ΣΩΜΑ ΕΠΟΠΤΕΙΑΣ ΝΕΟΛΟΓΙΣΜΩΝ ΤΗΣ ΝΕΑΣ ΕΛΛΗΝΙΚΗΣ: ΣΧΕΔΙΑΣΜΟΣ ΚΑΙ ΚΕΙΜΕΝΙΚΗ ΤΑΞΙΝΟΜΗΣΗ

Βασιλική Αφεντουλίδου & Αναστασία Χριστοφίδου
Ακαδημία Αθηνών

I am in support of human beings having human-like things to read
John Sinclair[1]

ABSTRACT

This paper introduces the corpus component of *Νεοδημία*, an ongoing research programme conducted at the *Research Centre for Scientific Terms and Neologisms* of the Academy of Athens, constantly developed to accomplish the tasks of (semi)automated detection and linguistic analysis of Greek Neologisms and Terminology. The *Monitor Corpus of Neologisms* (henceforth, MCN) is a dynamic collection of journalistic discourse harvested on a daily basis from the online versions of selected Greek newspapers. A detailed account of the objectives, design criteria, levels of analysis, as well as the compilation process of the corpus is provided. Firstly, the system of discovery and extraction of significant textual content (structured and enhanced with metadata), as incorporated into the general architecture of *Νεοδημία*, is briefly presented. The main part of the paper addresses issues of text encoding and markup theory, mainly focusing on the encoding decisions that had to be made concerning the annotation layers in the MCN, the representation of basic metadata and superordinate structural divisions of newspaper articles, and notably, the classification of texts. A document model for the corpus is defined as a customization within the framework of the TEI (*Text Encoding Initiative*). Following Χριστοφίδου (2012β), a dynamic text and topic classification scheme is proposed, which can be integrated into the textual model of the TEI and extend it.

Ευχαριστούμε τον καθηγητή κ. Ηρακλή Βαρλάμη (Χαροκόπειο Πανεπιστήμιο) για τη σημαντική συμβολή του στη συγκρότηση του ηλεκτρονικού σώματος κειμένων εφημερίδων στο ΚΕΕΟΝ.

1 Από την ομιλία του *From theory to practice*, όπως δημοσιεύθηκε στο Sinclair (1995: 108).

Examples of the encoded features are provided and discussed. The rest of the paper offers an implementation of the genre taxonomy designed for the MCN in terms of (semi)automated annotation practices. Lastly, a case study is briefly reported, in an effort to highlight the potential of the MCN in empirical, textually-informed analyses of the morphological tendencies of Modern Greek and pinpoint its significance to the study of neology.

1. Εισαγωγή

1.1 Μελέτη της νεολογίας μέσα από ηλεκτρονικά σώματα κειμένων

Η νεολογία αποτελεί έναν από τους πιο δυναμικούς τομείς του φαινομένου της γλώσσας. Η μελέτη της νεολογίας επιτρέπει στον ερευνητή να συναγάγει πολύτιμα συμπεράσματα για τη δυναμική και την παραγωγική πλευρά της μορφολογίας μέσω των ενεργών κανόνων σχηματισμού λέξεων, καθώς και για τη διεπιδραστική σχέση λέξης-σημασίας εντός του κοινωνικού γίγνεσθαι (μέσω κειμένων / λόγου). Περαιτέρω, η νεολογία όχι μόνο τέμνεται με κλάδους όπως της μορφολογίας / λεξικολογίας / ορολογίας, της σημασιολογίας και της κοινωνιογλωσσολογίας / κειμενογλωσσολογίας, αλλά αναδεικνύει ίσως και την πιο ενδιαφέρουσα πλευρά τους (βλ. και Kinne 1998: 73 κ.εξ.).

Όπως στην πλειονότητα των γλωσσολογικών ερευνών έτσι και στη μελέτη της νεολογίας έχει υπεισέλθει δυναμικά τις τελευταίες δεκαετίες η σωματοκειμενική προσέγγιση (*corpus-linguistic approach*), η οποία βασίζεται περισσότερο σε ποσοτικά προσανατολισμένες αναλύσεις χρησιμοποιώντας τις έννοιες της συχνότητας, του αυθεντικού περιβάλλοντος χρήσης, της καθιέρωσης, της παρακολούθησης (*establishment, entrenchment, monitoring*)[2] κ.λπ. Φυσικά αυτή η προσέγγιση δεν μπορεί παρά να λειτουργεί σε αγαστή αρμονία και αλληλεπίδραση με θεωρητικές υποθέσεις και με αντίστοιχες ποιοτικές αναλύσεις, αφού το να τίθενται σωστά τόσο τα ερευνητικά ερωτήματα, όσο και τα desiderata, αποτελεί ακόμη τη λυδία λίθο κάθε είδους έρευνας. Ωστόσο, βασική προϋπόθεση για την άρτια διεξαγωγή των ερευνών, ιδιαιτέρως σήμερα στην εποχή των «μεγαδεδομένων» (*big data*), είναι η ύπαρξη υπολογιστικών εργαλείων και μεθόδων που να εξασφαλίζουν αποτελεσματικότητα, διαφάνεια και εγκυρότητα στην επεξεργασία της γλωσσικής πληροφορίας. Ο σχεδιασμός, λοιπόν, ενός ηλεκτρονικού σώματος κειμένων (εφεξής ΗΣΚ)

2 Θεωρητικές έννοιες, όπως οι παραπάνω, μπορούν να ελεγχθούν πειραματικά και να διερευνηθούν εμπειρικά μέσα στα σώματα κειμένων (βλ. Stefanowitsch & Flach 2017, Gries 2017 και ειδικότερα σε σχέση με τη γλωσσική δημιουργικότητα και το φαινόμενο της νεολογίας βλ. Schmid 2008, Kerremans 2015).

scussed. The rest
onomy designed
practices. Lastly,
ght the potential
ses of the mor-
: its significance

ματα κειμένων

τομείς του φαινομε-
ερευνητή να συνα-
ταραγωγική πλευρά
ύ λέξεων, καθώς και
ινωνικού γίγνεσθαι
έμνεται με κλάδους
ιασιολογίας και της
αδεικνύει ίσως και
3 κ.εξ.).
έτσι και στη μελέτη
ιετίες η σωματοκει-
ισίζεται περισσότε-
ιώντας τις έννοιες
ης καθιέρωσης, της
ng)[2] κ.λπ. Φυσικά
αστή αρμονία και
χες ποιοτικές ανα-
ωτήματα, όσο και
έρευνας. Ωστόσο,
ύν, ιδιαιτέρως σή-
ύπαρξη υπολογι-
τελεσματικότητα,
κής πληροφορίας
ων (εφεξής ΗΣΚ)

ν πειραματικά και να
ι & Flach 2017, Gries
φαινόμενο της νεο-

εξακολουθεί να αποτελεί έργο εξαιρετικά πολυσύνθετο, απαιτώντας στιβαρό θεωρητικό υπόβαθρο, καθώς και ιδιαίτερη τεχνογνωσία και υπολογιστικούς πόρους, οι οποίοι πολλαπλασιάζονται εκθετικά, όταν η έρευνα στοχεύει στην αποτύπωση γλωσσικών φαινομένων εν εξελίξει.

Ειδικότερα η σχέση κειμένων και νεολογισμών – όπου το κείμενο αναδεικνύει, υποστηρίζει και ερμηνεύει τον νεολογισμό και ο νεολογισμός πυκνώνει, κάνει πιο συνοχικό το κείμενο μέσα από τη διευκόλυνση και πυροδότηση πολλαπλών κειμενικών λειτουργιών – έχει μια μακροχρόνια ευρωπαϊκή παράδοση (βλ. Dressler 1982, de Knop 1987, Boase-Beier 1987, Schröder 1978, 1983, Peschel 2002, για την Ελληνική Christofidou 1994, Χριστοφίδου 2001, 2008, Christofidou & Dimitropoulou in press). Αυτή η παράδοση έχει μπολιαστεί με τις νεότερες δυνατότητες των σωμάτων κειμένων (βλ. τον προβληματισμό και τα οφέλη που επισημαίνουν οι Teubert 1998, Sinclair 2005, Andersen & Hofland 2012, Kristiansen & Andersen 2012, Cabré & Nazar 2012, Renouf 2012), ώστε να παραγάγει πολλαπλασιαστικά αποτελέσματα. Μεταξύ άλλων αναφέρουμε τη μελέτη της σχέσης διαφόρων μορφολογικών τύπων νεολογισμού (π.χ. σύνθεση, παραγωγή, συμφυρμός κ.λπ.) και διαφορετικών θεματικών ή/και κειμενικών ειδών στα οποία αυτοί εμφανίζονται (βλ. Χριστοφίδου & Αφεντουλίδου 2016) ή τη διερεύνηση της δομικής θέσης του νεολογισμού, όπως στο σύστημα των τίτλων ή στα εναρκτικά τμήματα του εκάστοτε κειμένου εμφάνισης, λ.χ. στην εισαγωγική περίληψη (*lead*) δημοσιογραφικού άρθρου κ.λπ.

Δεδομένης της ουσιαστικής σημασίας που έχει για το φαινόμενο της νεολογίας η μελέτη των κειμένων, στα οποία αυτός εμφανίζεται, θα παρουσιάσουμε στην εργασία μας τη συγκρότηση ενός σώματος δημοσιογραφικών κειμένων βασισμένου σε παρακολούθηση / καταγραφή γλωσσικού υλικού από ελληνικές εφημερίδες, αυστηρώς για την ερευνητική σκοποθεσία του *Κέντρου Ερεύνης Επιστημονικών Όρων και Νεολογισμών* (ΚΕΕΟΝ) της Ακαδημίας Αθηνών. Συγκεκριμένα, θα αναλύσουμε (βλ. αντίστοιχες ενότητες) τις παραδοχές, τον τρόπο συλλογής των δεδομένων, την προτεινόμενη κωδικοποίησή τους, την οργάνωση της θεματικής και των κειμενικών ειδών στον δημοσιογραφικό λόγο και γενικότερα τον μακρο- και μικρο- σχεδιασμό του.

Πρώτα, όμως, θα υποστηρίξουμε την επιλογή μας για συγκρότηση σώματος κειμένων δημοσιογραφικού λόγου από το διαδίκτυο (μεθοδολογία *Web for corpus*, πρβλ. *Web as corpus*).

1.2 Web as/for Corpus

Η μεθοδολογία *Web as Corpus* (βλ. Lüdeling, Evert & Baroni 2007), δηλαδή η αντιμετώπιση του διαδικτύου ως διαρκώς ανανεώσιμου σώματος κειμένων ψηφιακού λόγου τεραστίων διαστάσεων, ελεύθερα προσβάσιμου και υπο-

λογιστικά αναγνώσιμου μέσω των μηχανών αναζήτησης (δηλαδή των αποτελεσμάτων σε αναζητήσεις μέσω *Google*, *Bing* κ.ά.) έχει πλεονεκτήματα για την έρευνα των νεολογικών σχηματισμών και γενικά της δημιουργικότητας στη γλώσσα, αλλά και σοβαρά μειονεκτήματα που σχετίζονται κυρίως με την αντιπροσωπευτικότητα και την ποιότητα του γλωσσικού του περιεχομένου (για επεξήγηση της μεθόδου και βιβλιογραφική επισκόπηση βλ. την πρώτη αναλυτική παρουσίαση της *Νεοδημίας*[3] – του ηλεκτρονικού περιβάλλοντος εντοπισμού, καταγραφής και επεξεργασίας νεολογισμών και ορολογίας του ΚΕΕΟΝ – στο Χριστοφίδου, Αφεντουλίδου, Καρασίμος & Δημητροπούλου 2013: 205 κ.εξ.). Η παρούσα μελέτη, επεκτείνοντας την προηγούμενη, εστιάζεται στην έρευνα για νεολογισμούς υπό το πρίσμα της συμπληρωματικής προσέγγισης – και βέλτιστης πρακτικής πλέον για τα διεθνή ερευνητικά προγράμματα νεολογισμών – *Web for Corpus* (βλ. Fletcher 2013: 5-6, Χριστοφίδου κ.ά. 2013: 207), δηλαδή την αξιοποίηση *στοχευμένων πηγών* του διαδικτύου – στην περίπτωσή μας δημοσιογραφικού λόγου από τις ηλεκτρονικές εκδόσεις ελληνικών εφημερίδων – για τη σύσταση ΗΣΚ.[4]

Ο σκοπός της δημιουργίας του σώματος κειμένων ως ειδικού υποσυστήματος στο οικοσύστημα της *Νεοδημίας* είναι τριπλός: Η πρόσβαση σε κειμενικό υλικό είναι απαραίτητη α) για την *ανίχνευση* νεολογισμών (εντοπισμός, συλλογή), β) για τη *μελέτη* τους μέσα σε εκτενή κειμενικά συμφραζόμενα (χαρακτηρισμός, ένταξη σε βάση δεδομένων) και γ) για την απρόσκοπτη παρακολούθηση των νεολογικών σχηματισμών (ελληνογενών και ξενόγλωσσων) σε δείγματα λόγου που ανανεώνονται στο πέρασμα του χρόνου (πρόκειται δηλαδή για διαχρονικού τύπου ανάλυση μέσα σε συγχρονικό πλαίσιο).[5] Το *Σώμα Εποπτείας Νεολογισμών* (εφεξής ΣΕΝ) ανήκει στις σωματοκειμενικές προσεγγίσεις στο φαινόμενο της νεολογίας και εντάσσεται στην κατηγορία των μεγάλων, ανοιχτών (μη πεπερασμένων) ΗΣΚ συνεχόμενης, μη διακοπτόμενης παρακολούθησης (*monitor corpora*).[6]

---

3 Συνεργάτες του ΚΕΕΟΝ που έχουν κατά καιρούς συμβάλει σε άλλα τμήματα της *Νεοδημίας* με αλφαβητική σειρά: Ρ. Βασιλειάδου, Ε. Δημητροπούλου, Α. Καρασίμος και φοιτητές της πρακτικής άσκησης.

4 Για την επιλογή αποκλειστικά της έντυπης δημοσιογραφίας ως θεσμοθετημένου, αντιπροσωπευτικού δείγματος γλωσσικής δραστηριότητας των Νεοελλήνων από το οποίο αντλούνται τα δεδομένα ανάλυσης βλ. Χριστοφίδου (2012α: 9-22).

5 Η στοχοθεσία επεκτείνεται αυτούσια και στην περιοχή της ορολογίας, η οποία αποτελεί τη δεύτερη βασική ερευνητική συνιστώσα του ΚΕΕΟΝ, ωστόσο, είναι προφανές ότι για τη μελέτη της ορολογίας και την εξαγωγή όρων είναι αναγκαία η επέκταση σε κείμενα που παράγονται σε άλλες περιοχές δραστηριοτήτων, πέρα από τον Τύπο, όπως η Επιστήμη ή η Διοίκηση.

6 Βλ. Μικρός (2004: 142), Γούτσος & Φραγκάκη (2015: 34). Πέρα από τον εξειδικευμένο χαρακτήρα του (νεολογικοί σχηματισμοί, δημοσιογραφικός λόγος) πρόκειται για το

1.3 Επισκόπηση συναφών έργων

Για λόγους περιορισμού σελίδων στην παρούσα εργασία θα αναφερθούμε πολύ συνοπτικά σε μία λεπτή διάκριση ανάμεσα σε υπάρχουσες υποδομές για την έρευνα των νεολογισμών με χρήση σωμάτων κειμένων. Κάποια ερευνητικά προγράμματα επιλέγουν έναν πιο στατικά κειμενοκεντρικό προσανατολισμό[7] (όπου η ταξινομημένη κειμενική πληροφορία χρησιμοποιείται ως τεκμήριο-παράθεμα που καταγράφει και επεξηγεί τον νεολογικό σχηματισμό και τα περιβάλλοντα εμφάνισής του), ενώ άλλα επεκτείνουν το ενδιαφέρον τους και σε μια πιο δυναμική παρακολούθηση των νεολογισμών στον άξονα του χρόνου ή βάσει άλλων μεταβλητών (ανοιχτά ΗΣΚ ελέγχου)[8] με εξαγωγή τόσο ποιοτικών όσο και ποσοτικών συμπερασμάτων για την εξέλιξή τους. Η δεύτερη επιλογή, βέβαια, απαιτεί συγκεκριμένο σχεδιασμό, συγκρότηση και αξιοποίηση των σωμάτων κειμένων. Η *Νεοδημία*, το ηλεκτρονικό περιβάλλον εντοπισμού, καταγραφής και παρακολούθησης νεολογισμών του ΚΕΕΟΝ (βλ. Χριστοφίδου κ.ά. 2013, Χριστοφίδου, Καρασίμος & Αφεντουλίδου 2014), με πρώτη για την Ελλάδα μηχανή ημι-αυτόματης ανίχνευσης, σχεδιάζεται και επεκτείνεται ώστε να συνδεθεί με τα μεγάλα ερευνητικά προγράμματα δυναμικής παρακολούθησης νεολογισμών, με την επιπρόσθετη μεθοδολογική διαφοροποίηση ότι εφαρμόζει και τη μεθοδολογία *Web as Corpus* (βλ. προηγούμενη ενότητα).

2. Σχεδιασμός

2.1 Δεδομένα

Στο ΣΕΝ της *Νεοδημίας* συλλέγονται συνεχόμενα οι ροές ειδήσεων (βλ.

πρώτο ελληνικό δυναμικό ΗΣΚ εποπτείας. Ακολουθώντας μάλιστα την πρόσφατη κατάταξη του Davies (2015: 12, 18, 22), το αρχείο που αναπτύσσεται ως υποσύστημα της *Νεοδημίας* εντάσσεται στα *υβριδικά* ΗΣΚ, δηλαδή στα μεγάλα ΗΣΚ που αντλούν δεδομένα από το διαδίκτυο αλλά αποκτούν προστιθέμενη αξία για τους χρήστες μόνο μέσω της (χρονοβόρας) μετα-επεξεργασίας τους και της προβολής τους από ισχυρά συστήματα με προσχεδιασμένη αρχιτεκτονική και εξειδικευμένες διεπαφές που να επιτρέπουν τη βέλτιστη διαχείριση της πληροφορίας. Πρόκειται, δηλαδή, για απαιτητικό έργο υποδομής που προϋποθέτει ειδικό σχεδιασμό, πειραματισμό και τεχνογνωσία σε κάθε βήμα υλοποίησης, αποτελώντας επέκταση του *Σώματος Αθησαύριστων Νεολογισμών* (ΣΑΝ) του ΚΕΕΟΝ (περιλαμβάνει τα κείμενα των παραθεμάτων των *Δελτίων Επιστημονικής Ορολογίας και Νεολογισμών* 9-10, 11 και 13).

7 Βλ. για τη Γερμανική *Die Wortwarte* (http://www.wortwarte.de), για την Ελληνική βλ. Αναστασιάδη-Συμεωνίδη, Αλεξιάδου & Νικολάου (2009).

8 Βλ. Renouf (1993 και 2012): πρόγραμμα APRIL, Cabré & Nazar (2012): OBNEO, Andersen & Hofland (2012) και Meurer (2012): *Norwegian Newspaper Corpus*, Kerremans (2015): *NeoCrawler*, Gérard (2014): *Logoscope*, Cartier (2017): *Neoveille* (εν εξελίξει, όπως και η *Νεοδημία*). Οι δύο τελευταίες μελέτες αναφέρονται σε προγράμματα με αξιοσημείωτες ομοιότητες με τη στοχοθεσία και τη μεθοδολογία ανάπτυξης της *Νεοδημίας*.

ενότητα 2.4) πέντε διαδικτυακών εκδόσεων ελληνικών εφημερίδων (*Έθνος, Η Καθημερινή, Πρώτο Θέμα, Το Βήμα, Τα Νέα*). Ως προς την επιλογή τους ακολουθήθηκε το πραγματολογικό κριτήριο της δημοφιλίας μετρήσιμο διά της αναγνωσιμότητας. Συγκεκριμένα, οι δικτυότοποι που επιλέχθηκαν έπρεπε να πληρούν τις ακόλουθες προϋποθέσεις:

(α) να αντιστοιχούν σε εφημερίδες ευρείας κυκλοφορίας με ταυτόχρονη έντυπη παρουσία στην ελληνική πραγματικότητα (μεγιστοποίηση της εμβέλειας του αναγνωστικού κοινού),

(β) να κατατάσσονται σε υψηλές θέσεις διαδικτυακής επισκεψιμότητας, σύμφωνα με τα στοιχεία (*website traffic statistics*) που συλλέγει η μηχανή *Alexa* για ελληνικούς δικτυότοπους.[9]

Καθίσταται, λοιπόν, σαφές ότι το αρχείο κειμένων της *Νεοδημίας* είναι εξαρχής ψηφιακά συγκροτημένο (*born digital*) και ως προς την αντιπροσωπευτικότητά του περιλαμβάνει όλον τον πληθυσμό των ειδήσεων προεπιλεγμένων πηγών που διακινούνται προς το αναγνωστικό κοινό με συγκεκριμένη τεχνολογία (βλ. ενότητα 2.4), συνεπώς ανήκει, χάρη στο εύρος και τη διαρκή ανανέωση των δεδομένων του, στα ονομαζόμενα «ΗΣΚ 10^{10}».[10] Ωστόσο, λόγω των ιδιαίτερων συνθηκών και του στόχου δημιουργίας του, καθώς κινούμαστε πλέον σε υπέρογκα μεγέθη[11] μαζικών δεδομένων ευνόητο είναι ότι βασικό μέλημα των δημιουργών του είναι τόσο η κωδικοποίηση και ο πολυεπίπεδος χαρακτηρισμός του, ώστε να είναι μηχαναγνώσιμη η πληροφορία που θέλουμε να πάρουμε από αυτό, όσο και η ενσωμάτωσή του στο συνολικό οικοσύστημα

9 Η αναγνωσιμότητα των ειδήσεων υπολογίζεται μέσω του εκάστοτε χρήστη (μονάδα δειγματοληψίας) που έχει επιλέξει εθελοντικά την εγκατάσταση του λογισμικού μέτρησης στον υπολογιστή του (http://www.avangate.com/avangate-resources/article/alexa-ranking.htm). Η *Alexa* παρέχει πληροφορίες σχετικά με τη γεωγραφική κατανομή των υπολογιστών που επισκέπτονται έναν συγκεκριμένο δικτυακό τόπο δίνοντας ποσοστά επισκεψιμότητας ανά χώρα και δικτυότοπο (http://www.alexa.com/topsites/countries/GR).

10 *Ten-Ten corpora* (δηλαδή αρχεία κειμένου που συλλέγονται από το διαδίκτυο με στόχο τις 10+ δισεκατομμύρια λέξεις). Για τον όρο 10^{10} και σχετικά ΗΣΚ βλ. την πλατφόρμα διαχείρισης / επεξεργασίας ΗΣΚ *Sketch Engine* (στο https://www.sketchengine.co.uk/documentation/tenten-corpora/, καθώς και Jakubíček, Kilgarriff, Kovář, Rychlý & Suchomel 2013), το *Jozef Stefan Institute Newsfeed Web Corpus* (στο http://newsfeed.ijs.si, με χρονοσήμανση των δεδομένων, καθώς και Trampus & Novak 2012), την πρωτοβουλία *COW* (*COrpora from the Web*, στο https://www.webcorpora.org), το *NOW* (*News on the Web*, στο corpus.byu.edu/now/, καθώς και Davies 2010).

11 Ενδεικτικά, ο όγκος των δεδομένων του αρχείου μας που συλλεγόταν αρχικά σε πιλοτικό επίπεδο για μία περίοδο δώδεκα μηνών (έτη 2013-2014) στο *Χαροκόπειο Πανεπιστήμιο* από τον καθηγητή Ηρακλή Βαρλάμη ανέρχεται σε 260 076 άρθρα εφημερίδων συνόλου 85 860 736 λέξεων (81 894 288 αλφαβητικές λέξεις, αν αφαιρεθούν τα αριθμητικά σύμβολα). Η μέτρηση έγινε με το λογισμικό *WordSmith Tools* v. 7.0 (Scott 2017).

