίκος	Κεκαρμένοι	Π
1960	Χειμωνάς Γιώργος	Πεισίστρατος	Π
1960	Καχτίτσης Νίκος	Η ομορφάσχημη	Π
1961	Τσίρκας Στρατής	Η λέσχη	Μ
1961	Σαμαράκης Αντώνης	Αρνούμαι	Π
1962	Ταχτσής Κώστας	Το τρίτο στεφάνι	Μ
1962	Σωτηρίου Διδώ	Ματωμένα χώματα	Μ
1963	Ζέη Άλκη	Το καπλάνι της βιτρίνας	Μ
1963	Καζαντζάκης Νίκος	Οι αδελφοφάδες	Μ
1964	Αθανασιάδης Νίκος	Το γυμνό κορίτσι	Π

1964	Χειμωνάς Γιώργος	Εκδρομή	Π
1965	Σαμαράκης Αντώνης	Το λάθος	Μ
1965	Αξιώτη Μέλπω	Το σπίτι μου	Π
1966	Σχινάς Αλέξανδρος	Αναφορά περιπτώσεων	Μ
1966	Χειμωνάς Γιώργος	Μυθιστόρημα	Π
1967	Καχτίτσης Νίκος	Ο ήρωας της Γάνδης	Π
1967	Τσιφόρος Νίκος	Ο Γκιούλιβερ στη χώρα των γιγάντων	Μ
1968	Βενέζης Ηλίας	Οι νικημένοι	Π
1968	Αβέρωφ-Τοσίτσας Γεώργιος	Γη δελφύς	Π
1969	Κοροβέσης Περικλής	Ανθρωποφύλακες	Μ
1969	Σαρή Ζωρζ	Ο θησαυρός της Βαγίας	Μ
1970	Ταχτσής Κώστας	Τα ρέστα	Μ
1970	Αναγνωστάκης Νίκος	Στόχος	Μ
1971	Τσιφόρος Νίκος	Η γυναίκα κουρσάρος	Μ
1971	Ζέη Άλκη	Ο μεγάλος περίπατος του Πέτρου	Μ
1972	Βαλτινός Θανάσης	Συναξάρι Αντρέα Κορδοπάτη	Μ
1972	Αξιώτη Μέλπω	Η Κάδμω	Π
1973	Ρίτσος Γιάννης	Γραγκάντα	Μ
1973	Καμπανέλλης Ιάκωβος	Το μεγάλο μας τσίρκο	Μ
1974	Αλεξάνδρου Άρης	Το κιβώτιο	Μ
1974	Σεφέρης Γιώργος	Έξι νύχτες στην Ακρόπολη	Μ
1975	Κουμανταρέας Μένης	Βιοτεχνία υαλικών	Μ
1975	Νικολαΐδης Αριστοτέλης	Εξαφάνιση	Π
1976	Σωτηρίου Διδώ	Εντολή	Μ
1976	Τσίρκας Στρατής	Η Χαμένη Άνοιξη	Μ
1977	Σουρούνης Αντώνης	Οι συμπαίκτες	Μ
1977	Λειβαδίτης Τάσος	Βιολί για μονόχειρα	Μ
1978	Ιορδανίδου Μαρία	Σαν τα τρελά πουλιά	Π
1978	Κουμανταρέας Μένης	Η κυρία Κούλα	Π
1979	Δούκα Μάρω	Η αρχαία σκουριά	Μ
1979	Ιορδανίδου Μαρία	Στου κύκλου τα γυρίσματα	Π
1980	Δρακονταειδής Φίλιππος	Προς Οφρύνιο	Π
1980	Λειβαδίτης Τάσος	Εγχειρίδιο ευθανασίας	Μ
1981	Ιορδανίδου Μαρία	Η αυλή μας	Π
1981	Κουμανταρέας Μένης	Σεραφείμ και Χερουβείμ	Μ
1982	Σωτηρίου Διδώ	Κατεδαφιζόμεθα	Μ
1982	Γώγου Κατερίνα	Το ξύλινο παλτό	Π
1983	Λειβαδίτης Τάσος	Ο τυφλός με τον λύχνο	Μ
1983	Βρεττάκος Νικηφόρος	Ο διακεκριμένος πλανήτης	Π