ν εφημερίδων (*Έθνος*, *Η*
ς την επιλογή τους ακο-
λίας μετρήσιμο διά της
επιλέχθηκαν έπρεπε να

ιφορίας με ταυτόχρονη
ητα (μεγιστοποίηση της

ιακής επισκεψιμότητας,
ς) που συλλέγει η μηχα-

ων της *Νεοδημίας* είναι
ς προς την αντιπροσω-
ων ειδήσεων προεπιλεγ-
κοινό με συγκεκριμένη
το εύρος και τη διαρκή
$< 10^{10}$».[10] Ωστόσο, λόγω
του, καθώς κινούμαστε
ητο είναι ότι βασικό μέ-
και ο πολυεπίπεδος χα-
ηροφορία που θέλουμε
συνολικό οικοσύστημα

εκάστοτε χρήστη (μονάδα
η του λογισμικού μέτρησης
rces/article/alexa-ranking.
ατανομή των υπολογιστών
ποσοστά επισκεψιμότητας
ries/GR).

ονται από το διαδίκτυο με
ά ΗΣΚ βλ. την πλατφόρμα
/www.sketchengine.co.uk/
garriff, Kovář, Rychlý &
(στο http://newsfeed.ijs.si,
k 2012), την πρωτοβουλία
g), το *NOW* (*News on the*

ου συλλεγόταν αρχικά σε
14) στο *Χαροκόπειο Πανε-*
50 076 άρθρα εφημερίδων
αφαιρεθούν τα αριθμητικά
7.0 (Scott 2017).

της *Νεοδημίας* που περιλαμβάνει βάση δεδομένων σχεσιακού τύπου για την καταγραφή των νεολογικών σχηματισμών.

Τι δεν είναι το ΣΕΝ της *Νεοδημίας*; Ασφαλώς και δεν αποτελεί ΗΣΚ αναφοράς (*reference corpus*) της Νέας Ελληνικής, εφόσον ελέγχονται νεολογισμοί συγκεκριμένων πηγών μιας ειδικής γλωσσικής ποικιλίας. Ούτε πρόκειται για ψηφιακή έκδοση του κειμένου των εφημερίδων με τον τρόπο που εκδίδεται μία φιλολογική έκδοση δοκιμίων ή ένα ανθολόγιο κειμένων (βλ. ενότητα 3.2.2). Η ανάγνωση δεν είναι ο πρωταρχικός στόχος, ούτε η εκδοτική αποκατάσταση. Ο στόχος είναι η δομημένη πρόσβαση στο υλικό αποκλειστικά για ακαδημαϊκούς, ερευνητικούς σκοπούς, συνεπώς η κωδικοποίηση της πληροφορίας συνδέεται με την έρευνα και την καθιέρωση των νεολογισμών (ελληνογενών και ξενόγλωσσων). Για τον παραπάνω λόγο, όπως το επιβάλλουν άλλωστε και οι κανόνες της πνευματικής ιδιοκτησίας,[12] τα κείμενα που αποθηκεύονται απαραιτήτως παραπέμπουν στο «ψηφιακό είδωλο» αφετηρίας τους, δηλαδή στον υπερσύνδεσμο στο εκάστοτε άρθρο της διαδικτυακής εφημερίδας, ώστε ανά πάσα στιγμή να είναι δυνατή είτε η πρόσβαση στο περιεχόμενό του είτε οποιαδήποτε άλλη τροποποιητική ενέργεια δικαιούται να ζητήσει η πηγή προέλευσης. Όπως επισημαίνει ο Πολίτης (2008α: 433-434) δίνοντας παραδείγματα από την ελληνική πραγματικότητα, η συγκρότηση σωμάτων δημοσιογραφικού λόγου είναι πολυσύνθετο, κοπιαστικό εγχείρημα και προσκρούει σε πλείστα επιστημολογικά ζητήματα.

2.2 Παραδοχές σχεδιασμού

Πέρα από την αναγκαιότητα συγκρότησης μίας δεξαμενής κειμενικού περιεχομένου που να τροφοδοτεί διαρκώς την έρευνα, ώστε να μπορούν να εντοπιστούν οι νεολογικές χρήσεις, η άρτια ταξινόμηση των δεδομένων σε ψηφιακό περιβάλλον είναι η μόνη συνθήκη που εξασφαλίζει τη βέλτιστη λειτουργία των υπολογιστικών εργαλείων κειμενικής ανάλυσης. Τα δεδομένα, δηλαδή, εκπληρώνουν τον στόχο συλλογής και αποθήκευσής τους από τη στιγμή που μπορούν να χρησιμοποιηθούν οργανωμένα από πολλούς αποδέκτες, είτε ανθρώπους είτε αυτόματα συστήματα επεξεργασίας της γλώσσας, ώστε να δίνουν απαντήσεις σε στοχευμένα ερωτήματα. Επιπλέον, η αξιοποίησή τους σε αναλύσεις γίνεται

12 Το ΗΣΚ είναι προσβάσιμο μόνο εντός του εσωτερικού δικτύου του ΚΕΕΟΝ για τις ανάγκες των ερευνητικών του προγραμμάτων (βλ. οδηγίες του Ευρωπαϊκού Κοινοβουλίου 2001/29 της 22ας Μαΐου 2001 και 96/9/ΕΟΚ της 11ης Μαρτίου 1996, καθώς και τον Ν. 2121/1993 άρθρο 45Α παρ. 6, τον Ν. 2819/2000 άρθρο 7 και κανόνες Διεθνούς Δικαίου που ισχύουν στην Ελλάδα). Το ενδιαφέρον εστιάζεται μόνο στη γλωσσική πληροφορία για τους σκοπούς της γλωσσικής ανάλυσης και δεν επιδιώκεται η απευθείας πρόσβαση στο πρωτότυπο υλικό των εφημερίδων που περιέχει κώδικα HTML ή άλλα σημειωτικά συστήματα.

μέσω μίας ιδιότυπης πράξης «ανάγνωσης» του ψηφιακού κειμένου: η απλή φυλλομέτρηση για τυχαία ανακάλυψη δίνει τη θέση της στην πλοήγηση βάσει προκαθορισμένων παραμέτρων αναζήτησης. Στα συστήματα διαχείρισης σωμάτων κειμένων (*Corpus Management Systems*), η πρόσβαση στο υλικό γίνεται μέσα από τη μελέτη συμφραστικών πινάκων εμφανίσεων και συνεμφανίσεων γλωσσικών πραγματώσεων (όπου ο μελετητής ρυθμίζει το εύρος του κειμένου που μπορεί να δει για κάθε αποτέλεσμα), καθώς και την προσπέλαση ευρετηρίων (όπου οι λεκτικές ακολουθίες συνυπάρχουν με μετρήσεις, δηλαδή απλά αριθμητικά δεδομένα συχνότητας ή περισσότερο σύνθετα, με χρήση προηγμένων στατιστικών μεθόδων). Οι γραφηματικές οπτικοποιήσεις σπανίως κομίζουν κάτι καινούργιο στην ανάλυση, αναπαριστούν, όμως, την πληροφορία με τρόπο ώστε να γίνεται πιο εύληπτη.

Τόσο η ιδιότυπη αυτή αναγνωστική συνθήκη ενός σώματος κειμένων, σε αντιδιαστολή με τον ολιστικό, λιγότερο αποσπασματικό τρόπο ανάγνωσης μίας ψηφιακής έκδοσης κειμένων (*digital scholarly edition*) όσο και η αναγκαιότητα βέλτιστης διαχείρισης της πληροφορίας σε όλο τον κύκλο επεξεργασίας των δεδομένων (συλλογή, αποθήκευση, ανάλυση) δημιουργεί για τον σχεδιαστή του ειδικά καθήκοντα:

1) Τα κείμενα, όπως οργανώνονται σε σώμα, αποπλαισιώνονται από το περιβάλλον δημιουργίας και προβολής τους, οπότε τίθεται θέμα *αναπαράστασής* τους μέσα από τη διαμεσολάβηση της *κωδικοποίησης*. Η κωδικοποίηση δεν είναι απλή μεταγραφή του κειμένου, αλλά αναδεικνύει μέσα από κατάλληλο χαρακτηρισμό (*markup*) στοιχεία δομής και περιεχομένου, περικειμενικά στοιχεία, αλλά και πιο σύνθετες «αναγνώσεις» και επισημειώσεις (*annotations*) ανάλογα με τα επίπεδα αναπαράστασης της πληροφορίας που ο ερευνητής επιλέγει να επεξεργαστούν οι μηχανές.

2) Το ψηφιακό σώμα πρέπει να οργανωθεί με απόλυτη *αυστηρότητα* ως προς τα ερευνητικά ζητούμενα, ώστε να δίνει τη δυνατότητα στους χρήστες να υποβάλλουν τα κατάλληλα ερωτήματα μέσω των μηχανών.[13]

3) Αναπόφευκτα, οι σχεδιαστικές αρχές εκκινούνται από τη θεωρητική αφετηρία ανάλυσης (ο σχεδιασμός ενός α-θεωρητικού συστήματος επεξεργασίας φυσικής γλώσσας είναι όπως αποδεικνύεται στην πράξη ουτοπικός στόχος), ωστόσο πρέπει να εξασφαλίζουν αρκετή *ευελιξία*, ώστε η επεξεργασία του όγκου των δεδομένων από υπολογιστικά συστήματα διαχείρισης της πληροφορίας να καλύπτει όσο το δυνατόν περισσότερες αναγνωστικές διαδρομές, καθώς και μελλοντικές λειτουργίες που τα ερευνητικά ερωτήματα δεν πρόβλεψαν εξαρχής.

13 Μείζονος σημασίας στα μεγάλα ΗΣΚ είναι η αρχιτεκτονική τους και η διεπαφή με τους χρήστες (Davies 2015: 18).

4) Στον βαθμό όπου υπεισέρχεται ο ανθρώπινος παράγοντας και η επέμβαση του μελετητή στην επεξεργασία του περιεχομένου των κειμένων, κυρίως όταν πρόκειται για συλλογικά έργα (το περιβάλλον προβολής και ανάγνωσης αποτελεί δηλαδή συγχρόνως περιβάλλον συγγραφής, συνεργασίας και εμπλουτισμού), απαιτούνται κανόνες και *πρότυπα* που να καθοδηγούν και να δεσμεύουν τις επιλογές επέμβασης στα κείμενα, ώστε να μειώνεται το ποσοστό ανθρώπινου σφάλματος και το κόστος επεξεργασίας.

5) Οι οποιεσδήποτε αποφάσεις για τις τεχνολογίες που επιλέγονται στην υλοποίηση ενός ηλεκτρονικού σώματος οφείλουν να συνυπολογίζουν τη *συμβατότητα* και τη *διαλειτουργικότητά* τους με άλλα συστήματα / υπολογιστικά εργαλεία που απαρτίζουν το περιβάλλον εργασίας και να είναι όσο το δυνατόν απαγκιστρωμένες από συγκεκριμένα λειτουργικά συστήματα και κλειστές διανομές που μπορεί να εγκαταλειφθούν στο μέλλον (*platform-independent, open software solutions*).

Οι σχεδιαστικές παραδοχές (αναπαράσταση μέσω κωδικοποίησης, αυστηρότητα, ευελιξία, προτυποποίηση, διαλειτουργικότητα)[14] έγιναν σεβαστές στη βασική απόφαση να χρησιμοποιηθεί η γλώσσα σήμανσης XML ως μορφότυπο για την πρόσκτηση αλλά και την αναπαράσταση των κειμενικών δεδομένων της *Νεοδημίας*, καθώς και το πρότυπο TEI (*Text Encoding Initiative*) για την κωδικοποίησή τους και την οργάνωσή τους σε σώμα.

2.3 XML – eXtensible Markup Language

Απαντώντας στα αιτούμενα σχεδιασμού, το ΣΕΝ αποτελεί εφαρμογή τεχνολογίας XML, της καθιερωμένης μεταγλώσσας σήμανσης για την ανταλλαγή πληροφοριών και τη δυναμική προβολή των εγγράφων. Θεωρείται πλέον πάγια τακτική στη σωματοκειμενική έρευνα και έχουν αναπτυχθεί βέλτιστες πρακτικές για την αξιοποίησή της σε έργα κάθε επιπέδου πολυπλοκότητας (Hardie 2014, Gries & Berez 2017, Kübler & Zinsmeister 2015, Rühlemann, Bagoutdinov & O' Donnell 2015, Rühlemann & O' Donnell 2012).

Τα πλεονεκτήματα της XML είναι πολλά: 1)Διαβάζεται από μηχανές (τελικός αποδέκτης), αλλά με τρόπο που ο άνθρωπος, διαχειριστής της μηχανής, μπορεί να κατανοήσει (τα ονόματα των ετικετών – των σημειωτικών ψηφίδων της XML μαζί με τα ορίσματά τους – εκφράζονται σε φυσική γλώσσα ενσωματώνοντας μέσα στα έγγραφα πληροφορίες σχετικά με την ερμηνεία και το νόημα των κει-

14 Ο όγκος της βιβλιογραφίας πάνω στο θέμα του σχεδιασμού είναι τεράστιος και απασχολεί κάθε ερευνητικό εγχείρημα, μεγάλο ή μικρό. Βλ. McEnery & Rayson (1997) ή ενδεικτικά το άρθρο του Leech (1997), από τους πρώτους που καθιέρωσε μία σειρά από «χρυσούς κανόνες» για την επισημείωση των σωμάτων κειμένων.

μένων). 2) Εξασφαλίζει τη βέλτιστη διαχείριση της πληροφορίας, δηλαδή την ταξινόμηση, ευρετηρίαση και ιεράρχηση της μάζας του κειμενικού περιεχομένου, όπως και την αναζήτηση σε πεδία που έχουν χαρακτηριστεί καταλλήλως και με απόλυτη ευελιξία (εξ ου και *eXtensible* δηλαδή *επεκτάσιμη*) σε πολλαπλά επίπεδα (γλώσσες XPath, XQuery). 3) Επιτρέπει αφενός την παραδοσιακή φυλλομέτρηση του υλικού, αφετέρου τη δημιουργία πολλαπλών δημοσιεύσεών του (και σε άλλα μέσα) με μετασχηματισμούς (γλώσσα XSLT). 4) Δέχεται πρότυπους κανόνες (σχήματα) που ορίζουν εξαρχής το λεξιλόγιο και τη γραμματική των εγγράφων (τι επιτρέπεται, με ποια σειρά, πόσες φορές) και πιστοποιούν την εγκυρότητά τους (πέρα από τον έλεγχο της ορθής μορφοποίησης από αναλυτές XML που εντοπίζουν συντακτικά λάθη). 5α) Πρόκειται για ενιαίο περιβάλλον συγγραφής και εμπλουτισμού κειμένων, ανεξάρτητο από συγκεκριμένα λειτουργικά συστήματα (σημαντικό, διότι πέρα από την εξοικονόμηση πόρων ένα δυναμικό ΗΣΚ εποπτείας αποτελεί συλλογικό έργο εν εξελίξει). 5β) «Συνεργάζεται» ικανοποιητικά με άλλα συστήματα, γλώσσες προγραμματισμού, διαθέσιμα υπολογιστικά εργαλεία επεξεργασίας κειμένων,[15] καθώς και συστήματα βάσεων δεδομένων (π.χ. η βάση δεδομένων του ΚΕΕΟΝ έχει αναπτυχθεί σε τεχνολογία PHP-MySQL, το σύστημα συλλογής κειμένων και ανίχνευσης σε Java).

2.4 Συλλογή των κειμένων

Για την τροφοδοσία της *Νεοδημίας* με γλωσσικό υλικό χρησιμοποιείται πρόγραμμα ανίχνευσης περιεχομένου και εξαγωγής πληροφοριών απευθείας συμβατό με την τεχνολογία υλοποίησης σε XML.[16] Η αναλυτική παρουσίαση του υποσυστήματος συλλογής, όπως προσαρμόστηκε στην έρευνα του ΚΕΕΟΝ (πε-

15 Μειονέκτημα, ωστόσο, αποτελεί το γεγονός ότι η συντριπτική πλειοψηφία τους δεν έχει δημιουργηθεί για να επεξεργάζεται εξαρχής και να εμπλουτίζει 1) αρχεία XML 2) μεγάλο όγκο δεδομένων, οπότε απαιτούνται χρονοβόρα βήματα μετα-επεξεργασίας των μεγάλων ΗΣΚ (πρόκειται για γενικότερο ζήτημα που αφορά κάθε είδους επισημείωση των ΗΣΚ, όταν για την προσπέλασή τους χρησιμοποιείται ειδικό λογισμικό, βλ. Smith, Hoffmann & Rayson 2008).

16 Αναπτύχθηκε στο *Τμήμα Πληροφορικής & Τηλεματικής, Χαροκόπειο Πανεπιστήμιο*, από τον καθηγητή Ηρακλή Βαρλάμη και προσαρμόζεται σταδιακά στις ανάγκες της γλωσσικής έρευνας, εφόσον δεν υλοποιήθηκε εξαρχής με αυτό τον στόχο. Στον εντοπισμό των άρθρων αξιοποιεί, μεταξύ άλλων, την τεχνολογία RSS (*Rich Site Summary*), η οποία στηρίζεται στο πρότυπο XML και την οποία χρησιμοποιούν τα ηλεκτρονικά μέσα ενημέρωσης για τη συνοπτική δημοσίευση, σε πραγματικό χρόνο, των ροών ειδήσεων συνοδευόμενων από μεταδεδομένα, όπως είναι ο τίτλος τους και η ημερομηνία δημοσίευσης. Οι κανόνες εξαγωγής πληροφοριών διατυπώνονται σε XPath (*XML Path Language*, βλ. Εικόνες 1, 16-17) και στοχεύουν σε συγκεκριμένες πληροφορίες για κάθε άρθρο. Τα τελικά αρχεία παράγονται στο μορφότυπο XML (βλ. Εικόνα 3).

ρυητής *NDCrawler*), αποτελεί αντικείμενο αυτοτελούς ανακοίνωσης, ωστόσο αναφερόμαστε περιληπτικά στον τρόπο λειτουργίας του, επειδή σχετίζεται άμεσα με τον σχεδιασμό του υποσυστήματος των κειμένων της *Νεοδημίας*. Όπως προαναφέρθηκε, ο όγκος του υλικού βάσει του οποίου επιτυγχάνεται η ανίχνευση των νεολογισμών (ελληνογενών και ξενόγλωσσων: βλ. Εικόνες 5-7) και η παρακολούθησή τους σε κειμενικά συμφραζόμενα είναι πολλών εκατομμυρίων λέξεων και ανανεώνεται αδιάκοπα. Οι ανάγκες, λοιπόν, επιβάλλουν ένα υπολογιστικό εργαλείο που να συλλέγει και να κωδικοποιεί τα δεδομένα με τρόπο ώστε όχι μόνο να μεγιστοποιεί την απόδοση της διαδικασίας συλλογής, αλλά κυρίως να ελαχιστοποιεί την ανθρώπινη παρέμβαση και τα βήματα μετα-επεξεργασίας: να δημιουργεί, ει δυνατόν, «ετοιμοπαράδοτο» το σώμα κειμένων και οι χρήστες να επισημειώνουν μόνο τα επίπεδα πληροφορίας που δεν μπορούν να αναλυθούν αυτόματα. Ασφαλώς, μία τέτοια στρατηγική θέτει περιορισμούς στην αναπαράσταση του κειμένου, ώστε να προκρίνεται κάθε φορά η κωδικοποίηση αφενός με το μικρότερο υπολογιστικό κόστος αφετέρου με τη μεγαλύτερη περιγραφική επάρκεια στην απεικόνιση και στην επεξεργασία των δεδομένων.

Βασικό πλεονέκτημα και συνάμα καινοτομία του υποσυστήματος συλλογής κειμένων της *Νεοδημίας* είναι η συμβατότητά του με όλες τις παραδοχές σχεδιασμού (βλ. σχετική ενότητα), εφόσον: (α) πρόκειται για σύστημα εξειδικευμένο στην εξαγωγή κειμενικού περιεχομένου από δημοσιογραφικά άρθρα (βλ. Varlamis, Tsirakis, Poulopoulos & Tsantilas 2014), (β) η διαδικασία είναι επαναλαμβανόμενη, καθότι ο ανιχνευτής περιεχομένου (*crawler*) επισκέπτεται συγκεκριμένες πηγές δεδομένων ανά τακτά διαστήματα (κάθε 20 λεπτά της ώρας ή αλλιώς κάθε 1 200 δευτερόλεπτα) και συλλέγει (συναθροίζει) ροές ειδήσεων και άρθρα από συγκεκριμένες στήλες της τρέχουσας επικαιρότητας που ανανεώνονται διαρκώς, (γ) οι μηχανισμοί του εντοπισμού,[17] της επιλογής και της σύνθεσης της πληροφορίας προϋποθέτουν την ανάλυση των οπτικών και δομικών χαρακτηριστικών των πηγών, ώστε να απομονώνεται *εξαρχής αποκλειστικά* η πληροφορία που είναι συναφής με τις ανάγκες του χρήστη και να εξάγεται το κείμενο ήδη χαρακτηρισμένο ως προς τα δομικά του χαρακτηριστικά (τίτλος, περιεχόμενο) και τα βασικά του μεταδεδομένα (όνομα πηγής, ημερομηνία συγγραφής, URL κ.ά. – βλ. Εικόνα 1, καθώς και Εικόνες 16-17 στο *Παράρτημα*),[18] (δ) ο τρόπος

17 Πρβλ. Koutsis, Kouklakis, Mikros & Markopoulos (2005) για παράλληλο προβληματισμό και παρουσίαση ενός εργαλείου δημιουργίας σωμάτων κειμένων από το διαδίκτυο συμβατό με την Ελληνική, αξιοποιώντας για την πρόσβαση στα κείμενα, τη δομή των URLs των δικτυότοπων. Πρβλ. επίσης το *BootCat*, διαθέσιμο στο http://bootcat.dipintra.it, εργαλείο κατασκευής ΗΣΚ κατ' απαίτηση από το διαδίκτυο, το οποίο είναι εν μέρει συμβατό με ελληνικά δεδομένα.

18 Κάτι που, ενώ σε πρακτικό επίπεδο παρακάμπτει την ενεργοβόρο για κάθε υπολογι-

εντοπισμού και εξαγωγής της πληροφορίας προσομοιάζει την ανθρώπινη συμπεριφορά (Varlamis κ.ά. 2014: 3) – η πληροφορία επιλέγεται και αποθηκεύεται ακολουθώντας μια συγκεκριμένη σειρά στη διαδρομή της (σειριακή ροή), όμοια με αυτή που θα επέλεγε ένας αναγνώστης μετακινούμενος «υπερκειμενικά» στη διαδικτυακή εφημερίδα από το γενικό στο μερικό, ακολουθώντας συνδέσμους· παρακολουθεί την τρέχουσα επικαιρότητα, στη συνέχεια κινείται από το πρώτο σέλιδο στις ειδικές κατηγορίες ειδήσεων, καταλήγοντας στη σελίδα του άρθρου που τον ενδιαφέρει, το οποίο όμως «διαβάζει» επιλεκτικά.

Το σύστημα συλλογής ενσωματώνει και μια σειρά από υπο-προγράμματα με επικουρικές λειτουργίες (απαλοιφή διπλότυπων – συγχώνευση άρθρων, κατάτμηση ανά περιόδους συλλογής π.χ. μήνας, ημέρα) για τη βέλτιστη διαχείριση της πληροφορίας. Για λόγους συμβατότητας με λογισμικά επεξεργασίας κειμένων, τα δεδομένα διατηρούνται και ως απλά αρχεία κείμενου TXT.

Kathimerini.properties
#Mon Feb 02 14:53:10 EEST 2016
javascript=true
rss=http://www.kathimerini.gr/rss
charset=utf-8
name=Kathimerini
tagsToRemove=script,div
xpathbody=//div[@class="large-9 columns"]/div[@class="freetext"]
dateFormat=EEE, dd MMM yyyy HH:mm:ss zzz

Εικόνα 1. NDCrawler: Αρχείο ρυθμίσεων για τις ροές ειδήσεων της εφημερίδας «Η Καθημερινή»

στικό σύστημα – και με περιθώρια σφάλματος – διαδικασία «καθαρισμού» των δεδομένων (*data cleaning*) από αλλογενή τμήματα κώδικα, δημιουργεί ωστόσο αυξημένες απαιτήσεις για τους χρήστες του συστήματος που πρέπει εξαρχής να δώσουν τους κανόνες εξαγωγής κειμενικού και μετακειμενικού περιεχομένου: έργο πολυσύνθετο διότι, όσο και αν η δημοσίευση των ειδήσεων ακολουθεί συγκεκριμένα πρότυπα (*templates*), κάθε εφημερίδα έχει το δικό της περιβάλλον συγγραφής σε HTML, το οποίο *δεν μένει αμετάβλητο* στο πέρασμα του χρόνου, οπότε κάθε φορά πρέπει να αναβαθμίζονται και οι κανόνες, ώστε να συλλέγονται απρόσκοπτα τα κείμενα. Επίσης, όχι σπάνια οι δημοσιογραφικοί οργανισμοί αλλάζουν την ποσότητα των ειδήσεων που διοχετεύουν στις ροές και η γενικότερη τάση που παρατηρήθηκε είναι η σταδιακή αποδέσμευση των ειδικών ενθέτων των εφημερίδων (π.χ. με θέμα τη μαγειρική, τη μόδα κ.ά.) από τις γενικές ροές, ώστε να προβάλλεται με αυτές μόνο ο «πυρήνας» των ειδήσεων (πολιτική, οικονομία, κ.λπ.). Για παρουσίαση των πλεονεκτημάτων αλλά και των μειονεκτημάτων της μεθόδου, την παραμετροποίηση των πηγών δεδομένων σε XPath, καθώς και των βημάτων μετα-επεξεργασίας και ονοματοδοσίας του συστήματος αρχείων για την τελική συγκρότηση corpus με τον συγκεκριμένο τρόπο συλλογής βλ. Αφεντουλίδου (υπό προετοιμ.). Γενικότερα για τη μεθοδολογία πρβλ. Camozzo (2013).

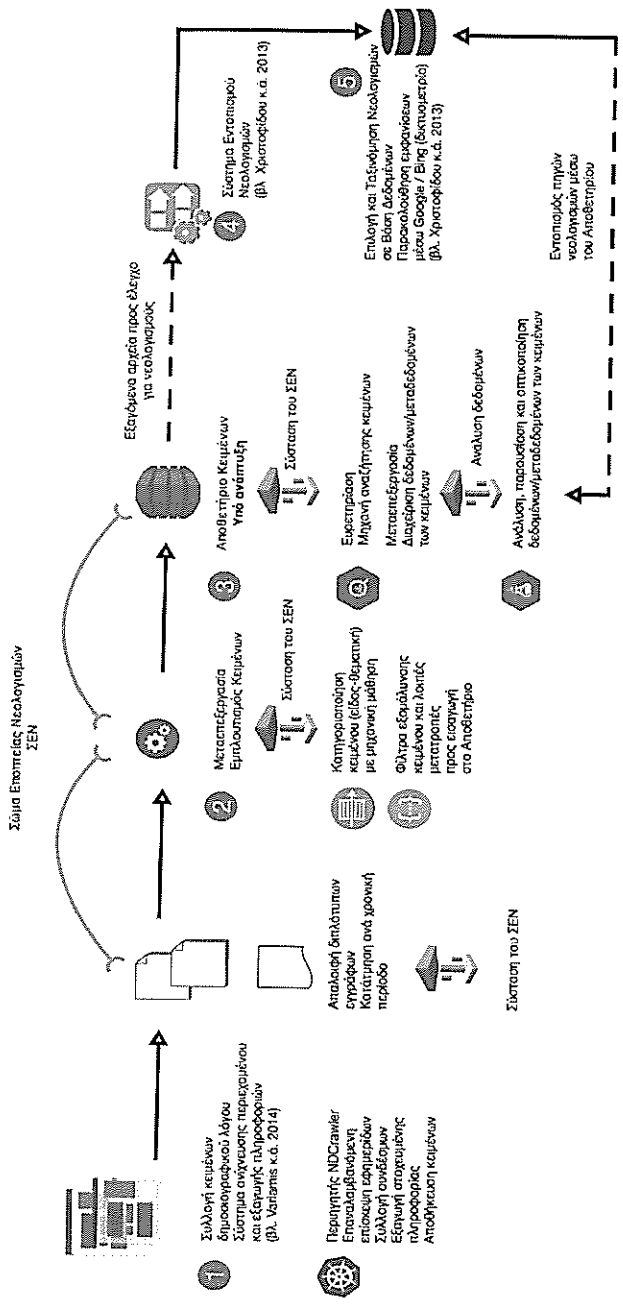
ισομοιάζει την ανθρώπινη ομ
α επιλέγεται και αποθηκεύετα
ομή της (σειριακή ροή), όμως
ούμενος «υπερκειμενικά» στη
, ακολουθώντας συνδέσμους
νέχεια κινείται από το πρωτο
οντας στη σελίδα του άρθρου
ιεκτικά.

ρά από υπο-προγράμματα με
συγχώνευση άρθρων, κατά
ι) για τη βέλτιστη διαχείριση
γισμικά επεξεργασίας κειμέ
κ κείμενου TXT.

class="freetext"]

ές ειδήσεων της εφημερίδας

«καθαρισμού» των δεδομένων
υστόσο αυξημένες απαιτήσεις
ισουν τους κανόνες εξαγωγής
ετο διότι, όσο και αν η δημο
lates), κάθε εφημερίδα έχει το
αμετάβλητο στο πέρασμα του
ινόνες, ώστε να συλλέγονται
τοί οργανισμοί αλλάζουν την
κότερη τάση που παρατηρή
ον εφημερίδων (π.χ. με θέμα
οβάλλεται με αυτές μόνο ο
υσίαση των πλεονεκτημάτων
ίηση των πηγών δεδομένων
ομοτοδοσίας του συστήμα
ριμένο τρόπο συλλογής βλ.
πρβλ. Camozzo (2013).



Εικόνα 2. Αρχιτεκτονική της Νεοδημίας (κυρίως: υποσύστημα συλλογής κειμένων και ΣΕΝ)

3. Κωδικοποίηση δεδομένων

3.1 Διεθνή πρότυπα κωδικοποίησης κειμένου – Text Encoding Initiative

Για την κωδικοποίηση των δεδομένων επιλέχθηκε ένα σχήμα XML: το διεθνές πρότυπο που καθιέρωσε η *Πρωτοβουλία για την Κωδικοποίηση Κειμένων* (*Text Encoding Initiative*, εφεξής TEI), στην 5η, την πιο πρόσφατη αναθεώρησή του (TEI P5, The TEI Consortium 2016). Αν και καθιερωμένο επί έτη (από το 1987) με εκτενέστατες *Οδηγίες* (*Guidelines for Electronic Text Encoding and Interchange*) και συμβατό με την XML ήδη από το 2002, η ελληνόγλωσση βιβλιογραφία για το TEI είναι ελάχιστη, όπως επισημαίνεται στο Ακριτίδου (2014). Ωστόσο, το ενδιαφέρον της ελληνικής επιστημονικής κοινότητας αυξάνει ολοένα.[19]

Είναι αλήθεια ότι η Επιστήμη της Γλωσσολογίας από νωρίς ανέπτυξε πληθώρα προτύπων επισημείωσης (ενδεικτικά βλ. Lehmberg & Wörner 2008, Ide 2008, Ide, Romary & de la Clergerie 2004, Kahrel, Barnett & Leech 1997, Kübler & Zinsmeister 2015, Stührenberg 2012). Συγκεκριμένα, ένα από αυτά, το XCES (*Corpus Encoding Standard for XML*)[20] διαδόθηκε αρκετά στην ελληνική ερευνητική κοινότητα και οδήγησε σε αντίστοιχες εφαρμογές και συμβατά γλωσσικά εργαλεία.[21] Η εξέλιξή του, ωστόσο, έχει σήμερα «παγώσει» από τους δημιουργούς του και γενικώς έχει περιοριστεί η εμβέλειά του στη διεθνή ερευνητική κοινότητα.

Τα πρότυπα επισημείωσης που προέρχονται από τον *Διεθνή Οργανισμό Τυποποίησης* ISO (ιδίως τα ISO/TC 37/SC4 «Ορολογία και άλλοι γλωσσικοί πόροι και πόροι περιεχομένου – Διαχείριση γλωσσικών πόρων», τα οποία δεν επιτρέπει ο περιορισμός των σελίδων να παρουσιάσουμε – βλ. Lehmberg & Wörner 2008) είναι εξαιρετικά λεπτομερή ως προς τις αναλυτικές μονάδες περιγραφής των γλωσσικών στοιχείων, διαθέτουν, ωστόσο, αρκετή πολυπλοκότητα, απαγορευτική ως προς την υλοποίησή τους στο παρόν εγχείρημα. Αντιθέτως, το πρότυπο TEI, όπως θα παρουσιαστεί στη συνέχεια της μελέτης, με την ευελιξία και την προσαρμοστικότητά του στις ανάγκες των ανθρωπιστικών επιστημών,[22] αλλά και τη βασική του

19 Βλ. Δημητρούλια & Τικτοπούλου (2015), το πρώτο, από όσο γνωρίζουμε, ελληνικό ακαδημαϊκό σύγγραμμα στο οποίο περιλαμβάνεται η αναλυτική παρουσίαση του προτύπου με παραδείγματα από λογοτεχνικά σώματα κειμένων.

20 Αποτελεί τροποποίηση και προσαρμογή του TEI P3 (SGML) στις ανάγκες της γλωσσικής ανάλυσης και της κωδικοποίησης ΗΣΚ.

21 Βλ. τα σώματα κειμένων και τα εργαλεία μορφοσυντακτικής ανάλυσης και λημματοποίησης του *Ινστιτούτου Επεξεργασίας του Λόγου*, όπως περιγράφονται και διατίθενται μέσω της Πύλης της ερευνητικής υποδομής *clarin:el*, στο http://www.clarin.gr

22 Τα παραδείγματα είναι πολυάριθμα και ενδεικτικά παρατίθενται στη βιβλιογραφία: *Digital Archive of Letters in Flanders* (Vanhoutte & Van den Branden 2003), *Deutsches Tex-*

παραδοχή για την ιεραρχική φύση των κειμένων[23] αποτέλεσε ισχυρό μεθοδολογικό εργαλείο για την υλοποίηση του ΣΕΝ.

3.2 Υλοποίηση – Η ανάπτυξη ψηφιακού σώματος δημοσιογραφικού λόγου – Στρατηγικές επισημείωσης του ΣΕΝ

3.2.1 Επίπεδα ανάλυσης

Το σχήμα επισημείωσης των κειμένων (για περαιτέρω λεπτομέρειες βλ. Εικόνες 10-15 στο *Παράρτημα*) αναπτύσσεται με γνώμονα τα παρακάτω επίπεδα ανάλυσης:

Α) Επίπεδο *περικείμενου*: περιγράφονται χαρακτηριστικά που συνδέονται με το καταστασιακό περιβάλλον του κειμένου, τον σκοπό και τις συνθήκες παραγωγής του, τον δημιουργό και τους αποδέκτες του, καθώς και στοιχεία εξωγλωσσικά που το προσδιορίζουν λειτουργικά και ειδολογικά σε σχέση με άλλα κείμενα: πρόκειται για το επίπεδο ανάλυσης που παραδοσιακά καλύπτεται από την κεφαλίδα των μεταδεδομένων, δηλαδή πληροφοριών που περιγράφουν τα δεδομένα.

Β) Επίπεδο *κειμένου*: συνδέεται με κειμενικά, μακροδομικά συστατικά των δημοσιογραφικών άρθρων, για παράδειγμα την οργάνωσή τους σε λειτουργικές ενότητες και ακολουθίες, την αφηγηματική διάσταση των γραπτών ειδήσεων, αλλά και τις διεκφωνηματικές σχέσεις μέσα στις παραγράφους, στο ευρύτερο συγκείμενο (λ.χ. το δίκτυο των αναφορικών δεσμών, δείκτες συνοχής-συνεκτικότητας κ.ά.).

Γ) Επίπεδο *εκφωνήματος*: στο μεταίχμιο μεταξύ μορφολογίας και κειμενικής ανάλυσης, οι λέξεις εξετάζονται στο άμεσο συγκείμενό τους, οι αναλύσεις επομένως εντάσσονται σε εκφωνήματα (ολιστική προσέγγιση). Το επίπεδο επιτρέπει την καταγραφή και άλλων παραμέτρων, πραγματολογικών και υφολογικών, όπως λεξικογραμματικά σχήματα, φραστικοί σχηματισμοί και λεξιλογικές συνάψεις (λ.χ. πολυλεκτικές σύνθετες νεολογικές εκφράσεις), σημασιοσυντα-

---

larchiv, Medieval Nordic Text Archive (Haugen 2008), *Van Gogh's Letters* (Jansen, Luijten & Bakker 2009), *Women Writers Project, The Shelley-Godwin Archive, Corpus Est Républicain* κ.ά. Βλ. και τον *Κατάλογο Ψηφιακών Εκδόσεων* της Greta Franzini, διαθέσιμο στο https://dig-ed-cat.eos.arz.oeaw.ac.at

23 Πρόκειται για την περίφημη *'OHCO (Ordered Hierarchy of Content Objects) thesis'*, θέση που διατυπώθηκε από τους DeRose, Durand, Mylonas & Renear (1990), Renear, Mylonas & Durand (1993) και διατρέχει σε φιλοσοφικό, οντολογικό επίπεδο το TEI μέχρι σήμερα. Επιπλέον, για τις γλωσσολογικές αναλύσεις και την ανάγκη να αναπαρασταθούν σε πολλαπλά επίπεδα επισημείωσης, το TEI έχει ήδη ενσωματώσει πιο περίπλοκες δυνατότητες αναπαράστασης, ώστε ο Stührenberg (2012: 8-9) να προβλέπει το σταδιακό πέρασμα στο μέλλον από τα ιεραρχικά, «δενδρικά» μοντέλα εγγράφων (*hierarchical data models*) στα μοντέλα γράφων (*graph-based formal models*), όσο οι τεχνολογίες το επιτρέπουν.

κτικά φαινόμενα και ρητορικές πρακτικές που γειτνιάζουν με τους νεολογικούς σχηματισμούς (λ.χ. μεταγλωσσικοί δείκτες επεξήγησης νέων λέξεων, αλλαγές στη συμφραστική συμπεριφορά μεμονωμένων λέξεων, αναφορικοί δεσμοί και επαναλήψεις νέων όρων).

Δ) Επίπεδο *λέξης*: το κατεξοχήν επίπεδο στο οποίο ανιχνεύονται και χαρακτηρίζονται οι νεολογισμοί ως αυτοτελή λεξήματα με πολλαπλό χαρακτηρισμό ως προς τις *μορφολογικές διαδικασίες σχηματισμού*, καθώς και τη *θεματική* τους (ένταξη σε σημασιολογικά πεδία, κατά το πρότυπο συστημάτων περιγραφής επώνυμων οντοτήτων, διευρυμένο ως προς την εφαρμογή του στο γενικό λεξιλόγιο).[24] Πρόκειται για το πιο απαιτητικό ως προς την υλοποίησή του επίπεδο, για τρεις λόγους: 1) Προϋποθέτει ειδική μελέτη για τη συμβατότητά του με το σχήμα της βάσης δεδομένων της *Νεοδημίας*, το οποίο έχει ήδη υλοποιηθεί σε τεχνολογία PHP-MySQL, ώστε η μία εφαρμογή να συμπληρώνει την άλλη. 2) Το TEI καλύπτει τέτοιου είδους αναλύσεις / αναπαραστάσεις γλωσσικών αντικειμένων με τις «δομές χαρακτηριστικών» (*feature structures–FS*).[25] Η πολυπλοκότητά τους, ωστόσο, κρίνεται «απαγορευτική» για έργα που εμπλέκουν τον ανθρώπινο παράγοντα στην επισημείωση των δεδομένων και προκρίνεται σε αυτόματα συστήματα επεξεργασίας. Σε κάθε περίπτωση, η εφαρμογή μίας τέτοιας δυνατότητας με τρόπο ώστε ο χρήστης να μπορεί να παρεμβαίνει στο έργο του χαρακτηρισμού καθορίζεται από τις εκάστοτε διαθέσιμες τεχνολογικές μεθόδους και οι κατηγορί-

24 Πρβλ. το σύστημα σημασιολογικού χαρακτηρισμού που αναπτύχθηκε για τον σημασιολογικό αναλυτή *Lancaster Semantic Tagger*, UCREL Semantic Analysis System (USAS, στο http://ucrel.lancs.ac.uk/usas/).

25 Πρόκειται για μακροσκελείς ακολουθίες χαρακτηριστικών-τιμών (*features-values*) που μπορούν να επεκταθούν *ad infinitum*. Προέρχονται από τον γραμματικό φορμαλισμό HPSG (*Head-Driven Phrase Structure Grammar*, βλ. Pollard & Sag 1994) για την υπολογιστική επεξεργασία της φυσικής γλώσσας. Οι δομές χαρακτηριστικών δημιουργούν «έλικες» χαρακτηρισμών που δεν προστίθενται σειριακά μέσα στο κείμενο (*inline*), αλλά σε ξεχωριστό σημείο (*stand-off αρχιτεκτονική*) και συνδέονται με τις μονάδες που χαρακτηρίζουν μέσω ειδικών αναγνωριστικών (αναφορικών δεσμών). Έτσι το κείμενο αποθηκεύεται ανέπαφο σε διαφορετικό σημείο από τη γλωσσική ανάλυση, για την οποία μπορούν να προστεθούν θεωρητικά άπειρα επίπεδα επισημείωσης, καθώς οι σημειωτικές ψηφίδες αποδεσμεύονται από τη σειριακή πραγμάτωσή τους (*serialization*), βλ. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/FS.html Πρόκειται για την πάγια τακτική διαχείρισης δεδομένων σε αυτόματα συστήματα επεξεργασίας ΗΣΚ, ωστόσο, η υποδοχή των FS από την ερευνητική κοινότητα δεν έγινε χωρίς σκεπτικισμό. Για ενστάσεις βλ. ενδεικτικά Bański 2010, καθώς και την ανοιχτή πρόταση–κάλεσμα της ομάδας του TEI SIG (*Special Interest Group*) for Linguists για την ανάπτυξη ενός πλαισίου που θα χαρακτηρίζει με περιγραφική απλότητα, επάρκεια και ευελιξία (*corpus TEI simple*) το επίπεδο της λέξης <w> και τα ορίσματά του, βλ. *'Strong proposal for (modestly) extending the set of linguistic attributes to <w>'*, διαθέσιμο στο https://github.com/LingSIG/wordAttributes/wiki

τους νεολογικούς
λέξεων, αλλαγές
ορικοί δεσμοί και

ύονται και χαρα-
λό χαρακτηρισμό
τη *θεματική* τους
περιγραφής επώ-
γενικό λεξιλόγι-
του επίπεδο, για
ι του με το σχήμα
ιηθεί σε τεχνολο-
λη. 2) Το TEI κα-
αντικειμένων με
πλοκότητά τους,
ανθρώπινο παρά-
τόματα συστήμα-
; δυνατότητας με
υ χαρακτηρισμού
; και οι κατηγορί-

θηκε για τον σημα-
ysis System (USAS,

features-values) που
φορμαλισμό HPSG
υπολογιστική επε-
«έλικες» χαρακτη-
ε ξεχωριστό σημείο
ίζουν μέσω ειδικών
ανέπαφο σε διαφο-
ροστεθούν θεωρη-
εσμεύονται από τη
ase/doc/tei-p5-doc/
ιν σε αυτόματα συ-
τική κοινότητα δεν
ώς και την ανοιχτή
iguists για την ανά-
άρκεια και ευελιξία
'*Strong proposal for*
https://github.com/

ες περιγραφής που είναι αναγκαίο να αναπτυχθούν για το επίπεδο Δ, όπως και η αρχιτεκτονική της κωδικοποίησης (*inline vs. offline*), αποτελούν αντικείμενο νέας ανακοίνωσης. 3) Η ενσωμάτωση αυτοματοποιημένων μεθόδων μορφοσυντακτικού χαρακτηρισμού και λημματοποίησης όλων των λεξικών μονάδων (*tokens*) του σώματος κειμένων αποτελεί πρόκληση και απώτατο στόχο για το επίπεδο Δ. Ωστόσο, πρόκειται για εγχείρημα που απαιτεί χρόνο και αρκετό πειραματισμό, καθότι οι υπάρχοντες μορφοσυντακτικοί αναλυτές χρειάζεται να παραμετροποιηθούν[26] ως προς τη δομή του αρχείου που παράγουν, ώστε να είναι φιλικοί αφενός με το πρότυπο της αναπαράστασης των δεδομένων αφετέρου με τις περιγραφικές ανάγκες ανάλυσης των νεολογικών σχηματισμών και της ορολογίας.

Δεδομένου του εύρους των στρωμάτων ανάλυσης και της ιδιαίτερης τεχνογνωσίας που απαιτεί καθένα από τα τέσσερα επίπεδα, η παρούσα μελέτη επικεντρώνεται στα δύο πρώτα. Συγκεκριμένα, παρουσιάζουμε την πρότασή μας για την κωδικοποίηση του σώματος κειμένων της *Νεοδημίας* κατά το διεθνές πρότυπο TEI, με έμφαση στον χαρακτηρισμό της ταυτότητας των κειμένων (την ειδολογική κατάταξή τους) και φυσικά στην αναπαράστασή τους ως δομικά οργανωμένου συνόλου, πάντοτε σε άμεση συνάρτηση με τον τρόπο, τις δυνατότητες πρόσκτησης του υλικού, καθώς και την ποσότητά του.

3.2.2 Ανάλυση εγγράφου

3.2.2.1 Κυρίως κείμενο (στοιχεία <text, body>) – Δεδομένα

Στις ηλεκτρονικές εκδόσεις εφημερίδων το περιβάλλον απεικόνισης της πληροφορίας είναι πολυτροπικό με πολλαπλές ζώνες ανάγνωσης: το δημοσιογραφικό κείμενο ως προϊόν λόγου διαφοροποιείται από τις στήλες σχολιασμού των αναγνωστών, τους υπερδεσμούς στα «συναφή άρθρα» που προτείνει αυτόματα το σύστημα ευρετηρίασης των ειδήσεων, τις διασυνδέσεις με τα μέσα κοινωνικής δικτύωσης, τα πλαίσια με τις διαφημίσεις και τις γενικότερες κατηγορίες πλοήγησης μέσα στον δικτυότοπο της εφημερίδας, τη συμπαράθεση πολυμεσικών στοιχείων από άλλα σημειωτικά συστήματα, π.χ. εικονογραφικού υλικού (φωτογραφίες, χάρτες, σχεδιαγράμματα, πληροφοριογραφήματα, σκίτσα, ευθυμογραφήματα, εικονομηνύματα), αρχείων κινούμενης εικόνας και ήχου (βίντεο, ηχητικά αποσπάσματα) κ.ά. Είναι αλήθεια ότι η εμπειρία της ανάγνωσης μιας ηλεκτρονικής εφημερίδας πραγματώνεται μέσα σε αυτό το υπερκειμενικό πλαί-

26 Καθοριστική στο στάδιο αυτό είναι η σύνδεση του Κέντρου με την ομάδα της *Neurolingo* και τον Διευθυντή της Δρ Χρήστο Τσαλίδη (τον οποίο ευχαριστούμε) με στόχο την ενσωμάτωση στη *Νεοδημία*, των γλωσσικών εργαλείων επεξεργασίας φυσικής γλώσσας που έχουν αναπτυχθεί για τη μορφοσυντακτική αναγνώριση και τη λημματοποίηση.

σιο γραφής, σε ένα δίκτυο πολλαπλών τρόπων και διακειμένων και το νόημα που αποκομίζει ο επισκέπτης του δικτυακού τόπου παράγεται μέσα στο περιβάλλον αυτό, όσο σύνθετο και αν είναι. Ωστόσο, για τη μελέτη του ψηφιακού κειμένου και την κωδικοποίησή του σε σώμα για τις ανάγκες της έρευνας των νεολογισμών – αλλά και κάθε γλωσσολογικής έρευνας – είναι μεθοδολογικά απαραίτητη η αποπλαισίωσή του από τον πολυσημειωτικό αυτό ιστό ανάγνωσης και η αναπλαισίωσή[27] του σε ένα νέο περιβάλλον μελέτης και υπομνηματισμού: ήδη κατά τη διαδικασία της συλλογής απομονώνεται η βασική κειμενική μονάδα ανάλυσης που κατά το πρότυπο TEI είναι το στοιχείο <text>.