1984	Κοντολέων Μάνος	Με στοιχεία προσωπικών συνεντεύξεων	Μ
1984	Ζατέλλη Ζυράννα	Περσινή αρραβωνιαστικιά	Π
1985	Μίσσιος Χρόνης	…καλά εσύ σκοτώθηκες νωρίς	Μ
1985	Καραπάνου Μαργαρίτα	Ο υπνοβάτης	Μ
1986	Κουμανταρέας Μένης	Η φανέλα με το εννιά	Π
1986	Σίνου Κίρα	Στην χώρα των μαμούθ	Μ
1987	Αμπατζόγλου Πέτρος	Τι θέλει η κυρία Φρίμαν	Μ
1987	Ζέη Άλκη	Η αρραβωνιαστικιά του Αχιλλέα	Μ
1988	Δοξιάδης Απόστολος	Μακαβέττας	Μ
1988	Μίσσιος Χρόνης	Χαμογέλα ρε...Τι σου ζητάνε	Μ
1989	Ξανθούλης Γιάννης	Ο χάρτινος Σεπτέμβρης της καρδιάς μας	Μ
1989	Μουρσελάς Κώστας	Βαμμένα κόκκινα μαλλιά	Μ
1990	Μάτεσις Παύλος	Η μητέρα του σκύλου	Π
1990	Σχινάς Αλέξανδρος	Η παρτίδα	Μ
1991	Βακαλόπουλος Χρήστος	Η γραμμή του ορίζοντος	Μ
1991	Ελύτης Οδυσσέας	Τα ελεγεία της οξώπετρας	Μ
1992	Καλιότσος Παντελής	Τα γουρούνια	Μ
1992	Ζέη Άλκη	Ο ψεύτης παππούς	Μ
1993	Σίμος Γρηγόρης	Ο ζωντανός τόπος των αναμνήσεων	Μ
1993	Δεσύλλας Χρήστος	Η κρύπτη	Μ
1994	Κουμανταρέας Μένης	Θυμάμαι την Μαρία	Μ
1994	Μανιώτης Γιώργος	Το πονηρό μονοπάτι	Μ
1995	Μάρκαρης Πέτρος	Νυχτερινό δελτίο	Μ
1995	Μήτσου Ανδρέας	Τα ανίσχυρα ψεύδη του Ορέστη Χαλκιόπουλου	Μ
1996	Λαμπαρίδου-Πόθου Μαρία	Πήραν την πόλιν πήραν την...	Μ
1996	Πολυράκης Γιώργος	Εκείνη η στιγμή	Μ
1997	Τσεμπερλίδου Κατερίνα	Όχι πια σεξ, μόνο φίλοι	Μ
1997	Καρυστιάνη Ιωάννα	Μικρά Αγγλία	Μ
1998	Παυλιώτης Αργύρης	Αρμαγεδών	Μ
1998	Ξανθούλης Γιάννης	Ύστερα, ήρθαν οι μέλισσες	Μ
1999	Κορτώ Αύγουστος	Ραμπαστέν	Μ
1999	Σωτηροπούλου Έρση	Ζιγκ-ζαγκ στις νεραντζιές	Μ

Abstract

Dionysis Goutsos, Christiana Nika, Konstantinos Perifanos, Georgia Fragaki

CMGL: A new linguistic resource for studying Greek literature

This paper presents the principles and procedures involved in the creation of a new linguistic resource for Greek, the *Corpus of Modern Greek Literature* (CMGL), designed to support the systematic diachronic study of twentieth-century Greek literature. We first outline the conceptual framework underpinning the development of CMGL and situate it in relation to comparable resources in other languages, which are briefly reviewed. We then describe the corpus compilation process, with particular emphasis on the *Logios* platform, developed specifically for the digitization of polytonic texts (Perifanos & Goutsos 2025). CMGL comprises 146 works of modern Greek literature, written in either the polytonic or monotonic orthographic system and published between 1927 and 1999, amounting to approximately 6.5 million words. The corpus includes novels, short-story collections, poetry collections, and theatrical plays. The article concludes by presenting preliminary findings that illustrate the analytical potential of corpus-based stylistic approaches. Specifically, we report frequency lists derived from CMGL, including lexical bundles, as well as measures of lexical density, average sentence length, and readability for the texts in the corpus. In addition, we present charts depicting the diachronic development of selected grammatical and lexical forms, pointing to research directions that can be further explored.