Συγκεκριμένα, ως γενικότερη στρατηγική αποφασίστηκε η αναπαράσταση του κειμένου σε XML να στοχεύει στην ελάχιστη έμφαση στη μορφή (στα οπτικά του χαρακτηριστικά και την απόδοση των τυπογραφικών συμβάσεων των εφημερίδων) και στην όσο το δυνατόν πληρέστερη αποτύπωση της δομής του κειμένου και ασφαλώς του περιεχομένου του. Έτσι το δημοσιογραφικό άρθρο αποθηκεύεται εξαρχής «απογυμνωμένο» από εικονογραφικά και άλλα πολυτροπικά στοιχεία, καθώς, φυσικά και από κάθε εξωκειμενική, εξωγλωσσική πληροφορία μη σχετική με τη σύσταση του ΣΕΝ (μενού πλοήγησης, σχόλια αναγνωστών, υπερσύνδεσμοι σε εξωτερικές διαδικτυακές πηγές, δηλαδή συνοδευτικά κείμενα ή άλλα ψηφιακά αντικείμενα). Το έγγραφο που παράγεται αυτόματα από το υποσύστημα ανίχνευσης περιεχομένου και εξαγωγής πληροφοριών ακολουθεί το εξής πρότυπο (*document model*):

1) Το κυρίως κείμενο περικλείεται στο στοιχείο <body> και διακρίνουμε δύο ζώνες ανάγνωσης. Οι ζώνες αυτές κωδικοποιούνται με ιδιαίτερα στοιχεία που παρέχει το πρότυπο αναπαράστασης. Συγκεκριμένα:

27 Για τη μεταφορά της έννοιας της αναπλαισίωσης (*recontextualization* κατά Bernstein 1996) στα ηλεκτρονικά περιβάλλοντα και στα σώματα κειμένων βλ. Κουτσογιάννης (2009: 156-160).

28 Αποφύγαμε την κωδικοποίηση τέτοιων στοιχείων με απλή πρόσθεση της τιμής "caption" στο γνώρισμα @type του ήδη χρησιμοποιημένου στοιχείου <head>, προκειμένου 1) να διαφανεί με τον τρόπο αυτό η διαφοροποίηση των επικεφαλίδων αυτών από τα άλλα λειτουργικά στοιχεία που απαρτίζουν το σύστημα των τίτλων, καθότι οι λεζάντες συνοδεύουν διαφορετικούς σημειωτικούς κώδικες από ό,τι οι παραδοσιακές επικεφαλίδες, 2) να μπορούν να προστεθούν στο μέλλον, αν οι ανάγκες για προστιθέμενη πληροφορία το απαιτήσουν, νέα ορίσματα στο στοιχείο <caption>, π.χ. το γνώρισμα @source για σύνδεση του κειμένου με το ψηφιακό τεκμήριο που συνοδεύει.

Ζώνες ανάγνωσης	Πραγματώσεις (δομικά μέρη του κειμένου)	Κωδικοποίηση σε TEI
Κείμενο των τίτλων(headings)		με ιδιότητα @type και τιμή "headings"
	Α. «Παραδοσιακό» σύστημα των τίτλων (κυρίως τίτλος άρθρου, υπέρτιτλος, υπότιτλος, μεσότιτλος δηλ. τίτλος παραγράφου/ων)	<head> με ιδιότητα @type και τιμές "surTitle", "mainTitle", "subTitle", "midTitle"
	Β. Εισαγωγική παράγραφος (*lead*) εν είδει επικεφαλίδας στο κείμενο που ακολουθεί	<head> με ιδιότητα @type και τιμή "lead"
	Γ. Τίτλοι πολυμεσικού υλικού (λεζάντες σε εικονογραφικό υλικό και αρχεία βίντεο που υποστηρίζουν το περιεχόμενο του γραπτού κειμένου)	<caption>[28]
Κυρίως κείμενο (mainText)		με ιδιότητα @type και τιμή "mainText"
	Α. Παράγραφοι	<p>
	Β. Άλλη διάταξη (με αρίθμηση ή χωρίς)[29]	<list> σε συνδυασμό με το υποστοιχείο <item>

Πίνακας 1. Κωδικοποίηση του στοιχείου <body>

2) Αμιγώς παρακειμενικά στοιχεία (συνοδευτικά του κειμένου χωρίς να ανήκουν οργανικά σε αυτό) θεωρήθηκαν (α) το όνομα του συγγραφέα του άρθρου, (β) πιθανή αναφορά στον αρχικό διανομέα της είδησης (π.χ. *Αθηναϊκό-Μακεδονικό Πρακτορείο Ειδήσεων*), (γ) δηλώσεις για την ιδιότητα του συγγραφέα, εάν δεν είναι επαγγελματίας δημοσιογράφος (συνήθως αναφέρονται στο επάγγελμά του και σημειώνονται με αστερίσκο εκτός κειμένου), (δ) ενδείξεις χρόνου δημοσίευσης του κειμένου μέσω συναφών αυτοματισμών (*timestamps*), τέλος, (ε) αναφορά προέλευσης της δημοσίευσης, αν δηλαδή το κείμενο έχει δημοσιευθεί και στην έντυπη έκδοση της εφημερίδας. Όλες οι ανωτέρω πληροφορίες αποφασίστηκε να καταγράφονται εκτός κυρίως κειμένου, στην κεφαλίδα των μεταδεδομένων (βλ. επόμενη ενότητα).

3) Επισημαίνουμε ότι στη σήμανση των δομικών και λειτουργικών τμημάτων του δημοσιογραφικού κειμένου το πρότυπο εγγράφου που προτείνουμε για το ΣΕΝ αποτελεί προϊόν νοητικής αφαίρεσης του κειμένου και αναγωγής του σε δύο βασικές ζώνες ανάγνωσης, εφόσον συσταδοποιεί σε ενότητες (*divisions*) τα στοιχεία τίτλων και τα διαχωρίζει από το κυρίως κείμενο. Μία τέτοια αρχιτεκτονική στην ιεραρχία των κειμενικών μονάδων υιοθετήθηκε για τους εξής λόγους: (α) σε θεωρητικό επίπεδο προσομοιάζει την ίδια τη λειτουργία της ανάγνωσης, κα-

29 Επισημειώνονται προς το παρόν ημιαυτόματα με «εύρεση-αντικατάσταση».

θότι το εφημεριδικό σύστημα των τίτλων έχει αυθύπαρκτη παρουσία και διαβάζεται ανεξάρτητα από τα κυρίως δημοσιογραφικά κείμενα,[30] (β) σε υπολογιστικό επίπεδο είναι συμβατή με τη βέλτιστη λειτουργία του προγράμματος συλλογής των κειμένων,[31] (γ) παρακάμπτει το πρόβλημα των αυθαίρετων διαιρέσεων (από μία μηχανή) των κειμένων σε ενότητες με βάση την ύπαρξη μεσότιτλου.[32]

3.2.2.2 Κεφαλίδα εγγράφου <header> – Μεταδεδομένα

Πρόκειται για το τμήμα του εγγράφου στο οποίο περιγράφονται τα κείμενα μιας οποιασδήποτε συλλογής με περαιτέρω πληροφορίες, τα ονομαζόμενα «μεταδεδομένα». Ο σχεδιασμός της κεφαλίδας έγινε με γνώμονα τα εξής:

1) Την επέμβαση στη λογική δομή του εγγράφου με στόχο την ελαχιστοποίηση του «μεικτού περιεχομένου» (*mixed content*) και την αντικατάστασή του με στοιχεία χωρίς περιεχόμενο («κενά στοιχεία», *empty elements*), εφόσον η σύνταξη της XML το επιτρέπει.[33] Έτσι μειώσαμε κατά πολύ την έκταση των αρχείων που παρά-

30 Οι τίτλοι, κατά κανόνα, δεν συντάσσονται από τους ίδιους τους συγγραφείς των άρθρων (κατά τη στιγμή δηλαδή παραγωγής του άρθρου) αλλά από «πεπειραμένους συντάκτες ή αρχισυντάκτες» (Τσελίγκα-Γκαζιάνη 2008: 601, Χατζησαββίδης 1999: 109).

31 Δηλαδή τον (σειριακό) τρόπο που παράγει αρχεία XML από τις πηγές ειδήσεων που επισκέπτεται.

32 Ένας εναλλακτικός τρόπος κωδικοποίησης (και περισσότερο συμβατός με το TEI, ευχαριστούμε την Elli Mylonas για την επισήμανση) θα διαιρούσε τα κείμενα όχι με βάση τη λειτουργία τους (διχοτόμηση σε τμήματα που λειτουργούν ως επικεφαλίδες και σε τμήματα συνεχούς κειμένου), αλλά με βάση τη δομική τους αλληλουχία (τα τμήματα καθορίζονται από την παρουσία/απουσία τίτλων που διακόπτουν το κυρίως κείμενο και ως εκ τούτου η/οι παράγραφος/οι που ακολουθούν έναν μεσότιτλο εγκιβωτίζονται όλες μαζί σε μια ενότητα , μαζί με τον μεσότιτλο, διαφοροποιούνται δηλαδή από το τμήμα του κειμένου που προηγείται). Ωστόσο, η ανωτέρω διαίρεση προϋποθέτει την ανθρώπινη διακριτική ικανότητα που μπορεί να εντοπίσει τις πραγματικές ιεραρχίες, κατά πόσο δηλαδή έντιτλες παράγραφοι μέσα σε κατά κανόνα σύντομα κείμενα αποτελούν υποενότητες ποιων (υπο)ενοτήτων. Συνεπώς, μια τέτοια ανάλυση, δεδομένου και του όγκου του υλικού, προτείνεται στην περίπτωση της *Νεοδημίας* μόνο ως απότοκο μετα-επεξεργασίας των κειμένων από ανθρώπους-επισημειωτές.

33 Πρόκειται για δυνατότητα της XML να απεικονίζει για τις βασικές μονάδες της (στοιχεία) όχι μόνο ιεραρχίες (μέσω της σήμανσης στοιχείων και υποστοιχείων σε σχέση γονέα-παιδιού) αλλά και «χαλαρότερες», «επίπεδες» δηλώσεις *με μεικτό περιεχόμενο*, δηλαδή με τη συνύπαρξη των υποστοιχείων, στο ίδιο επίπεδο, με απλές λέξεις και τμήματα κειμένου ως εξής: <στοιχείο>[[<υποστοιχείο>Η περίοδος αυτή περιέχει στοιχεία κειμένου (PCDATA) που διαβάζονται από τους ανθρώπους, ενώ η XML διαβάζεται από τις μηχανές</υποστοιχείο>]]</στοιχείο> όπου οι αγκύλες [[]] οριοθετούν το μεικτό περιεχόμενο. Αν δεν υπήρχε η δυνατότητα αυτή, δεν θα μπορούσε να χρησιμοποιηθεί η XML στα σώματα κειμένων. Ωστόσο, επειδή ακριβώς υπάρχει η δυνατότητα αυτή δημιουργούνται «χαλαρότερες ιεραρχικά» ακολουθίες δεδομένων, δομικό χαρακτηριστικό της XML που, όπως επισημαίνει ο Fraser (2011) για το *Greek Lexicon Project* (*Cambridge Greek Lexicon*) αποδυναμώνει τη λογική πειθαρχία και την περιγραφική ακρίβεια των ιεραρχικών σχέσεων μεταξύ στοιχείων-υποστοιχείων.

γονταν διατηρώντας το βάθος της ιεραρχίας των στοιχείων. Επιπλέον, σε αντίθεση με την πάγια τακτική του TEI να προκρίνει το μεικτό περιεχόμενο παντού, διαχωρίσαμε πληροφοριακά το τμήμα των μεταδεδομένων από τα δεδομένα (στα οποία, αντιθέτως, επικρατεί το μεικτό περιεχόμενο και αποφεύγονται τα κενά στοιχεία). Τέλος, με την ανωτέρω στρατηγική πετύχαμε τον βέλτιστο βαθμό συμβατότητας του ΣΕΝ με υπολογιστικά εργαλεία επεξεργασίας κειμένων.[34]

2) Αξιοποιήθηκαν οι καθιερωμένες σημειωτικές ψηφίδες του TEI για την κωδικοποίηση των μεταδεδομένων στα τέσσερα βασικά υποσύνολα πληροφορίας της ψηφιακής κεφαλίδας, με τροποποιήσεις και προσθήκες στο τρίτο επίπεδο (βλ. Εικόνα 3):

Α) Τμήμα περιγραφής αρχείου <fileDesc> (*file description*): δήλωση τίτλου ηλεκτρονικού αρχείου <titleStmt> (*title statement*), δήλωση υπευθυνότητας για την κωδικοποίηση του κειμένου <respStmt> (*statement of responsibility*), δήλωση δημοσίευσης <publicationStmt> (*publication statement*), πρόσβαση και διανομή <availability> <licence>, περιγραφή της πηγής προέλευσης <sourceDesc> (*source description*).

Β) Τμήμα δήλωσης των συμβάσεων κωδικοποίησης <encodingDesc> (*encoding description*): ταυτότητα του ερευνητικού έργου <projectDesc> (*project description*), περιγραφές των συστημάτων ταξινόμησης <classDecl> (*classification declarations*), βλ. επόμενη παράγραφο.

Γ) Τμήμα αναπαράστασης της ταυτότητας των κειμένων <profileDesc> (*profile description*): επιμέρους περιγραφή κειμένου <textDesc> (*text description*), βλ. επόμενη παράγραφο.

Δ) Τμήμα περιγραφής του ιστορικού αναθεωρήσεων <revisionDesc> (*revision description*): αποτύπωση σημαντικών επεμβάσεων στο ΣΕΝ.

3) Στο τμήμα περιγραφής της ταυτότητας των κειμένων <profileDesc> (*profile description*) και στο υποστοιχείο του <textDesc> (*text description*), όπου δηλώνονται επιμέρους περιγραφικά μεταδεδομένα, μας απασχόλησε η αποτύπωση παραμέτρων σχετικά με τα βασικά στοιχεία της ταυτότητας των κειμένων του ΣΕΝ. Αναπόφευκτα, ένα σύστημα κατηγοριοποίησης κειμένων τέμνεται με τη θεωρητική θέση για το τι συνιστά κειμενικό είδος. Για την ανάπτυξη του σχήματος επισημείωσης για το επίπεδο αυτό πληροφορίας:

Α) Μελετήσαμε τους τρόπους με τους οποίους το TEI δίνει τη δυνατότητα κειμενικής ταξινόμησης: (α) στο τμήμα περιγραφής κειμένου <profileDesc> με το στοιχείο <textClass> (*text classification*) και τα υποστοιχεία του, βάσει

34 Για παράδειγμα, στην ορθή καταμέτρηση των λέξεων και στη μεγαλύτερη ευελιξία στις αναζητήσεις. Τα εργαλεία που δοκιμάστηκαν επιτυχώς είναι το *WordSmith Tools* v. 7.0, το *TXM – Projet Textométrie* (Heiden, Magué & Pincemin 2010) και η *Sketch Engine* (Kilgarriff κ.ά. 2014).

των οποίων δηλώνονται οι τιμές οιασδήποτε θεματικής ή κειμενικής ταξινομίας υιοθετείται, είτε καθιερωμένης στην επιστημονική κοινότητα, όπως το πρότυπο Dublin Core, είτε προκαθορισμένης από τους χρήστες βάσει λέξεων-κλειδιών ή ελεγχόμενων λεξιλογίων (*controlled vocabularies*), (β) στην περιοχή <encodingDesc> (*encoding description*), όπου περιγράφονται τα σχήματα κωδικοποίησης των δεδομένων μέσω του υποστοιχείου <classDecl> (*classification declarations*) και με τερματικό στοιχείο το <catDesc> (*category description*), το οποίο δέχεται είτε απλούς χαρακτηρισμούς σε ελεύθερο κείμενο είτε προκύπτει ως συνισταμένη παραμέτρων που διακρίνουν τα κείμενα στην καταστασιακή τους διάσταση.[35] Οι δύο διαδρομές είναι ανεξάρτητες μεταξύ τους αλλά μπορούν να συνυπάρχουν στο ίδιο ΗΣΚ. Η διττή δυνατότητα αναπαράστασης μιας τέτοιας πληροφορίας, αν και συνοδεύεται από ορισμένο βαθμό σημειωτικής ασάφειας (χρησιμοποιείται άλλοτε για τη θεματική κατάταξη, άλλοτε για να αποδώσει την ειδολογική ταυτότητα του κειμένου) πιστοποιεί, ακριβώς, τη σπουδαιότητά της για το TEI και υλοποιείται με τρόπο ώστε να διασφαλίζεται η διαλειτουργικότητα με όλα τα καθιερωμένα διεθνή βιβλιογραφικά πρότυπα θεματικής και τυπολογικής περιγραφής τεκμηρίων.[36]

Για το ΣΕΝ προτιμήσαμε έναν εναλλακτικό τρόπο κωδικοποίησης, στο πνεύμα της δεύτερης διαδρομής κατηγοριοποίησης, τροποποιώντας το ίδιο το πρότυπο TEI. Το σκεπτικό που μας οδήγησε σε αλλαγές ήταν η ανάγκη ενοποίησης ομογενών στοιχείων και οι επιταγές της περιγραφικής επάρκειας: κατά τη γνώμη μας, η ένταξη ενός κειμένου σε θεματικές περιοχές αποτελεί στοιχείο αλληλένδετο με την κειμενική ταυτότητά του και αποκτά αξία για ένα σύστημα περιγραφής κειμένων σε συνέργεια με τις παραμέτρους που περιγράφουν το περιβάλλον δημιουργίας τους, όχι ξεχωριστά από αυτές. Το επίπεδο αυτό πληροφορίας (θεματική κατηγοριοποίηση) μπορεί να ενοποιηθεί σε ενιαίο τμήμα υπό το στοιχείο <textDesc>, το οποίο ήδη παρέχεται από το πρότυπο στην ενότητα για τα σώματα κειμένων. Άλλωστε, ανάμεσα στη θεματική ενός κειμένου και στη γενικότερη περιγραφή του θα διακρίναμε σχέση υπωνυμίας και όχι σχέση ιδίου επιπέδου, όπως εισηγείται η ιεραρχία <profileDesc> <textClass> </textClass> <textDesc> </textDesc> </profileDesc>.

---

35 Για τεχνικές λεπτομέρειες βλ. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD43 (2.4.3 The Text Classification) και http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD55 (2.3.7 The Classification Declaration). Επίσης, http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-catDesc.html

36 Εναλλακτικά, για την κειμενική ταξινόμηση μπορούν να αξιοποιηθούν και οι δομές χαρακτηριστικών (*feature structures–FS*), οι οποίες δεν υιοθετούνται στα πρώτα επίπεδα ανάλυσης του ΣΕΝ και συνεπώς δεν απασχολούν την παρούσα μελέτη (βλ. ενότητα 3.2.1, υποσημείωση 25).

ειμενικής ταξο-
όητα, όπως το
ς βάσει λέξεων-
3) στην περιοχή
:α σχήματα κω-
I> (*classification*
description), το
ενο είτε προκύ-
στην καταστα-
ταξύ τους αλλά
ιναπαράστασης
αθμό σημειωτι-
ιξη, άλλοτε για
ιεί, ακριβώς, τη
ι διασφαλίζεται
αφικά πρότυπα

κοποίησης, στο
οντας το ίδιο το
η ανάγκη ενο-
επάρκειας: κατά
τοτελεί στοιχείο
για ένα σύστη-
ου περιγράφουν
ο επίπεδο αυτό
εί σε ενιαίο τμή-
ο πρότυπο στην
τική ενός κειμέ-
ωνυμίας και όχι
· <textClass> </

oc/tei-p5-doc/en/
r.tei-c.org/release/
laration). Επίσης,

θούν και οι δομές
:α πρώτα επίπεδα
βλ. ενότητα 3.2.1,

Β) Το τμήμα περιγραφής κειμένου <textDesc> προστέθηκε προσφάτως στο TEI, αποκλειστικά στην ενότητα για τα σώματα κειμένων και υιοθετεί την εξής οπτική: τα κείμενα περιγράφονται βάσει ενός συνόλου γνωρισμάτων, χωρίς να υπονοείται μία εκ των προτέρων κατάταξη σε κειμενικό είδος. Αντιθέτως, ο χαρακτηρισμός σε κειμενικό επίπεδο είναι συνισταμένη οχτώ διαφορετικών καταστασιακών παραμέτρων.[37]

Γ) Εφόσον η ανάπτυξη ενός συστήματος κατηγοριοποίησης κειμένων τέμνεται με τη θεωρητική τοποθέτηση του μελετητή για το τι συνιστά κειμενικό είδος (βλ. ενότητα 2.2) ακολουθήσαμε την πρόταση Χριστοφίδου (2012β, 2013 και υπό δημ.), κατά την οποία το κειμενικό είδος προκύπτει ως αποτέλεσμα συνδυασμού διαφορετικών παραμέτρων σε (Α) Επικοινωνιακό και (Β) Δομικό/Γνωσιακό επίπεδο (βλ. ενότητα 4). Σύμφωνα με τη δυνατότητα που προσφέρει το TEI διευρύναμε και τροποποιήσαμε τις καταστασιακές αυτές παραμέτρους.[38]

Δ) Συγκεκριμένα, αφαιρέθηκαν 9 από τα 14 στοιχεία της ενότητας <textDesc>, από τα 5 που διατηρήθηκαν όλα τροποποιήθηκαν (π.χ. εξειδικεύθηκαν με συγκεκριμένα ορίσματα, άλλαξαν σειρά στην προτεινόμενη ακολουθία του TEI, έγιναν επαναλαμβανόμενα), ακολούθως 3 νέα στοιχεία (θεματική, τρόπος λόγου, κειμενικός τύπος) δημιουργήθηκαν με δικά τους γνωρίσματα και τιμές για την περιγραφή του υλικού με διακριτά μεταδεδομένα. Τα νέα στοιχεία δηλώθηκαν σε ειδικό χώρο ονομάτων (*namespace*) ονόματι ndc (*NeoDemiaCorpus*) ως ndc:subject, ndc: discourseType, ndc:textType (για την πρόταση κωδικοποίησης βλ. Εικόνες 3, 11 και για την τροποποίηση του TEI βλ. Αφεντουλίδου 2016). Σχετικά με την τελική κατονομασία του κειμενικού είδους (σύνθετη απόφαση που προκύπτει από τον συνδυασμό και τη διεπίδραση των επιπέδων κατηγοριοποίη-

37 Ο όρος *περιγραφή κειμένου* <textDesc> χρησιμοποιείται καταχρηστικώς στο TEI, εφόσον στο επίπεδο αυτό ανάλυσης κωδικοποιείται η σχέση του εκάστοτε κειμένου με πλείστες *περικειμενικές* παραμέτρους. Για περαιτέρω επεξηγήσεις βλ. Module 15. *Language Corpora* (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/CC.html). Όπως επισημαίνεται: "The <textDesc> element [...] is provided as an alternative or a supplement to the common use of descriptive taxonomies used to categorize texts".

38 Στην περίπτωσή μας δηλαδή, το θεωρητικό πρότυπο που υιοθετήθηκε συνάδει στην αρχιτεκτονική του με το μοντέλο κωδικοποίησης των δεδομένων. Βεβαίως, και το ίδιο το TEI στηρίζεται σε θεωρητικές παραδοχές που προέκυψαν από την ενασχόληση των μελετητών με τα κείμενα, με επιρροές, όπως δηλώνεται (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/CC.html) από τους Hymes, Halliday, Crystal και Davy. Στο κείμενο των *Οδηγιών* του TEI επισημαίνεται η πολυφωνία και η ποικιλομορφία της κειμενικής ταξινόμησης και προτείνονται τα εξής: "Rather than attempting the task of proposing a single taxonomy of text-types (or the equally impossible one of enumerating all those which have been proposed previously), the closed set of situational parameters [..] can be used in combination to supply useful distinguishing descriptive features of individual texts, without insisting on a system of discrete high-level text-types" (ό.π.).

σης) αποφασίσαμε να κινηθούμε όσο το δυνατόν πιο κοντά στο ίδιο το πρότυπο: (α) αξιοποιήθηκε το ήδη υπάρχον στοιχείο <catDesc> στο τμήμα της περιγραφής κωδικοποίησης <encodingDesc>, (β) προστέθηκε η ιδιότητα @genre στο στοιχείο <textDesc> με τιμές δείκτες (*pointers*), οι οποίοι (γ) συνδέονται με την ταξονομία του κειμενικού είδους, όπως δηλώνεται στο (α) (βλ. Εικόνα 3).[39]

Ε) Το σχήμα που υλοποιήθηκε αναδεικνύει την εγγενή δυναμικότητα του TEI στην περιγραφή των κειμένων σε συνεχές πραγματώσεων σε συνάρτηση με τα περιβάλλοντα λόγου στα οποία παράγονται. Κατά τη γνώμη μας, η δυνατότητα – που παρέχεται αποκλειστικά στην ενότητα των σωμάτων κειμένων (*language corpora module*) – εύλογα μπορεί να γενικευθεί και να ενσωματωθεί στον ίδιο τον πυρήνα του TEI (*core module*), διαθέσιμη *εξαρχής* για όλα τα προς κωδικοποίηση κείμενα (π.χ. επιστολογραφία).[40] Τα πλεονεκτήματα, παραδόξως, της ενοποιητικής μας πρότασης τα εντοπίζει το ίδιο το TEI, χωρίς, ωστόσο, να εγκαταλείπει την κατάτμηση σε εξειδικευμένες ενότητες (*modules, modularity* λ.χ. για λεξικά,

39 Ιδανικά θα προτιμούσαμε όλα τα μεταδεδομένα για την κειμενική ταξινόμηση να αποθηκεύονται σε ένα τμήμα του XML, επιφορτισμένο με την περιγραφή της ταυτότητας του κειμένου (<profileDesc>) και όχι σε δύο ετερόκλητα (<encodingDesc> και <profileDesc>), με εισαγωγή νέου στοιχείου (*wrapper element*) για το κειμενικό είδος υπό το <textDesc> ή ακόμη και αντικαθιστώντας το. Ωστόσο, με τον τρόπο που ακολουθήσαμε (δημιουργία ανοιχτής ταξονομίας για το κειμενικό είδος στο τμήμα δηλώσεων κωδικοποίησης <encodingDesc> + προσθήκη ιδιότητας στο ήδη υπάρχον <textDesc> στο τμήμα περιγραφής της ταυτότητας κειμένου <profileDesc> με τιμές που συνδέονται απευθείας με την ανοιχτή ταξονομία) εξισορροπήσαμε την πρότασή μας με τα όσα ήδη ισχύουν στο TEI. Έτσι αποφεύχθηκαν μεγάλες δομικές αλλαγές αλλά και διαφοροποιήσεις στη σημασία και τη χρήση των ετικετών (*tags*) χαρακτηρισμού, πρακτική που θα δημιουργούσε απόπειρες κωδικοποίησης που μοιάζουν με το TEI χωρίς ουσιαστικά να είναι TEI (τον κίνδυνο επισημαίνουν και οι Vanhoutte & Van den Branden 2009: 84-85). Εξάλλου, με τον τρόπο αυτό διασφαλίσαμε τον εμπλουτισμό της ταξονομίας των κειμενικών ειδών από τους επισημειωτές του ΣΕΝ με αυστηρότητα και συνέπεια: κάθε νέο κειμενικό είδος που εντοπίζεται, δηλώνεται στο τμήμα κωδικοποίησης <encodingDesc> και δεν είναι ανάγκη να τροποποιηθεί το σχήμα του ΣΕΝ. Τέλος, η συγκεκριμένη χρήση δεικτών πληροί τις προϋποθέσεις της πρότασης των Haaf & Bański *'Attributes for lightweight linguistic annotation'*, βλ. υποσημείωση 25 και https://github.com/TEIC/TEI/issues/1670

40 Εντύπωση προκαλεί το γεγονός ότι στην τρίτη έκδοση του *British National Corpus* (BNC XML Edition 2007, βλ. http://www.natcorp.ox.ac.uk) συνυπάρχουν δύο διαφορετικές τυπολογίες κειμενικής ταξινόμησης, η αρχική της πρώτης έκδοσης BNC-1 (1994, βλ. http://www.natcorp.ox.ac.uk/docs/catRef.xml) και η μετέπειτα κατάταξη των κειμενικών ειδών του Lee (2001) για το BNC-2 ή BNC World (2001). Αυτό που ουσιαστικά προτείνουν οι κατατάξεις είναι ένα σύνολο από περιγραφικά χαρακτηριστικά που αποδίδονται στα κείμενα. Ως προς τον τρόπο υλοποίησης στο TEI ακολουθείται ένας από τους εναλλακτικούς δρόμους δήλωσης κατηγοριοποίησης που περιγράψαμε στο 3Α της ενότητας 3.2.2.2 (το BNC άλλωστε είναι μακρόπνοο έργο που προσαρμόστηκε κατά καιρούς σε διαφορετικά πρότυπα κωδικοποίησης), η ουσία είναι όμως ότι το κειμενικό είδος προσεγγίζεται πολυπαραγοντικά.

χειρόγραφα κείμενα, αναπαράσταση κριτικού υπομνήματος κ.ά.) στις οποίες, άλλωστε, ως πρότυπο φιλολογικής προέλευσης οφείλει και μέρος της δημοφιλίας του: «Μια τέτοια προσέγγιση έχει τα εξής συγκεκριμένα προτερήματα: ~ επιτρέπει έναν σχετικά συνεχή, δυναμικό χαρακτηρισμό των κειμένων (σε αντίθεση με διακριτές κατηγορίες βασισμένες σε τύπους ή θέματα) ~ επιτρέπει λογικές συγκρίσεις μεταξύ των σωμάτων κειμένων ~ επιτρέπει στους ερευνητές να κατασκευάζουν και να συγκρίνουν τους δικούς τους κειμενικούς τύπους βασισμένοι σε ειδικές παραμέτρους που τους αφορούν ~ είναι εξίσου εφαρμόσιμο σε προφορικά, γραπτά και νοηματικής γλώσσας κείμενα» (Module 15. Language Corpora).

4. Σύστημα κατηγοριοποίησης κειμένων δημοσιογραφικού λόγου

Όπως προαναφέρθηκε, για την περιγραφή της ταυτότητας των κειμένων δημιουργήθηκαν τα ειδικά στοιχεία *θεματική* και *κειμενικός τύπος, τρόπος λόγου* εκτός TEI. Στην ενότητα που ακολουθεί παρουσιάζουμε τις αποφάσεις μας σχετικά με τις τιμές των ορισμάτων τους. Η ανάπτυξη ταξονομιών κατηγοριοποίησης κειμένου είναι έργο απαιτητικό από μόνο του και ξεπερνά το θέμα της παρούσας ανακοίνωσης. Σε επισκόπηση ερευνητικών προγραμμάτων συναφών με το αντικείμενο της μελέτης μας (βλ. ενδεικτικά Hagen 2012, Gérard, Falk & Bernhard 2014, Kerremans 2015) διαπιστώσαμε ότι όλες οι κατηγοριοποιήσεις διαφέρουν μεταξύ τους, κάποιες θα ονομάζαμε ομαδοποιητικές, κάποιες αρκετά αναλυτικές (για παράδειγμα το *Logoscope* ανιχνεύει νεολογισμούς από ροές ειδήσεων του γαλλικού Τύπου και κατατάσσει τα άρθρα σε 68 διαφορετικά ειδικά θέματα και σε 16 γενικότερα, ενώ ο Hagen περιγράφει πώς για το *Norwegian Newspaper Corpus* ξεκίνησαν με 29 κατηγορίες και κατέληξαν σε 8· τέλος, για τα διαδικτυακά κείμενα που συλλέγονταν μέσω του *NeoCrawler* στο Μόναχο για την ανεύρεση νεολογισμών της Αγγλικής έχει αναπτυχθεί μία τυπολογία με 11 κατηγορίες θεμάτων και 10 κατηγορίες πηγών προέλευσης).

Ως προς τη θεματική ταξινόμηση των άρθρων σχεδιάσαμε τη δική μας κατηγοριοποίηση, λαμβάνοντας υπ' όψιν μας, όσο το δυνατόν, το σκεπτικό των εφημερίδων στην απόδοση θέματος (σύμφωνα με Χριστοφίδου & Αφεντουλίδου 2016), μετά από μακροχρόνια ανάλυση του ημερήσιου ελληνικού Τύπου. Επιπλέον, ως προς το δεύτερο επίπεδο κατάταξης, αυτό της κειμενικής ταξινόμησης, ακολουθώντας τη Χριστοφίδου (2012β, 2013, υπό δημ.), αναπτύξαμε σύνολα τιμών (*values*) για κάθε κείμενο (βλ. πιο κάτω). Συγκεκριμένα η πρόταση της Χριστοφίδου για την απόδοση *κειμενικού τύπου* καθ' όσον γνωρίζουμε, δεν έχει κωδικοποιηθεί ως μεταβλητή ανάλυσης σε διεθνή προγράμματα ανίχνευσης και καταγραφής νεολογικών σχηματισμών και ορολογίας. Κατά την υλοποίηση του σχήματος κατηγοριοποίησης, ακολουθήσαμε δύο διαδρομές, ανάλογα με τις τεχνολογικές δυνατότητες που είχαμε κάθε φορά στη διάθεσή μας.

Η πρώτη (*rule-based* κατά Hagen 2012: 112), ημι-αυτόματη διαδικασία κατάταξης στηρίχθηκε στην επαυξημένη πληροφορία, σημασιολογικού τύπου, που μπορεί να αποκομίσει κανείς αναλύοντας τα URLs (πρβλ. Halkidi, Nguyen, Varlamis & Vazirgiannis 2003, Varlamis, Vazirgiannis, Halkidi & Nguyen 2004, Tyrkkö 2010) των δημοσιογραφικών άρθρων. Όλες οι εφημερίδες διαθέτουν σταθερά συστήματα ονοματοδοσίας για την ταξινόμηση των άρθρων στις διάφορες στήλες τους π.χ. στην Πολιτική, στην Οικονομία, στα Άρθρα Γνώμης. Μελετήσαμε διεξοδικά το περίπλοκο δίκτυο των διευθύνσεων URL των εφημερίδων που συγκροτούν το ΣΕΝ της *Νεοδημίας* και καταρτίσαμε λεπτομερείς καταλόγους, τους οποίους στη συνέχεια ομαδοποιήσαμε ανά θεματική κατηγορία. Ακολούθως, με διαδικασίες εύρεσης-αντικατάστασης, κατατάξαμε a posteriori τα άρθρα που συλλέγονταν. Η πρώτη δοκιμή κατηγοριοποίησης υλοποιήθηκε πιλοτικά σε σώμα 116409 άρθρων από τη διαδικτυακή έκδοση του *Πρώτου Θέματος*, έκτασης 35752146 λέξεων. Η διαδικασία, ασφαλώς, αγγίζει το 100% στην ανάκληση (*recall*), διότι οι κανονιστικές εκφράσεις επιλέγουν με απόλυτη επιτυχία τα τμήματα των URLs που ζητούνται, ωστόσο, δεν είναι πάντοτε ακριβής (*precision*), διότι από ενδεικτικές περιπτώσεις που εξετάσαμε η ονοματοδοσία των URLs των άρθρων μπορεί να είναι εσφαλμένη (δηλαδή από την ίδια την εφημερίδα, ένα αθλητικό άρθρο να κατατάσσεται εκ παραδρομής στην Πολιτική κ.ά.).[41] Τέλος, η κατηγοριοποίηση βάσει URL εξαρτάται απολύτως από τον τρόπο που οι δημοσιογραφικοί οργανισμοί ταξινομούν τα κείμενά τους και ορίζουν τις θεματικές: για παράδειγμα, δεν έχουν όλες οι εφημερίδες εξειδικευμένες ενότητες για Υγεία και Περιβάλλον. Αρκετά προβληματική θεωρούμε, επίσης, την πάγια επιλογή των εφημερίδων να διακρίνουν ως ειδική θεματική την ενότητα *Άρθρο Γνώμης*, αν και πρόκειται για κειμενικό είδος, όπως η επιφυλλίδα, το ρεπορτάζ κ.ά. (βλ. Πίνακα 2 και 3) και όχι θεματική κατηγορία (βλ. και Πολίτης 2008β: 274). Η κατηγοριοποίηση των άρθρων βάσει διεύθυνσης σελίδας εφαρμόστηκε πιλοτικά στην πρώτη επέκταση του συστήματος συλλογής κειμένων της *Νεοδημίας* κατά το πρότυπο TEI P5 (βλ. Εικόνα 16 στο *Παράρτημα*).[42]

Η δεύτερη διαδρομή κατάταξης των κειμένων στηρίζεται σε στατιστικά μοντέλα μηχανικής μάθησης (*probabilistic, pattern-matching*) και αποτελεί έργο εν

41 Οι περιπτώσεις αυτές είναι μεμονωμένες, απρόβλεπτες (και πολλές φορές διορθώνονται από τις ίδιες τις εφημερίδες) και απαιτείται εξονυχιστική – ειδική – έρευνα για να μετρηθούν. Κατά τη γνώμη μας το πρόβλημα δεν έγκειται τόσο στο ότι εμφανίζονται όσο στο ότι δεν μπορούν να αποφευχθούν λόγω του τρόπου κατηγοριοποίησης (βάσει URL).

42 Στην έκδοση 5.0 του *NDCrawler* (βλ. υποσημείωση 56), η κατηγοριοποίηση βάσει ονόματος URL επιτελείται δηλαδή αυτόματα, ήδη από το στάδιο της εξόρυξης δεδομένων.

εξελίξει.[43] Έχει ήδη σχεδιαστεί μία καινούργια λειτουργία που ενσωματώθηκε πρόσφατα στο ηλεκτρονικό περιβάλλον της *Νεοδημίας* και κατηγοριοποιεί τα άρθρα ως εξής: εισάγονται αρχεία XML στην εφαρμογή και το σύστημα τα επιστρέφει επισημειωμένα με ετικέτες (*tags*) κατηγοριοποίησης. Η διαδικασία που υποστηρίζεται τώρα είναι να δίνεται κάποιο σύνολο εκπαίδευσης και ο αλγόριθμος να εκπαιδεύεται σε αυτές τις κατηγορίες. Κάθε νέο κείμενο το τοποθετεί σε τουλάχιστον μία κατηγορία, δίνοντας παράλληλα τη δυνατότητα διόρθωσης ή έγκρισης. Ό,τι διορθώνεται ή εγκρίνεται, ενσωματώνεται κι αυτό στο σύνολο εκπαίδευσης (ημι-αυτόματη διαδικασία), έχοντας άμεσο αντίκτυπο στην απόδοση του μοντέλου. Υπάρχει η δυνατότητα να εφαρμοστεί το μοντέλο χωρίς επίβλεψη (αυτόματη διαδικασία) σε νέα κείμενα, αλλά χρειάζεται να έχει προηγηθεί επιτυχώς η αξιολόγηση της ημι-αυτόματης διαδικασίας. Για τη δεύτερη αυτή διαδρομή κατηγοριοποίησης προτείναμε τις εξής κατηγορίες ανάλυσης, οι οποίες θα κωδικοποιούνται στο ΣΕΝ σύμφωνα με Χριστοφίδου & Αφεντουλίδου (2016) για τη Θεματική Ταξινόμηση και με Χριστοφίδου (2012β, 2013 και υπό δημ.) για την Κειμενική Ταξινόμηση:

ΘΕΜΑΤΙΚΗ ΤΑΞΙΝΟΜΗΣΗ (για τον δημοσιογραφικό λόγο)

ΘΕΜΑΤΙΚΗ	ΕΠΕΞΗΓΗΣΗ (επιμέρους θέματα)
Ελλάδα: Πολιτική	αμιγώς πολιτικά θέματα
Ελλάδα: Κοινωνία	αμιγώς κοινωνικά θέματα, επίσης θέματα που σχετίζονται με τομείς όπως εργασία, παιδεία, δικαιοσύνη, θρησκεία, άλλους θεσμούς (σύστημα υγείας, περιβάλλον κ.ά.)
Ελλάδα: Οικονομία	αμιγώς οικονομικά θέματα
Διεθνή: Πολιτική	αντίστοιχα θέματα επικαιρότητας στο εξωτερικό της χώρας (πολιτικές, διεθνείς σχέσεις, διπλωματία, εξοπλισμοί κ.ά.)
Διεθνή: Κοινωνία	αντίστοιχα θέματα επικαιρότητας στο εξωτερικό της χώρας (ευρύτερα κοινωνικά ζητήματα, εξεγέρσεις, πραξικοπήματα, τρομοκρατία σε σχέση με την κοινωνία κ.ά.)
Διεθνή: Οικονομία	αντίστοιχα θέματα επικαιρότητας στο εξωτερικό της χώρας
Επιστήμη - Τεχνολογία	φυσικές και ανθρωπιστικές επιστήμες, διάστημα, υγεία, πληροφορική, διαδίκτυο και γενικά εφαρμογές, περιβάλλον κ.ά. υπό την οπτική γωνία της επιστήμης και της τεχνολογίας

43 Υλοποιείται σε συνεργασία με το *Τμήμα Πληροφορικής & Τηλεματικής, Χαροκόπειο Πανεπιστήμιο*: Διπλωματική εργασία Τηλέμαχου Πετσόπουλου. Τα αποτελέσματα θα συμπεριληφθούν σε ειδική μελέτη, όταν ολοκληρωθεί η εκπαίδευση του συστήματος και αξιολογηθεί η απόδοσή του. Επιβλέπων: καθηγητής Ηρακλής Βαρλάμης.

Πολιτισμός	τέχνες και γράμματα, θεάματα κ.ά.
Αθλητισμός	όλα τα αθλήματα (περιλαμβάνει και τον μηχανοκίνητο αθλητισμό)
Βιοτροπία: Δραστηριότητες Ελεύθερου Χρόνου	ταξίδια, κυνήγι, ψάρεμα κ.ά.
Βιοτροπία: Διατροφή – Σώμα	ό,τι σχετίζεται με το φαίνεσθαι, επιλογές ζωής, «λάιφ στάιλ» (π.χ. διατροφικές συνήθειες, δίαιτα, γαστρονομία, μαγειρική, ευεξία, υγεία μέσα από περιποίηση σώματος, άθληση)
Βιοτροπία: Κοινωνικά – Κοσμικά – Τηλεοπτικά Νέα	κοσμικά νέα, σχέσεις
Κατανάλωση: Σπίτι – Οικογένεια – Εμφάνιση	ό,τι σχετίζεται με καταναλωτικές συνήθειες του ανθρώπου που έχουν να κάνουν συνήθως με την εμφανισιακή του εικόνα (π.χ. ένδυση, μόδα, καλλωπισμός, διακόσμηση, γενικώς θέματα shopping)
Κατανάλωση: Εμπορική Τεχνολογία	έμφαση στην καταναλωτική πλευρά της τεχνολογίας (π.χ. gadgets, gaming, γενικώς παρουσίαση εφαρμογών με αποδέκτες το αγοραστικό κοινό)
Κατανάλωση: Αυτοκίνηση	η καταναλωτική, υλιστική πλευρά της αυτοκίνησης
Ξενόγλωσσα	αγγλικά κείμενα κυρίως
Άλλο	ωροσκόπια, προγνώσεις καιρού κ.ά.

ΚΕΙΜΕΝΙΚΗ ΤΑΞΙΝΟΜΗΣΗ (για τον δημοσιογραφικό λόγο)

ΚΕΙΜΕΝΙΚΟΣ ΤΥΠΟΣ	ΚΕΙΜΕΝΙΚΟ ΕΙΔΟΣ
Δημοσιογραφικό Άρθρο	άρθρο γνώμης / επιφυλλίδα / editorial / ρεπορτάζ / ενημερωτικό άρθρο κ.λπ.
Κριτική	κριτική βιβλίου, κριτική μουσείου / έκθεσης, κριτική θεάτρου κ.λπ.
Επιστολή	ανοιχτή επιστολή συντελεστή / επιστολή αναγνώστη
Συνέντευξη	συνέντευξη προσωπική / συνέντευξη πολιτική
Συνταγή	συνταγή μαγειρικής / συνταγή καλλυντικών κ.ά.
Άλλο	άλλο

Πίνακας 2. Πρόταση για διττή εκπαίδευση αυτόματου συστήματος κατηγοριοποίησης

Συγκεκριμένα η Χριστοφίδου (ό.π.) έχει προτείνει μία πολυεπίπεδη προσέγγιση για την οργάνωση και ταξινόμηση των κειμενικών ειδών, την οποία και υιοθετούμε. Ορμώμενη από τη διττή λειτουργία της γλώσσας, δηλαδή την επικοινωνιακή και τη γνωσιακή (Dressler 1987) προτείνει κατ' αρχήν δύο βασικά

επίπεδα ανάλυσης, το *Επικοινωνιακό* και το *Γνωσιακό* (εμπεριέχει και το δομικό). Στο πρώτο επίπεδο (α) εντάσσει τους παράγοντες που αφορούν Ι. τον *Λειτουργικό*, ΙΙ. τον *Καταστασιακό*, ΙΙΙ. τον *Θεματικό* τομέα (πρβλ. και Heinemann & Viehweger 1991) και στο δεύτερο επίπεδο (β) τους παράγοντες που αφορούν Ι. τους *τρόπους λόγου* (προτείνεται η οικονομική διάκριση τριών βασικών τρόπων, δηλαδή του αφηγηματικού / περιγραφικού / επιχειρηματολογικού λόγου) ΙΙ. τη διαρθρωτική οργάνωση, δηλαδή τον *κειμενικό τύπο* και ΙΙΙ. τη *λεξικογραμματική* οργάνωση (βλ. προτεινόμενο σχήμα στον Πίνακα 3).

ΚΕΙΜΕΝΙΚΟ ΕΙΔΟΣ			
Α. Επικοινωνιακό Επίπεδο	Ι. Λειτουργικός Τομέας (σκοπός)		*εκφράζομαι, χειραγωγώ / προωθώ, πληροφορώ, διδάσκω κ.λπ.*
	ΙΙ. Καταστασιακός Τομέας	*α. πολιτισμικό-κοινωνικό πλαίσιο*	α. *Περιοχές δραστηριοτήτων:* Επιστήμης, Λογοτεχνίας, Διοίκησης, Τύπου, Διαφήμισης, Καθημερινής ζωής κ.λπ.
		β. περίσταση επικοινωνίας	β. *Δίαυλοι / Ρόλοι:* γραπτό / προφορικό / ηλεκτρονικό, κοινωνική σχέση πομπού / δέκτη, χωρόχρονο παραγωγής / πρόσληψης κ.ά.
	ΙΙΙ. Θεματικός Τομέας		*γλώσσα, αθλητισμός, διασκέδαση κ.λπ.*
Β. Δομικό/ Γνωσιακό Επίπεδο	Ι. Οργάνωση: Τρόποι Λόγου (ή Είδη Λόγου)		*αφήγηση περιγραφή επιχειρηματολογία*
	ΙΙ. Οργάνωση: Κειμενικός Τύπος (πρότυπο) *αρχή / κυρίως μέρος (διάρθρωση) / τέλος* (μόνο γραμμική οργάνωση κειμένου)		*τύπος επιστολής επιστημονικού άρθρου, εγχειρίδιου, συνταγής, συνέντευξης κ.λπ.*
	ΙΙΙ. Οργάνωση: Λεξικογραμματική		*γραμματική / σύνταξη / λεξιλόγιο (καθημερινό, επιστημονικό κ.λπ.)*

Πίνακας 3. Πολυεπίπεδη παρουσίαση των παραγόντων που συνδιαμορφώνουν τα εκάστοτε κειμενικά είδη ενός πολιτισμού

Συγκεκριμένα, οι παραδοσιακές γενικές φιλολογικές κατηγορίες, όπως *επιστολή, συνέντευξη, άρθρο, συνταγή* κ.λπ. εντάσσονται στο παραπάνω σχήμα ως *κειμενικός τύπος*, αφού πρόκειται για περιπτώσεις που ακολουθούν μια συγκεκριμένη κειμενική οργάνωση, συνήθως με προτυπική, δομική κυρίως, διάρθρωση. Παραδείγματος χάριν η *επιστολή* έχει προτυπικά αποστολέα, προσφώνηση κ.λπ., η *συνέντευξη* είναι προτυπικά διαλογική συζήτηση με ερωτήματα και απαντήσεις, το *άρθρο* έχει συνήθως μια συνοπτική μονολογική αποτύπωση γνώμης ή γεγονότων με αναγραφόμενη την πηγή ή τον συγγραφέα, η *συνταγή* περιγράφει με σειρά τα συστατικά και τη διαδικασία παρασκευής κ.ο.κ.

Στις νεώτερες κειμενογλωσσολογικές προσεγγίσεις, όμως, μία πιο πολυδιάστατη και πιο συνεπής κατηγοριοποίηση των κειμένων φαίνεται απαραίτητη. Έτσι σύμφωνα με το σχήμα προτείνεται το εξής: διατηρώντας τις παραδοσιακές γενικές «ετικέτες» ως *κειμενικούς τύπους* (διαρθρωτική δομή) να αποδοθεί η πολυεπίπεδη δυναμική – που προτείνεται από τις νεώτερες θεωρήσεις (Bax 2011, βλ. και Swales 1990, Bhatia 2004)[44] – στην έννοια του *κειμενικού είδους*, το οποίο προκύπτει από την ευέλικτη διεπίδραση όλων των τομέων οι οποίοι συμβάλλουν στην τελική διαμόρφωση κάθε κειμένου, τόσο των επικοινωνιακών όσο και των γνωσιακών (βλ. Πίνακα 3). Σύμφωνα με την πρόταση της Χριστοφίδου (ό.π.), λοιπόν, το εκάστοτε *κειμενικό είδος* προκύπτει με δυναμικό τρόπο από πολλαπλούς παράγοντες και αποτελεί ένα σχήμα σχετικά ευέλικτο και ανοιχτό.

Ακολούθως εφαρμόζουμε το σχήμα στον δημοσιογραφικό λόγο:[45]

Α. Το πρώτο *Επικοινωνιακό Επίπεδο* είναι εν μέρει περιγεγραμμένο, αφού αναφερόμαστε σε κείμενα εφημερίδων, οπότε ως προς τον *Λειτουργικό Τομέα* ισχύουν τα *χειραγωγώ, προωθώ* ή/και *πληροφορώ*. Ως προς τον *Καταστασιακό Τομέα* ισχύει η *Περιοχή (δραστηριοτήτων) του Τύπου* με *ηλεκτρονικό δίαυλο* (στην περίπτωσή μας). Ο *Θεματικός Τομέας* ποικίλλει, όπως φαίνεται πιο πάνω.

Β. Ως προς το δεύτερο *Γνωσιακό Επίπεδο* οι επιλογές εναλλάσσονται περισσότερο και έτσι προκύπτουν διαφορετικά *κειμενικά είδη*: στον ίδιο *κειμενικό τύπο* του *δημοσιογραφικού άρθρου* ένα *άρθρο γνώμης* επιλέγει προφανώς τον επιχειρηματολογικό τρόπο λόγου, ενώ ένα *ρεπορτάζ* κυρίως την αφήγηση, μια *κριτική έκθεσης* συνδυάζει περιγραφή και επιχειρηματολογία. Αντίστοιχα εναλ-

44 Ειδικότερα ο Bhatia (2004) ασχολήθηκε με την ομαδοποίηση και υποκατηγοριοποίηση των κειμενικών ειδών και καθιέρωσε την έννοια της αποικίας κειμένων. Στο Χριστοφίδου (2012β, 2013 και υπό δημ.) αναλύεται η προβληματική της υπερ- και υποκατηγοριοποίησης κειμένων σε μια αρκετά διαφοροποιημένη πρόταση από αυτή του Bhatia.

45 Βλ. μεταξύ άλλων και Lenk & Chesterman (2005).

λάσσονται και η λεξικογραμματική οργάνωση, αφού το *άρθρο γνώμης* έχει συχνά διαφορετική γραμματική, σύνταξη και ύφος από ένα *ρεπορτάζ* ή από μια *κριτική*.[46] Διαπιστώνουμε ότι τα δύο επίπεδα διαπλέκονται δυναμικά, αφού τα προαναφερθέντα διακριτά *κειμενικά είδη* διαφέρουν και ως προς τον *Λειτουργικό Τομέα*: τα *άρθρα γνώμης χειραγωγούν* ενώ *ενημερωτικά άρθρα πληροφορούν*. Ο δε *Θεματικός Τομέας* φαίνεται να καθορίζει αποφασιστικά το λεξιλόγιο.

Αντίστοιχα, μια *επιστολή διαμαρτυρίας* αναγνώστη διαφέρει ως προς το *κειμενικό είδος* από την *ανοιχτή επιστολή* που συχνά δημοσιεύουν επίσημα σε μια εφημερίδα οι συντελεστές της ή άλλα σημαίνοντα πρόσωπα και συλλογικότητες με δημόσιο λόγο και κοινό ενδιαφέρον, τόσο ως προς την περίσταση επικοινωνίας, όσο και ως προς τους *τρόπους λόγου* αλλά και τη *λεξικογραμματική οργάνωση*. Παρ' όλ' αυτά έχουν και τα δύο κείμενα τον ίδιο βασικό *κειμενικό τύπο* της επιστολής, όπως άλλωστε και οι προσωπικές / φιλικές επιστολές ή οι συστατικές επιστολές, που αποτελούν διαφορετικό κειμενικό είδος αντίστοιχα, αφού άλλωστε ανήκουν σε διαφορετικές *Περιοχές δραστηριοτήτων* από αυτήν του *Τύπου*, όπως *Καθημερινής ζωής* κ.λπ.

5. *Εφαρμογή του μοντέλου*

Μετά την ανωτέρω διεξοδική συζήτηση η μορφή του κωδικοποιημένου κειμένου σύμφωνα με το μοντέλο εγγράφου που παρουσιάστηκε πιο πάνω διαμορφώνεται ως εξής, βάσει των ενδεικτικών Εικόνων 3-7, όπως παρατίθενται αμέσως πιο κάτω:

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model href="NeoDemiaCorpus.rnc" type="application/relax-ng-compact-syntax"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0"
xmlns:ndc="http://www.academyofathens.gr/kecon/NeoDemiaCorpus/ns/1.0">
<!--ΜΕΤΑΔΕΔΟΜΕΝΑ -->
   <teiHeader>
      <fileDesc>
         <titleStmt>
               <title xml:id="Kathimerini_2015-11-14_150.xml"/>
               <respStmt>
                  <resp>Collected with RSS/web crawling by the</resp>
                  <orgName>Research Centre for Scientific Terms and Neologisms-Academy of
                     Athens</orgName>
               </respStmt>
         </titleStmt>
            <publicationStmt>
               <authority
                  resp="Research Centre for Scientific Terms and Neologisms-Academy of Athens"/>
```

46 Η λεξικογραμματική οργάνωση εμπεριέχει υφομετρικές μεταβλητές βαρύνουσας σημασίας για διακρίσεις ανάμεσα στα κειμενικά είδη, οι οποίες έχουν χρησιμοποιηθεί από αρκετά νωρίς σε συστήματα αυτόματης κατηγοριοποίησης κειμένου, βλ. Tambouratzis, Markantonatou, Hairetakis & Carayannis (2000).

```
                <availability status="restricted">
                    <p>Available to the Research Centre for Scientific Terms and Neologisms for academic research
                        purposes only</p>
                    <licence target="http://creativecommons.org/licenses/by-nc-nd/4.0/">
                        <p>This work is licensed under a Creative Commons
                            Attribution-NonCommercial-NoDerivatives 4.0 International License.</p>
                    </licence>
                </availability>
            </publicationStmt>
            <sourceDesc>
                <bibl>
                    <author who="ΓΙΑΝΝΗΣ ΠΑΛΑΙΟΛΟΓΟΣ"/>
                    <publisher who="Η ΚΑΘΗΜΕΡΙΝΗ"/>
                    <date when="2015-11-14" dayinweek="Sat" quarter="fourth"/>
                    <time when="14:18:49"/>
                    <ref target="http://www.kathimerini.gr/838612/article/epikairothta/ellada/en-dynamei-
                        oikogeneies-sthn-anamonh-xwris-kinhtra"/>
                    <edition type="Online and printed"/>
                </bibl>
            </sourceDesc>
        </fileDesc>
        <encodingDesc>
            <projectDesc>
                <p> NeoDemiaCorpus [A monitor corpus of newspaper articles collected from the web
                    for identifying neologisms. Corpus and research design: Anastasia Christofidou,
                    Vassiliki Afentoulidou. Crawling technology developed by Iraklis Varlamis.] </p>
            </projectDesc>
            <classDecl>
                <taxonomy xml:id="Κειμενικό_Είδος"> <!-- Ταξονομία -->
                    <category xml:id="Αρ_Εν "> <!-- Κειμενικό Είδος -->
                    <catDesc>Άρθρο Ενημερωτικό</catDesc> <!-- Επεξήγηση -->
                    </category>
                    <category xml:id="Αρ_Γν"> <!-- Κειμενικό Είδος -->
                    <catDesc>Άρθρο Γνώμης</catDesc> <!-- Επεξήγηση -->
                    </category>
                    [...]
                </taxonomy>
            </classDecl>
        </encodingDesc>
        <profileDesc>
            <textDesc genre="#Αρ_Εν"> <!—Σύνδεση με το Κειμενικό Είδος -->
            <!-- ΕΠΙΚΟΙΝΩΝΙΑΚΟ ΕΠΙΠΕΔΟ -->
            <purpose type="inform" degree="primary"/> <!-- Λειτουργικός Τομέας: Σκοπός -->
            <purpose type="persuade" degree="secondary"/>
            <domain type="Press"/> <!-- Καταστασιακός Τομέας (πολιτισμικό-κοινωνικό πλαίσιο): Περιοχές δραστηριοτήτων -->
            <channel mode="w" type="eNewspaper"/> <!-- Καταστασιακός Τομέας (περίσταση επικοινωνίας): Δίαυλοι/Ρόλοι -->
            <interaction type="none"/>
            <ndc:subject key="Ελλάδα:Κοινωνία"/> <!-- Θεματικός Τομέας -->
            <!-- ΔΟΜΙΚΟ - ΓΝΩΣΙΑΚΟ ΕΠΙΠΕΔΟ -->
            <ndc:discourseMode type="description" degree="primary"/> <!-- Οργάνωση: Τρόποι Λόγου ή Είδη Λόγου -->
            <ndc:discourseMode type="argumentation" degree="secondary"/>
            <ndc:textType class="Δημοσιογραφικό Άρθρο"/> <!-- Οργάνωση: Κειμενικός Τύπος -->
            </textDesc>
        </profileDesc>
        <revisionDesc status="crawled">
            <change/>
        <revisionDesc>
    </teiHeader>
<!-- ΔΕΔΟΜΕΝΑ -->
    <text> <!— Οργάνωση: Λεξικο-γραμματική -->
        <body>
            <div type="newspaper_text">
                <div type="headings">
                    <head type="surTitle"/>
                    <head type="mainTitle">Εν δυνάμει οικογένειες στην αναμονή, χωρίς κίνητρα</head>
                    <head type="subTitle"/>
                    <head type="lead"/>
```

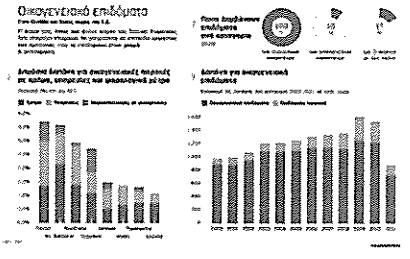
```
                    <caption/>
                </div>
                <div type="mainText">
                    <p>Καταβλήθηκαν τελικά προ ημερών, με καθυστέρηση εβδομάδων, τα επιδόματα τέκνου γ' τριμήνου σε 643.290 δικαιούχους. Στο διάστημα που προηγήθηκε έγινε πολύς λόγος για τα νέα συμπτώματα ταμειακής αδυναμίας του κράτους. Αγνοήθηκε (όπως συνηθίζεται) το ευρύτερο ζήτημα της ανεπαρκούς στήριξης που παρέχει η Πολιτεία σε άτομα που θέλουν να κάνουν οικογένεια. </p>
                    <p>Οι συνέπειες της ανεπάρκειας αυτής είναι πολυσχιδείς: από τη φτωχοποίηση μη πολύτεκνων νοικοκυριών και τα εμπόδια που πρέπει να ξεπεράσει μια γυναίκα -ιδιαίτερα στον ιδιωτικό τομέα- για να συνδυάσει τη μητρότητα με την εργασία έως την αναβολή (συχνά επ' αορίστου) της τεκνοποίησης από ζευγάρια που δεν βλέπουν πώς θα μπορέσουν να παράσχουν τα απαραίτητα για τα παιδιά τους όταν δυσκολεύονται να τα εξασφαλίσουν για τους ίδιους.</p>
                    <p>Το διαχρονικό ζήτημα των οικογενειακών επιδομάτων συνδέεται, ήδη πριν από την κρίση, με τις επιλεκτικές ευαισθησίες του κράτους πρόνοιας ελληνικής κοπής.</p>
                    <p>Όπως σημείωνε ο Μάνος Ματσαγγάνης του Οικονομικού Πανεπιστημίου Αθηνών προ διετίας («Η πολιτική κατά της φτώχειας στην Ελλάδα της κρίσης», Ενημερωτικό Δελτίο Ομάδας Ανάλυσης Δημόσιας Πολιτικής), οι μόνες κατηγορίες που λάμβαναν αξιόλογες οικογενειακές παροχές ήταν οι πολύτεκνοι (νοικοκυριά με τρία ή περισσότερα παιδιά), οι υπάλληλοι των ΔΕΚΟ, των τραπεζών και -ώς έναν βαθμό- του Δημοσίου.</p>
                    <p>Αντίθετα, η μεγάλη πλειονότητα, τα νοικοκυριά των εργαζομένων του ιδιωτικού τομέα με ένα ή δύο παιδιά και χαμηλά εισοδήματα, λάμβανε πενιχρά επιδόματα (8 ευρώ τον μήνα για ένα παιδί, 25 ευρώ για δύο) ή δεν λάμβανε τίποτα.</p>
                </div>
                <div type="headings">
                    <head type="midTitle">Ευρωπαϊκή πρακτική</head>
                </div>
                <div type="mainText">
                    <p>Με τον νόμο 4093/2012 εισήχθη το ενιαίο επίδομα τέκνων, το ύψος του οποίου κυμαίνεται από 13 έως 40 ευρώ τον μήνα ανά παιδί, ανάλογα με το εισόδημα της οικογένειας και τον αριθμό των παιδιών. Το επίδομα αυτό, ακολουθώντας την πάγια ευρωπαϊκή πρακτική, ισχύει από το πρώτο παιδί, διορθώνοντας κάποιες από τις αδικίες του προηγούμενου καθεστώτος.</p>
                    <p>Παράλληλα θεσπίστηκε νέο ειδικό πολυτεκνικό επίδομα, ύψους 500 ευρώ ανά παιδί πέραν των δύο πρώτων ετησίως. Το εισοδηματικό κριτήριο σε αυτή την περίπτωση είναι 45.000 ευρώ (για τρίτεκνες οικογένειες) και προσαυξάνεται κατά 3.000 ευρώ για το τέταρτο παιδί και κατά 4.000 ευρώ για κάθε παιδί πέραν των τεσσάρων. Καταργήθηκε ωστόσο επί πρωθυπουργίας Αντώνη Σαμαρά σειρά από άλλες παροχές προς τους πολύτεκνους, από την ισόβια σύνταξη της μητέρας ώς το αυξημένο επίπεδο αφορολογήτου.</p>
                    <p>Σημειώνεται, χάριν σύγκρισης, ότι την ίδια χρονιά που θεσπίστηκε το ενιαίο επίδομα τέκνων, η Συντηρητική κυβέρνηση στη Βρετανία επέβαλε κι αυτή εισοδηματικά κριτήρια στο αντίστοιχο επίδομα στο Ηνωμένο Βασίλειο. Με βάση τους νέους όρους, το δικαίωμα στο επίδομα χάνεται όταν το ετήσιο εισόδημα ενός εκ των δύο γονέων υπερβεί τις 60.000 λίρες (84.000 ευρώ).</p>
                    <p>Για τους δικαιούχους το επίδομα είναι 20,50 λίρες (28,70 ευρώ) την εβδομάδα για το πρώτο παιδί και 13,55 λίρες την εβδομάδα για καθένα από τα άλλα παιδιά.</p>
                </div>
                <div type="headings">
                    <head type="midTitle">Η ανισότητα επιμένει</head>
                </div>
                <div type="mainText">
                    <p>Εν τω μεταξύ, η χώρα μας -όπως και άλλες χώρες της Νότιας Ευρώπης- χαρακτηρίζεται από μια πατριαρχική αντίληψη σχετικά με τον ρόλο της γυναίκας στην οικογένεια, που αντικατοπτρίζεται σε σειρά διατάξεων της εργασιακής νομοθεσίας και (κυρίως) πρακτικής και στα κενά των υπηρεσιών φροντίδας παιδιών.</p>
                    <p>
                    [...]
                    </p>
                </div>
            </div>
        </body>
    </text>
</TEI>
```

Εικόνα 3. Δείγμα κωδικοποίησης (TEI P5) του κειμένου-πηγή (ακολουθεί)

ΕΛΛΑΔΑ

Εν δυνάμει οικογένειες στην αναμονή, χωρίς κίνητρα

ΠΑΝΑΓΗΣ ΠΑΠΑΔΟΠΟΥΛΟΣ



ΔΙΑΒΑΣΤΕ ΕΠΙΣΗΣ

180 χρόνια Πανεπιστήμιο Αθηνών

ΑΓΓΕΛΟΣ ΤΣΟΥΚΑΛΑΣ

Καλύτερα να εφαρμοζόταν το «e-mail Χαρδούβελη»

Τέλος στην πολυετή ασυλία των τραπεζών στις επιχειρήσεις «ζόμπι»

ΑΝΔΡΕΑΣ ΡΗΓΟΠΟΥΛΟΣ

Αίγυπτος: Ενός λεπτού σιγή στη συνέλευση της ελληνικής κοινότητας Κοΐρου

Καταβλήθηκαν τελικά προ ημερών, με καθυστέρηση εβδομάδων, τα επιδόματα τέκνου γ' τριμήνου σε 643.290 δικαιούχους. Στο διάστημα που προηγήθηκε έγινε πολύς λόγος για τα νέα συμπτώματα ταμειακής αδυναμίας του κράτους. Αγνοήθηκε (όπως συνηθίζεται) το ευρύτερο ζήτημα της ανεπαρκούς στήριξης που παρέχει η Πολιτεία σε άτομα που θέλουν να κάνουν οικογένεια.

Οι συνέπειες της ανεπάρκειας αυτής είναι πολυσχιδείς: από τη φτωχοποίηση μη πολύτεκνων νοικοκυριών και τα εμπόδια που πρέπει να ξεπεράσει μια γυναίκα -ιδιαίτερα στον ιδιωτικό τομέα- για να συνδυάσει τη μητρότητα με την εργασία έως την αναβολή (συχνά επ' αορίστου) της τεκνοποίησης από ζευγάρια που δεν βλέπουν πώς θα μπορέσουν να παράσχουν τα απαραίτητα για τα παιδιά τους όταν δυσκολεύονται να τα εξασφαλίσουν για τους ίδιους.

Το διαχρονικό ζήτημα των οικογενειακών επιδομάτων συνδέεται, ήδη πριν από την κρίση, με τις επιλεκτικές ευαισθησίες του κράτους πρόνοιας ελληνικής κοπής.

Οπως σημείωνε ο Μάνος Ματσαγγάνης του Οικονομικού Πανεπιστημίου Αθηνών προ διετίας («Η πολιτική κατά της φτώχειας στην Ελλάδα της κρίσης», Ενημερωτικό Δελτίο Ομάδας Ανάλυσης Δημόσιας Πολιτικής), οι μόνες κατηγορίες που λάμβαναν αξιόλογες οικογενειακές παροχές ήταν οι πολύτεκνοι (νοικοκυριά με τρία ή περισσότερα παιδιά), οι υπάλληλοι των ΔΕΚΟ, των τραπεζών και -ως έναν βαθμό- του Δημοσίου.

Αντίθετα, η μεγάλη πλειονότητα, τα νοικοκυριά των εργαζομένων του ιδιωτικού τομέα με ένα ή δύο παιδιά και χαμηλά εισοδήματα, λάμβανε πενιχρά επιδόματα (8 ευρώ τον μήνα για ένα παιδί, 25 ευρώ για δύο) ή δεν λάμβανε τίποτα.

Ευρωπαϊκή πρακτική

Με τον νόμο 4093/2012 εισήχθη το ενιαίο επίδομα τέκνων, το ύψος του οποίου κυμαίνεται από 13 έως 40 ευρώ τον μήνα ανά παιδί, ανάλογα με το εισόδημα της οικογένειας και τον αριθμό των παιδιών. Το επίδομα αυτό, ακολουθώντας την πάγια ευρωπαϊκή πρακτική, ισχύει από το πρώτο παιδί, διορθώνοντας κάποιες από τις αδικίες του προηγούμενου καθεστώτος.

Παράλληλα θεσπίστηκε νέο ειδικό πολυτεκνικό επίδομα, ύψους 500 ευρώ ανά παιδί πέραν των δύο πρώτων ετησίως. Το εισοδηματικό κριτήριο σε αυτή την περίπτωση είναι 45.000 ευρώ (για τρίτεκνες οικογένειες) και προσαυξάνεται κατά 3.000 ευρώ για το τέταρτο παιδί και κατά 4.000 ευρώ για κάθε παιδί πέραν των τεσσάρων. Καταργήθηκε ωστόσο επί πρωθυπουργίας Αντώνη Σαμαρά σειρά από άλλες παροχές προς τους πολύτεκνους, από την ισόβια σύνταξη της μητέρας ως το αυξημένο επίπεδο αφορολογήτου.

Σημειώνεται, χάριν σύγκρισης, ότι την ίδια χρονιά που θεσπίστηκε το ενιαίο επίδομα τέκνων, η Συντηρητική κυβέρνηση στη Βρετανία επέβαλε κι αυτή εισοδηματικά κριτήρια στο αντίστοιχο επίδομα στο Ηνωμένο Βασίλειο. Με βάση τους νέους όρους, το δικαίωμα στο επίδομα χάνεται όταν το ετήσιο εισόδημα ενός εκ των δύο γονέων υπερβεί τις 60.000 λίρες (84.000 ευρώ)

Για τους δικαιούχους το επίδομα είναι 20,50 λίρες (28,70 ευρώ) την εβδομάδα για το πρώτο παιδί και 13,55 λίρες την εβδομάδα για καθένα από τα άλλα παιδιά.

Η ανισότητα επιμένει

Εν τω μεταξύ, η χώρα μας -όπως και άλλες χώρες της Νότιας Ευρώπης- χαρακτηρίζεται από μια πατριαρχική αντίληψη σχετικά με τον ρόλο της γυναίκας στην οικογένεια, που αντικατοπτρίζεται σε σειρά διατάξεων της εργασιακής νομοθεσίας και (κυρίως) πρακτικής και στα κενά των υπηρεσιών φροντίδας παιδιών.

Οπως σημειώνει ο Συνήγορος του Πολίτη στην τελευταία ετήσια έκθεσή του, «στον ιδιωτικό τομέα, η μητρότητα φαίνεται να αποτελεί για τους εργοδότες βάρος από το οποίο προέχει να απαλλαγούν, και όχι πεδίο προστασίας.

Καταγγελίες συμβάσεων εργασίας, εξώθηση σε παραίτηση μέσω δυσμενών μεταβολών των συνθηκών εργασίας εργαζομένων στη διάρκεια της περιόδου προστασίας λόγω εγκυμοσύνης και μητρότητας, υποχρεωτική συνταξιοδότηση με χρήση ευεργετικών, κατ' αρχήν, διατάξεων με σκοπό την απομάκρυνση γυναικών από το εργασιακό περιβάλλον χωρίς τη συναίνεσή τους αποτελούν χαρακτηριστικές συμπεριφορές ιδιωτών εργοδοτών».

Παραλληλα, σύμφωνα με στοιχεία της Eurostat, η Ελλάδα βρίσκεται μακριά από τους ευρωπαϊκούς στόχους και σημαντικά κάτω από τον μέσο όρο της Ευρωπαϊκής Ενωσης στο ποσοστό των παιδιών (τόσο ώς την ηλικία των τριών όσο και από τα τρία ώς την Α΄ Δημοτικού) σε οργανωμένα ιδρύματα παιδικής φροντίδας.

Δεν είναι τυχαίο ότι το ποσοστό συμμετοχής των γυναικών στην αγορά εργασίας στην Ελλάδα είναι το χαμηλότερο στην Ευρωπαϊκή Ενωση (στοιχεία 2011), ενώ υπολείπεται περισσότερο του ποσοστού των ανδρών από ό,τι σε οποιοδήποτε άλλο κράτος-μέλος πλην της Ιταλίας και της Μάλτας

Εντυπη έκδοση

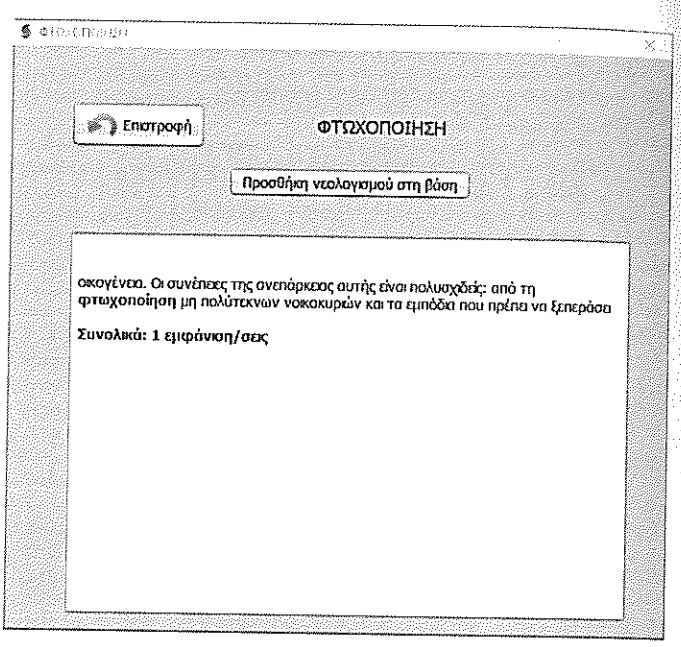
A.

```
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0" xmlns:ndc="http://www.academyofathens.gr/keeon/NeoDemiaCorpus/ns/1.0">
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title xml:id="Kathimerini_November_2015.xml"/>
                <!-- τυπώνεται ο τίτλος του ηλεκτρονικού αρχείου που παράγεται μέσω της λειτουργίας συγχώνευσης άρθρων ανά
                    περίοδο συλλογής -->
```

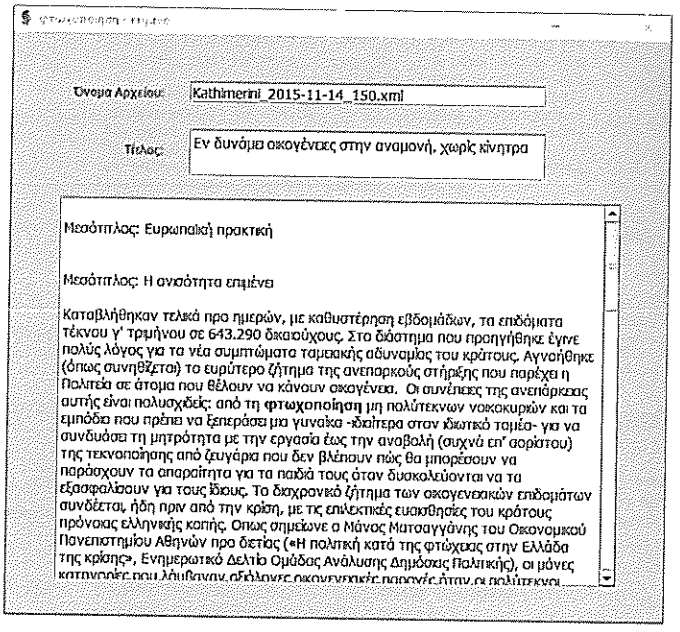
B.

```
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0" xmlns:ndc="http://www.academyofathens.gr/keeon/NeoDemiaCorpus/ns/1.0">
  <teiHeader>
      <fileDesc>
          <titleStmt>
              <title xml:id="Kathimerini_2015-11-14.xml"/>
              <!-- τυπώνεται ο τίτλος του ηλεκτρονικού αρχείου που παράγεται μέσω της λειτουργίας διαχωρισμού άρθρων ανά
                  ημέρα συλλογής -->
```

Εικόνα 4. Δημιουργία αρχείων ανά χρονική περίοδο συλλογής (λ.χ. ειδήσεις του μήνα, της ημέρας κ.ά.). Παράγονται πολυμερή τεκμήρια και κωδικοποιούνται με τη χρήση του στοιχείου <teiCorpus>



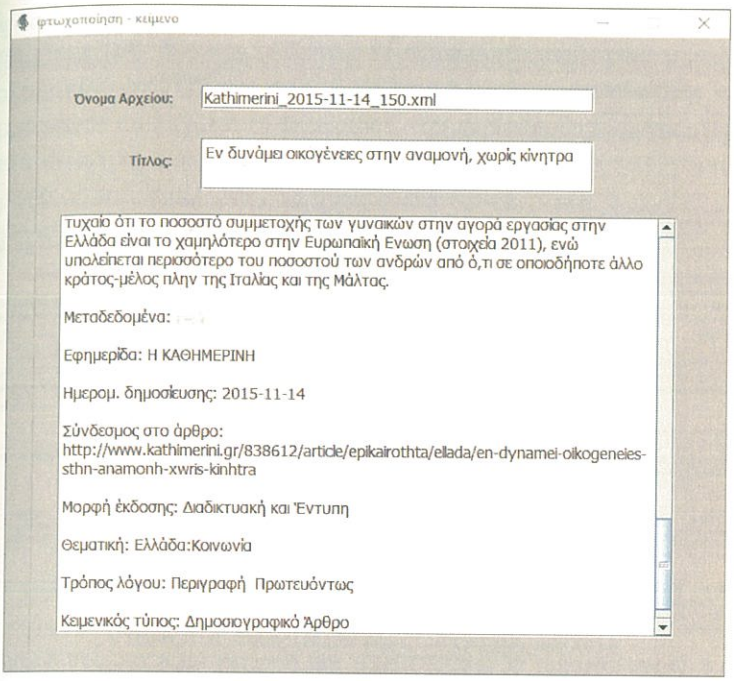
Εικόνα 5



Εικόνα 6

να 5

α 6

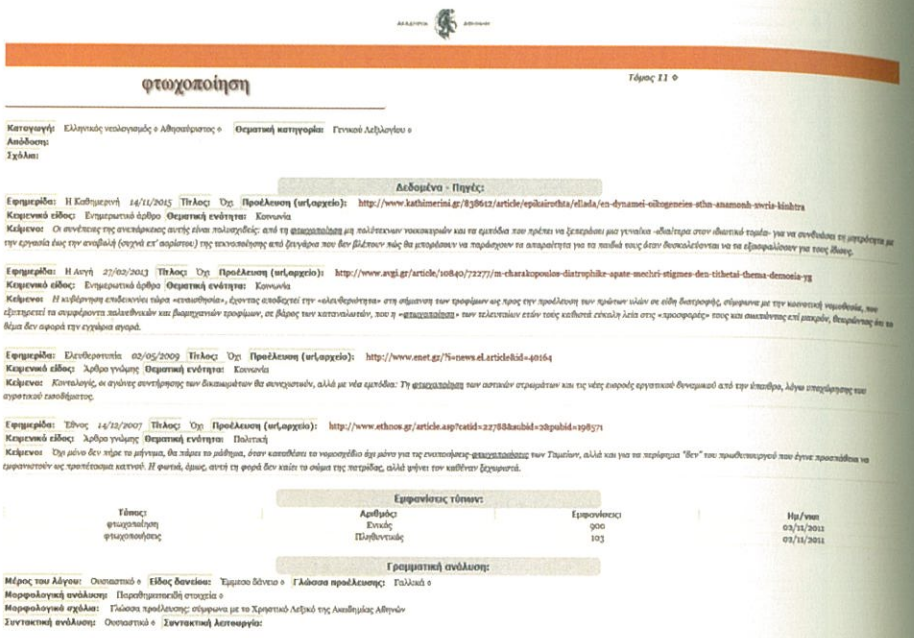


Εικόνα 7

Εικόνες 5-7. Στιγμιότυπα οθόνης από το Σύστημα Ανίχνευσης Νεολογισμών *της* Νεοδημίας, *το οποίο προσφάτως: (α) επεκτάθηκε με προσθήκη δύο νέων καταλόγων επώνυμων οντοτήτων στον έλεγχο των λεξικών, καθώς και με τη δυνατότητα πληκτρολόγησης από τους χρήστες, σε κατάλογο αποκλεισμού, επιλεγμένων λέξεων (αρκτικόλεξα, κύρια ονόματα κ.ά.) ή τυπογραφικών κ.ά. λαθών που εμφανίζονται ως «θόρυβος» στα δεδομένα (β) συνδέθηκε με το ΣΕΝ*[47]

Μετά την ολοκλήρωση του ημι-αυτόματου εντοπισμού του υποψήφιου νεολογισμού και εφόσον εγκριθεί από τον ερευνητή, ο λημματικός τύπος καταγράφεται με επιλογή παραδειγμάτων στη βάση *Νεοδημία*. Σε αυτό το στάδιο επιλέγεται από τον ερευνητή και το κειμενικό είδος, αφού λάβει υπ' όψιν του όλες τις παραμέτρους, όπως αναδείχθηκαν στο προηγούμενο μοντέλο (βλ. Εικόνα 8).

47 Υλοποίηση: φοιτητής Τηλέμαχος Πετσόπουλος, *Τμήμα Πληροφορικής & Τηλεματικής, Χαροκόπειο Πανεπιστήμιο*. Επιβλέποντες: Ηρακλής Βαρλάμης & Αναστασία Χριστοφίδου.



φτωχοποίηση

Τόμος ΙΙ ◊

Καταγωγή: Ελληνικός νεολογισμός ◊ Αθησαύριστος ◊ Σημασιακή κατηγορία: Γενικού Λεξιλογίου ◊
Απόδοση:
Σχόλια:

Δεδομένα - Πηγές:

Εφημερίδα: Η Καθημερινή 14/11/2015 Τίτλος: Όχι Προέλευση (url,αρχείο): http://www.kathimerini.gr/838612/article/epikairothta/ellada/en-dynamei-oikogeneies-sthn-anamonh-xwria-kinhtra
Κειμενικό είδος: Ενημερωτικό άρθρο Θεματική ενότητα: Κοινωνία
Κείμενο: Οι συνέπειες της ανεπάρκειας αυτής είναι πολυσχιδείς: από τη φτωχοποίηση μη πολύτεκνων νοικοκυριών και τα εμπόδια που πρέπει να ξεπεράσει μια γυναίκα «βαπτίζει στον ιδιωτικό τομέα» για να συνδυάσει τη μητρότητα με την εργασία έως την αναβολή (ακόμη κι αναίτιου) της τεκνοποίησης από ζευγάρια που δεν βλέπουν πώς θα μπορέσουν να παράσχουν τα απαραίτητα για τα παιδιά τους όταν θυσκολεύονται να τα εξασφαλίσουν για τους ίδιους.

Εφημερίδα: Η Αυγή 27/02/2013 Τίτλος: Όχι Προέλευση (url,αρχείο): http://www.avgi.gr/article/10840/7227/m-charakopoulos-diatrophika-apate-mechri-stigmes-den-tithetai-thema-demosia-yg
Κειμενικό είδος: Ενημερωτικό άρθρο Θεματική ενότητα: Κοινωνία
Κείμενο: Η κυβέρνηση επιδεικνύει τώρα «ευαισθησία», έχοντας αποδεχτεί την «αλεπαθρωπότητα» στη σημανση των τροφίμων ως προς την προέλευση των πρώτων υλών σε είδη διατροφής, σύμφωνα με την κοινοτική νομοθεσία, που εξετηρεστεί τα συμφέροντα πολυεθνικών και βιομηχανιών τροφίμων, σε βάρος των καταναλωτών, που η «φτωχοποίηση» των τελευταίων ετών τούς κεδυστά εκκλίνει λεία στις «προσφορές» τους και σωπάνουμε επί μακρόν, θεωρώντας ότι το θέμα δεν αφορά την εγχώρια αγορά.

Εφημερίδα: Ελευθεροτυπία 02/05/2009 Τίτλος: Όχι Προέλευση (url,αρχείο): http://www.enet.gr/?i=news.el.article&id=40164
Κειμενικό είδος: Άρθρο γνώμης Θεματική ενότητα: Κοινωνία
Κείμενο: Κανταλογής, οι αγώνες συντήρησης των δικαιωμάτων θα συνεχιστούν, αλλά με νέα εμπόδια: Τη φτωχοποίηση των αστικών στρωμάτων και τις νέες ευρυές εργασιακού θυναμικού από την ύπαιθρο, λόγω υπερχώρησης του αγροτικού εισοδήματος.

Εφημερίδα: Έθνος 14/12/2007 Τίτλος: Όχι Προέλευση (url,αρχείο): http://www.ethnos.gr/article.asp?catid=22788&subid=2&pubid=198571
Κειμενικό είδος: Άρθρο γνώμης Θεματική ενότητα: Πολιτική
Κείμενο: Όχι μόνο δεν πήρε το μήνυμα, θα πάρει το μάθημα, όσων καταδίκασε το νομοσχέδιο όχι μόνο για τις ενισχύσεις-φτωχοποίησεις των Τμημάτων, αλλά και για τα περίφημα "δεν" του πρωθυπουργού που έγινε προσπάθεια να εμφανιστούν ως τροπέτσουμα κατινοί. Η φωτιά, όμως, αυτή τη φορά δεν κάιει τα σώματα της πατρίδας, αλλά ψήνει τον καθένα ξεχωριστά.

Εμφανίσεις τύπων:

Τύπος:	Αριθμός:	Εμφανίσεις:	Ημ/νια
φτωχοποίηση	Ενικός	900	03/11/2011
φτωχοποιήσεις	Πληθυντικός	103	03/11/2011

Γραμματική ανάλυση:

Μέρος του λόγου: Ουσιαστικό ◊ Είδος δανείου: Έμμεσο δάνειο ◊ Γλώσσα προέλευσης: Γαλλικά ◊
Μορφολογική ανάλυση: Παραθηματοειδή στοιχεία ◊
Μορφολογικό σχόλιο: Γλώσσα προέλευσης: σύμφωνα με το Χρηστικό Λεξικό της Ακαδημίας Αθηνών
Συντακτική ανάλυση: Ουσιαστικό ◊ Συντακτική λειτουργία:

Εικόνα 8. Ο νεολογισμός ΦΤΩΧΟΠΟΙΗΣΗ *και η καταγραφή του στη βάση δεδομένων της* Νεοδημίας *με επικαιροποιημένα στοιχεία*[48]

6. *Ενδεικτικά σενάρια αναζήτησης*

Ένα κείμενο επισημειωμένο σε XML παρέχει στους συντάκτες του διευρυμένες δυνατότητες αναζήτησης, τουλάχιστον τόσες, όσες οι σημειωτικές ψηφίδες που χρησιμοποιήθηκαν. Όπως επισημάνθηκε, βασικός στόχος ήταν η συμβατότητα με διαθέσιμα υπολογιστικά εργαλεία ανάλυσης σωμάτων κειμένων, ελλείψει προς το παρόν δυνατότητας να μεταφερθεί το ΣΕΝ σε μια NoSQL βάση δεδομένων,[49] ή στην *NoSketchEngine*,[50] την ελεύθερη εκδοχή του δημοφιλούς συστήματος διαχείρισης ΗΣΚ.[51]

48 Για τα σχεδιαστικά χαρακτηριστικά του περιβάλλοντος βλ. Χριστοφίδου κ.ά. (2013).

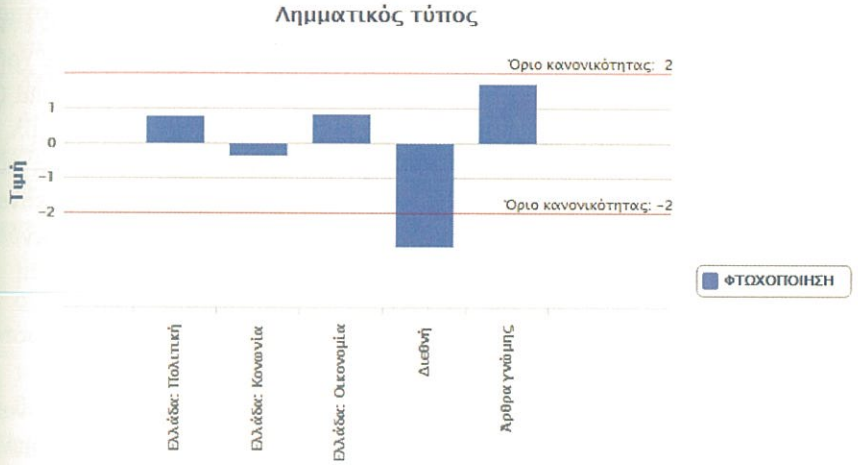
49 Όπως η *eXistdb*, βλ. http://www.exist-db.org/exist/apps/homepage/index.html

50 Βλ. https://nlp.fi.muni.cz/trac/noske

51 Προς την κατεύθυνση αυτή έχει προσφάτως δρομολογηθεί η χρήση της μηχανής αναζήτησης και ανάλυσης δεδομένων σε πραγματικό χρόνο *Elasticsearch* (βλ. Εικόνα 2). Η συγκεκριμένη μηχανή βασίζεται στο έργο ανοικτού κώδικα Apache Lucene, χρησιμοποιείται από μεγάλες εταιρείες/οργανισμούς παγκοσμίως (π.χ. https://www.elastic.co/use-cases) και θα επιτρέψει να εκτελούνται πληρέστερα οι αναζητήσεις στο ΣΕΝ, καθώς και να αναλύονται τα δεδομένα σε πραγματικό χρόνο, εφόσον μάλιστα το *IMS Open Corpus*

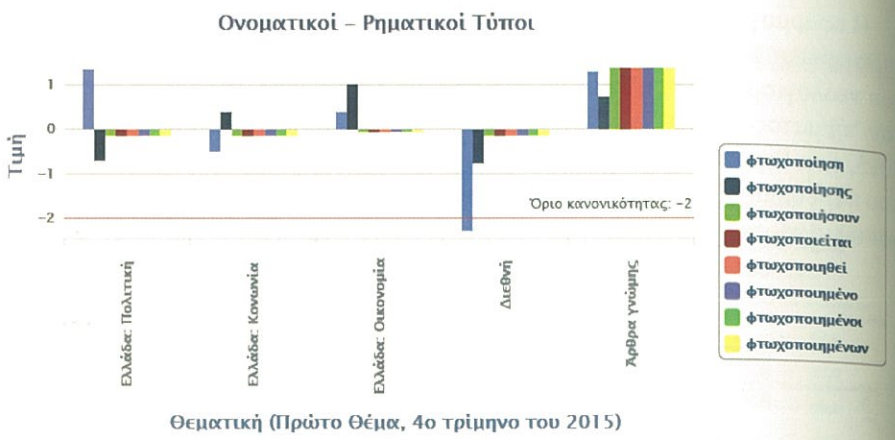
Η εφαρμογή σεναρίων αναζήτησης αποτελεί αντικείμενο νέας μελέτης. Εδώ ενδεικτικά, σε σύνδεση με το κωδικοποιημένο κείμενο που φέρει το παράδειγμα του νεολογισμού ΦΤΩΧΟΠΟΙΗΣΗ παραθέτουμε στιγμιότυπα από την κατανομή του λήμματος σε επιλεγμένες εφημερίδες, την περίοδο Οκτώβριος-Δεκέμβριος 2015, ανά θεματική ενότητα, με βάση το πρώιμο σύστημα κατηγοριοποίησης που παρουσιάσαμε βάσει URL. Ελπίζουμε σύντομα να είμαστε σε θέση να εφαρμόσουμε και το νέο σύστημα κατηγοριοποίησης, με μηχανική μάθηση.

Έθνος, Η Καθημερινή, Το Βήμα: 4ο τρίμηνο του 2015



Θεματική (Πρώτο Θέμα, 4ο τρίμηνο του 2015)

Workbench (το γνωστότερο ανοιχτό σύστημα διαχείρισης ΗΣΚ, βλ. http://cwb.sourceforge.net) θεωρείται πλέον πεπερασμένο υπολογιστικά, ως προς τον όγκο των δεδομένων που μπορεί να επεξεργαστεί.



Εικόνες 9α, β, γ. Στιγμιότυπα οθόνης (TXM)

Όπως επιτρέπει η λειτουργία *Analyse des Spécificités*[52] του προγράμματος *TXM – Projet Textométrie*[53] να διαφανεί, στο ΣΕΝ της *Νεοδημίας* ο νεολογισμός παρουσιάζει ελαφρά φθίνουσα πορεία ως προς τη χρήση του το συγκεκριμένο τρίμηνο του 2015 (Εικόνα 9α). Επιπλέον, φαίνεται να εξειδικεύεται πιο πολύ στα άρθρα γνώμης, ενώ αίσθηση προκαλεί η στατιστικώς σημαντική (< -2) απουσία του στο λεξιλόγιο των άρθρων της διεθνούς επικαιρότητας (Εικόνα 9β). Στην Εικόνα 9γ γίνεται φανερή η διαφοροποίηση της απόδοσης της έννοιας 'φτωχοποίηση' στα άρθρα γνώμης (διαφορετικό κειμενικό είδος) με χρήση τόσο ρηματικών όσο και ονοματικών τύπων σε αντίθεση με τα υπόλοιπα (επί το πλείστον) ενημερωτικά άρθρα όπου κυριαρχούν οι ονοματικοί τύποι (βλ. και ενότητα 4, σημείο Β πιο πάνω). Σε κάθε περίπτωση, η έννοια την οποία κωδικοποιούν γλωσσικά αυτοί οι δύο νεολογισμοί – ρήμα / ουσιαστικό – φαίνεται να απασχολεί ιδιαιτέρως την ελληνική επικαιρότητα, κάτι που αναδεικνύεται ανάγλυφα μέσα από τη σωματοκειμενική και κειμενοκεντρική προσέγγιση.

7. Επίλογος

Στην παρούσα εργασία προσπαθήσαμε να περιγράψουμε κατά το δυνατόν ευσύνοπτα τις εν εξελίξει πολύπλευρες προσπάθειές μας για τον σχεδιασμό και τη συγκρότηση ενός σώματος κειμένων από επιλεγμένες διαδικτυακά εμφανιζόμενες εφημερίδες με σκοπό την ημι-αυτόματη ανίχνευση, οργανωμένη καταγραφή και διαρκή παρακολούθηση νεολογισμών της Νέας Ελληνικής.

52 http://txm.sourceforge.net/doc/manual/manual1.xhtml

53 http://textometrie.ens-lyon.fr/?lang=fr_

Τους δύο βασικούς πυλώνες της έρευνας αποτέλεσαν η προσαρμογή του γνωστού προτύπου κωδικοποίησης κειμένων TEI στη δική μας γλωσσολογικά και κειμενογλωσσολογικά προσανατολισμένη προσέγγιση, καθώς και η επιλογή ενός μοντέλου κειμενικής ταξινόμησης με διακριτά επίπεδα ανάλυσης (επικοινωνιακά και δομικά/γνωσιακά) αλλά και καθαρές εννοιολογικές γραμμές μεταξύ των πολυχρησιμοποιημένων ετικετών *τρόποι (είδη) λόγου, κειμενικοί τύποι* και *κειμενικά είδη*.

Θεωρούμε ότι μόνο ένα σώμα κειμένων που έχει οργανωθεί βάσει στέρεων (κειμενο)γλωσσολογικών αρχών μπορεί να συμβάλει με επιστημονική ασφάλεια στη διερεύνηση των φαινομένων μιας φυσικής γλώσσας.

ΕΛΛΗΝΟΓΛΩΣΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑ[54]

Ακριτίδου, Μ. (2014). Οι νεοελληνικές σπουδές μπροστά στην ψηφιακή πραγματικότητα. Η ανάπτυξη ψηφιακού σώματος νεοελληνικής ποίησης από το ΚΕΓ με το διεθνές πρότυπο Text Encoding Initiative. Στο Κ. Δημάδης (Επιμ.), *Πρακτικά του Ε΄ Ευρωπαϊκού Συνεδρίου Νεοελληνικών Σπουδών της Ευρωπαϊκής Εταιρείας Νεοελληνικών Σπουδών, Τόμος Δ΄* (σσ. 651-679). Αθήνα: Ευρωπαϊκή Εταιρεία Νεοελληνικών Σπουδών.

Αναστασιάδη-Συμεωνίδη Α., Αλεξιάδου, Χ. & Νικολάου, Γ. (2009). *Ηλεκτρονική βάση νεολογισμών της Νέας Ελληνικής*. Στο Α. Χριστοφίδου (Επιμ.), *Δελτίο Επιστημονικής Ορολογίας και Νεολογισμών, 9-10*, 419-439. Αθήνα: Ακαδημία Αθηνών.

Αφεντουλίδου, Β. (2016). *Τροποποίηση σχήματος TEI για τη Νεοδημία – ΣΕΝ* (Εσωτερικό παραδοτέο). ΚΕΕΟΝ, Ακαδημία Αθηνών.

Αφεντουλίδου, Β. (υπό προετοιμ.) *Όψεις της ομοιότητας στη Νέα Ελληνική. Η περίπτωση του δείκτη «σαν» και η σχέση του με συναφείς εκφράσεις* (Αδημοσίευτη διδακτορική διατριβή). Πανεπιστήμιο Αθηνών.

Γούτσος Δ. & Φραγκάκη, Γ. (2015). *Εισαγωγή στη γλωσσολογία σωμάτων κειμένων*. Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Ανακτήθηκε από https://repository.kallipos.gr/handle/11419/1932

Δημητρούλια, Τ. & Τικτοπούλου, Κ. (2015). *Ψηφιακές λογοτεχνικές σπουδές*. Αθήνα: ΣΕΑΒ. Ανακτήθηκε από https://repository.kallipos.gr/handle/11419/5827

Κουτσογιάννης, Δ. (2009). Κειμενικά είδη, σώματα κειμένων και γλωσσική εκπαίδευση: Από το μερικό στο ολικό. *Επιστημονική Επετηρίδα Φιλοσοφικής Σχολής Θεσσαλονίκης, Τεύχος Τμήματος Φιλολογίας, 11*, 149-183.

54 Τελευταίος έλεγχος των διαδικτυακών διευθύνσεων: Ιούνιος 2017.

Μικρός, Γ. (2004). Ηλεκτρονικά σώματα κειμένων και ορολογία. Στο Μ. Κατσογιάννου & Ε. Ευθυμίου (Επιμ.), *Ελληνική ορολογία: Έρευνα και εφαρμογές* (σσ. 139-164). Αθήνα: Καστανιώτης.

Πολίτης, Π. (2008α). Επιστημολογικά προβλήματα στη μελέτη του λόγου της μαζικής επικοινωνίας. Στο Μ. Θεοδωροπούλου (Επιμ.), *Θέρμη και φως*. Αφιερωματικός τόμος στη μνήμη του Α.-Φ. Χριστίδη (σσ. 425-436). Θεσσαλονίκη: Κέντρο Ελληνικής Γλώσσας.

Πολίτης, Π. (2008β). Το κύριο άρθρο ελληνικών εφημερίδων. Στο Π. Πολίτης (Επιμ.), *Ο λόγος της μαζικής επικοινωνίας. Το ελληνικό παράδειγμα* (σσ. 271-328). Θεσσαλονίκη: ΑΠΘ-ΙΝΣ.

Τσελίγκα-Γκαζιάνη, Θ. (2008). Μορφή και λειτουργία των τίτλων ηλεκτρονικών εφημερίδων: Η διαδικτυακή έκδοση του Έθνους. Στο Π. Πολίτης (Επιμ.), *Ο λόγος της μαζικής επικοινωνίας. Το ελληνικό παράδειγμα* (σσ. 600-623). Θεσσαλονίκη: ΑΠΘ-ΙΝΣ.

Χατζησαββίδης, Σ. (1999). *Ελληνική γλώσσα και δημοσιογραφικός λόγος*. Αθήνα: Gutenberg.

Χριστοφίδου, Α. (2001). *Ο ποιητικός νεολογισμός και οι λειτουργίες του*. Αθήνα: Gutenberg.

Χριστοφίδου, Α. (2008). Συνοχικές λειτουργίες νεολογισμών και κειμενικές κατηγορίες. Στο Α. Μόζερ, Αικ. Μπακάκου-Ορφανού, Χ. Χαραλαμπάκης & Δ. Χειλά-Μαρκοπούλου (Επιμ.), *Γλώσσης χάριν*. Τιμητικός τόμος για τον Γ. Μπαμπινιώτη (σσ. 475-486). Αθήνα: Ελληνικά Γράμματα.

Χριστοφίδου, Α. (Επιμ.). (2012α). *Δελτίο Επιστημονικής Ορολογίας και Νεολογισμών, 11*. Αθήνα: Ακαδημία Αθηνών.

Χριστοφίδου, Α. (2012β). *Κειμενογλωσσολογία* [Πανεπιστημιακές σημειώσεις (χγφ)]. ΕΚΠΑ, ΠΤΔΕ.

Χριστοφίδου, Α. (2013). *Κειμενογλωσσολογία* [Πανεπιστημιακές σημειώσεις (χγφ)]. ΕΚΠΑ, ΠΤΔΕ.

Χριστοφίδου, Α. (υπό δημ.). Κειμενικά είδη και τρόποι λόγου. Στο Κ. Ντίνας (Επιμ.), *Figura in praesentia*. Τιμητικός τόμος Θ. Νάκα. Αθήνα: ΕΚΠΑ, ΠΤΔΕ.

Χριστοφίδου, Α., Αφεντουλίδου, Β., Καρασίμος, Θ. & Δημητροπούλου, Ε. (2013). Ηλεκτρονικό πρόγραμμα *Νεοδημία*. Προκλήσεις και δικτυο-λύσεις. Στο Α. Χριστοφίδου (Επιμ.), *Δημιουργία και μορφή στη γλώσσα, Δελτίο Επιστημονικής Ορολογίας και Νεολογισμών, 12*, 198-243. Αθήνα: Ακαδημία Αθηνών.

Χριστοφίδου, Α. & Αφεντουλίδου, Β. (2016). Νεολογία και κειμενικό είδος. Μια σωματοκειμενική προσέγγιση. Εισήγηση στην Ημερίδα «Ο Δημιουργικός Λόγος στην Εκπαίδευση», 27-28 Ιουνίου, Μαράσλειο Διδασκαλείο. ΕΚΠΑ, Εργαστήριο Εφαρμοσμένης Γλωσσολογίας & Μελέτης της Λογοτεχνίας και της Ρητορικής & Ελληνική Ένωση για την Προώθηση της Ρητορικής στην Εκπαίδευση. Υπό την αιγίδα του ΥΠΠΕΘ.

Χριστοφίδου, Α., Καρασίμος, Θ. & Αφεντουλίδου, Β. (2014). Έλεγχος, παρακολούθηση και ταξινόμηση νεολογισμών με το ηλεκτρονικό πρόγραμμα *Νεοδημία*: Η προσέγγιση των νέων δανείων. Στο G. Kotzoglou, K. Nikolou, E. Karantzola, K. Frantzi, I. Galantomos, M. Georgalidou, V. Kourti-Kazoullis, C. Papadopoulou & E. Vlachou (Επιμ.), *Selected Papers of the 11th International Conference on Greek Linguistics* (σσ. 1850-1868). Rhodes: University of the Aegean.

ΞΕΝΟΓΛΩΣΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

Andersen, G. & Hofland, K. (2012). Building a large corpus based on newspapers from the web. In G. Andersen (Ed.), *Exploring newspaper language. Using the web to create and investigate a large corpus of modern Norwegian* (pp. 1-28). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Bański, P. (2010). Why TEI standoff annotation doesn't quite work: And why you might want to use it nevertheless. *Proceedings of Balisage: The Markup Conference. Balisage Series on Markup Technologies, Vol. 5.* Retrieved from https://www.balisage.net/Proceedings/vol5/html/Banski01/BalisageVol5-Banski01.html

Bax, S. (2011). *Discourse and genre: Using language in context.* Hampshire: Palgrave Macmillan.

Bernstein, B. (1996). *Pedagogy, symbolic control and identity. Theory, research, critique.* London: Taylor and Francis.

Bhatia, V. (2004). *Worlds of written discourse.* London: Continuum.

Boase-Beier, J. (1987). *Poetic compounds.* Berlin: Niemeyer.

Burnard, L. & Sperberg-McQueen, M.C. (2012). *TEI Lite: Encoding for Interchange: An introduction to the TEI. Final revised edition for TEI P5. TEI Consortium.* Retrieved from http://www.tei-c.org/Guidelines/Customization/Lite/

Cabré, T. & Nazar, R. (2012). Towards a new approach to the study of neology. *Neologica, 6,* 63-80.

Cammozzo, A. (2013). Collecting corpora on the web: The dubious joys of scraping web pages. Presented at the 1st *IQLA Summer School on Quantitative Analysis of Textual Data*, Padua.

Cartier, E. (2017). Neoveille, a web platform for neologism tracking. In A. Peñas & A. Martins (Eds.), *Proceedings of the EACL 2017 Software Demonstrations* (pp. 95-98). Valencia, Spain: Association for Computational Linguistics.

Christofidou, A. (1994). *Okkasionalismen in poetischen Texten.* Tübingen: Gunter Narr.

Christofidou, A. & Dimitropoulou, E. (in press). Nonce formations in Greek children's literature. Contrastive analysis and language teaching. *Proceedings of the International Conference 'The Child and the Book'*. Athens: UOA, Department of Primary Education.

Corpus Est Républicain, Centre National de Ressources Textuelles et Lexicales (2016, January 12). Retrieved from http://www.cnrtl.fr/corpus/estrepublicain/

COW (COrpora from the Web), Freie Universität Berlin. (2017, April 18). Retrieved from https://webcorpora.org

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing, 25*(4), 447-464.

Davies, M. (2015). Corpora: An introduction. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 11-31). Cambridge: Cambridge University Press.

DeRose, S.J., Durand, D.G., Mylonas, E. & Renear, A.H. (1990). What is text, really? *Journal of Computer Documentation, 21*(3), 1-24.

Deutsches Textarchiv Guidelines, Berlin-Brandenburg Academy of Sciences and Humanities. (2016, January 12). Retrieved from http://www.deutschestextarchiv.de/doku

Dressler, W.U. (1982). Zum Verhältnis von Wortbildung und Textlinguistik (mit Beispielen aus der schönen Literatur). In J. Petöfi (Eds.), *Text vs. sentence continued* (pp. 96-106). Hamburg: Buske.

Dressler, W.U. (1987). Word formation (WF) as part of natural morphology. In W.U. Dressler (Eds.), *Leitmotifs in natural morphology* (pp. 99-126). Amsterdam: John Benjamins.

Fletcher, W. (2013). Corpus analysis of the World Wide Web. In C.A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 339-347). Oxford: Wiley-Blackwell.

Fraser, B. (2011). The TEI dictionary schema for XML: Its worth and weakness for the lexicographer. Retrieved from http://people.ds.cam.ac.uk/blf10/links/TEI.html

Gérard, C., Falk, I. & Bernhard, D. (2014). Traitement automatisé de la néologie: Pourquoi et comment intégrer l' analyse thématique? In F. Neveu, P. Blumenthal, L. Hriba, A. Gerstenberg, J. Meinschäfer & S. Prévost (Eds.), *4e Congrès Mondial de Linguistique Française (CMLF 2014), SHS Web of Conferences, Vol. 8* (pp. 2627-2646). France: EDP Sciences. Retrieved from https://www.shs-conferences.org/articles/shsconf/pdf/2014/05/shsconf_cmlf14_01208.pdf

Gries, S.T. (2017). Corpus approaches. In B. Dancygier (Ed.), *The Cambridge handbook of cognitive linguistics* (pp. 590-606). Cambridge: Cambridge University Press.

Gries, S.T. & Berez, A.L. (2017). Linguistic annotation in/for corpus linguistics. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation* (pp. 379-409). Berlin & New York: Springer.

Hagen, T.M. (2012). Automatic topic classification of a large newspaper corpus. In G. Andersen (Ed.), *Exploring newspaper language. Using the web to create and investigate a large corpus of modern Norwegian* (pp. 111-130). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Halkidi, M., Nguyen, B., Varlamis, I. & Vazirgiannis, M. (2003). THESUS: Organizing web document collections based on link semantics. *The International Journal on Very Large Data Bases, 12*(4), 320-332.

Hardie, A. (2014). Modest XML for corpora: Not a standard, but a suggestion. *ICAME, 38*(1), 73-103.

Haugen, O.E. (Ed.). (2008). *The Menota handbook: Guidelines for the electronic encoding of Medieval Nordic primary sources. Version 2.0.* Retrieved from http://www.menota.org/HB2_index.xml

Heiden, S., Magué, J.-P. & Pincemin, B. (2010). TXM: Une plateforme logicielle open-source pour la textométrie – conception et développement. In S. Bolasco, I. Chiari & L. Giuliano (Eds.), *10th International Conference on the Statistical Analysis of Textual Data – JADT 2010, Vol. 2* (pp. 1021-1032). Roma: Edizioni Universitarie di Lettere Economia Diritto.

Heinemann, W. & Viehweger, D. (1991). *Textlinguistik.* Berlin: Walter de Gruyter.

Ide, N. (2008). *XCES: Corpus Encoding Standard for XML. Version 1.0.4. Expert Advisory Group on Language Engineering Standards (EAGLES).* Retrieved from http://www.xces.org

Ide, N., Romary, L. & de la Clergerie, E. (2004). International Standard for a Linguistic Annotation Framework. *Natural Language Engineering, 10*(3-4), 211-225.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The TenTen corpus family. In A. Hardie & R. Love (Eds.), *7th International Corpus Linguistics Conference – CL 2013* (pp. 125-127). Lancaster: UCREL.

Jansen, L., Luijten, H. & Bakker, N. (Eds.). (2009). *Vincent van Gogh – the letters.* Van Gogh Museum, Amsterdam & Huygens ING, The Hague. Retrieved from http://vangoghletters.org/vg/

Kahrel, P., Barnett, R. & Leech, G. (1997). Towards cross-linguistic standards or guidelines for the annotation of corpora. In R. Garside, G. Leech & T. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 231-242). New York: Longman.

Kerremans, D. (2015). *A web of new words. A corpus-based study of the conventionalization process of English neologisms.* Frankfurt am Main: Peter Lang.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography 1*(1): 1-30.

Kinne, M. (1998). Der lange Weg zum deutschen Neologismenwörterbuch. In W. Teubert (Ed.), *Neologie und Korpus* (pp. 63-110). Tübingen: Gunter Narr.

Knop, de S. (1987). *Metaphorische Komposita in Zeitungsüberschriften.* Hamburg: Niemeyer.

Koutsis, I., Kouklakis, G., Mikros, G.K. & Markopoulos, G. (2005). MINOTAVROS: A tool for the semi-automated creation of large corpora from the web. *Proceedings from the Corpus Linguistics Conference Series, Vol. 1* (pp. 1-8). Birmingham: CCR. Retrieved from http://www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2005-conf-e-journal.aspx

Kristiansen, M. & Andersen, G. (2012). Corpus approaches to neology and their relevance for dynamic domains. *Neologica, 6*, 43-62.

Kübler, S. & Zinsmeister, H. (2015). *Corpus linguistics and linguistically annotated corpora.* London: Bloomsbury Academic.

Lee, D. (2001). Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology, 5*(3), 37-72. Retrieved from http://llt.msu.edu/vol5num3/lee/

Leech, G. (1997). Introducing corpus annotation. In R. Garside, G. Leech & T. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 1-18). New York: Longman.

Lehmberg, T. & Wörner, K. (2008). Annotation standards. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (Vol. 1, pp. 484-500). Berlin: Walter de Gruyter.

Lenk, H. & Chesterman, A. (Eds.). (2005). *Pressetextsorten im Vergleich – Contrasting text types in the Press.* Zürich: Olms.

Lüdeling, A., Evert, S. & Baroni, M. (2007). Using web data for linguistic purposes. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 7-24). Amsterdam/New York: Rodopi.

McEnery, T. & Rayson, P. (1997). A corpus / annotation toolbox. In R. Garside, G. Leech & T. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 194-208). New York: Longman.

Meurer, P. (2012). Corpuscle – a new corpus management platform for annotated corpora. In G. Andersen (Ed.), *Exploring newspaper language. Using the web to create and investigate a large corpus of modern Norwegian* (pp. 31-49). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Michelfeit, J., Rychlý,
ears on. *Lexicography*

gismenwörterbuch. In
übingen: Gunter Narr.
Zeitungsüberschriften.

s, G. (2005). MINO-
of large corpora from
Conference Series, Vol.
http://www.birming-
conference-archives/

s to neology and their

nd linguistically anno-

l styles: Clarifying the
gle. *Language Learning*
su.edu/vol5num3/lee/
Garside, G. Leech & T.
nation from computer

. In A. Lüdeling & M.
lbook (Vol. 1, pp. 484-

orten im Vergleich –

ata for linguistic pur-
ds.), *Corpus linguistics*
pi.
oolbox. In R. Garside,
Linguistic information
Longman.
platform for annotat-
ner language. Using the
Norwegian (pp. 31-49).
g Company.

Peschel, C. (2002). *Zum Zusammenhang von Wortneubildung und Textkonstitution.* Berlin: Niemeyer.

Pollard, C.J. & Sag, I.A. (1994). *Head-driven phrase structure grammar.* Studies in Contemporary Linguistics, Chicago, London: University of Chicago Press.

Renear, A.H., Mylonas, E. & Durand, D. (1993). Refining our notion of what text really is: The problem of overlapping hierarchies. Final version, 6 January 1993. Retrieved from http://cds.library.brown.edu/resources/stg/monographs/ohco.html

Renouf, A. (1993). Sticking to the text: A corpus linguist's view of language. *ASLIB Proceedings, 45*(5), 131-136.

Renouf, A. (2012). Defining neology to meet the needs of the translator: A corpus-based perspective. *Neologica, 6*, 17-41.

Rühlemann, C., Bagoutdinov, A. & O' Donnell, M.B. (2015). Modest XPath and XQuery for corpora: Exploiting deep XML annotation. *ICAME, 39*(1), 47-84.

Rühlemann, C. & O' Donnell, M.B. (2012). Introducing a corpus of conversational stories. Construction and annotation of the narrative corpus. *Corpus Linguistics and Linguistic Theory, 8*(2), 313-350.

Schmid, H.-J. (2008). New words in the mind: Concept-formation and entrenchment of neologisms. *Anglia – Zeitschrift für englische Philologie, 126*(1), 1-36.

Schröder, M. (1978). Über textverflechtende Wortbildungselemente. *Deutsch als Fremdsprache, 15*, 85-92.

Schröder, M. (1983). Zum Anteil der Wortbildungskonstruktionen an der Konstitution von Texten. *Beiträge zur Erforschung der deutschen Sprache, 3*, 108-119.

Scott, M. (2017). WordSmith Tools (Version 7). Stroud: Lexical Analysis Software.

Sinclair, J. (1995). From theory to practice. In G. Leech, G. Myers & J. Thomas (Eds.), *Spoken English on computer: Transcription, mark-up and application* (pp. 99-109). London: Longman.

Sinclair, J. (2005). Corpus and text – basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1-16). Oxford: Oxbow Books. Retrieved from https://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm

Smith, N., Hoffmann, S. & Rayson, P. (2008). Corpus tools and methods, today and tomorrow: Incorporating linguists' manual annotations. *Literary and Linguistic Computing, 23*(2), 163-180.

Stefanowitsch, A. & Flach, S. (2017). The corpus-based perspective on entrenchment. In Schmid, H.-J. (Ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge* (pp. 101-127). Washington, DC: American Psychological Association & Walter de Gruyter.

Stührenberg, M. (2012). The TEI and current standards for structuring linguistic data. *Journal of the Text Encoding Initiative (Online), 3*. Retrieved from https://jtei.revues.org/523

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Tambouratzis, G., Markantonatou, S., Hairetakis, N. & Carayannis, G. (2000). Automatic style categorisation of corpora in the Greek Language. *Proceedings of the 2nd International Conference on Language Resources and Evaluation* (pp. 135-140). Athens. Retrieved from http://www.lrec-conf.org/proceedings/lrec2000/pdf/301.pdf

Teubert, W. (Ed.). (1998). *Neologie und Korpus*. Tübingen: Gunter Narr.

The Shelley-Godwin Archive. Retrieved from http://shelleygodwinarchive.org

The TEI Consortium. (2016). *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.0.0. Last updated on 29th March 2016, revision 89ba24e*. Retrieved from http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/Guidelines.pdf

The Women Writers Project, Northeastern University. (2017, April 18). Retrieved from http://www.wwp.northeastern.edu/

Trampus, M. & Novak, B. (2012). The internals of an aggregated web news feed. In M. Bohanec, M. Gams, D. Mladenić, M. Grobelnik, M. Heričko, U. Kordeš, O. Markič, M. Smrdu, Z. Pirtošek, J. Lenarčič, L. Žlajpah, A. Gams, V. Rajkovič, T. Urbančič & M. Bernik (Eds.), *Proceedings of 15th International Multiconference on Information Society, Vol. A* (pp. 221-224). Ljubljana: Institut "Jožef Stefan".

Tyrkkö, J. (2010). Hyperlinks: Keywords or key words? In Bondi, M. & Scott, M. (Eds.), *Keyness in texts* (pp. 79-91). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Vanhoutte, E. & Van den Branden, R. (Eds.). (2003). *DALF (Digital Archive of Letters in Flanders) guidelines for the description and encoding of modern correspondence material Version 1.0*. Gent: Centrum voor Teksteditie en Bronnenstudie, Koninklijke Academie voor Nederlandse Taal- en Letterkunde. Revised 4 May 2005. Retrieved from http://ctb.kantl.be/project/dalf/dalfpdf/docall.pdf

Vanhoutte, E. & Van den Branden, R. (2009). Describing, transcribing, encoding, and editing modern correspondence material: A textbase approach. *Literary and Linguistic Computing, 24*(1), 77-98.

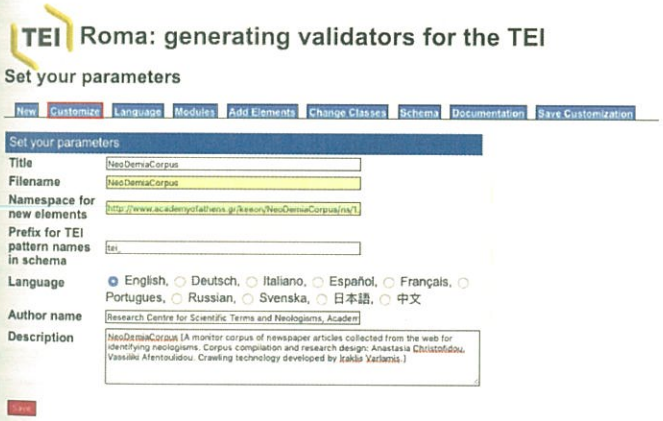
Varlamis, I., Tsirakis, N., Poulopoulos, V. & Tsantilas, P. (2014). An automatic wrapper generation process for large scale crawling of news websites. In

S. Katsikas, M. Hatzopoulos, T. Apostolopoulos, D. Anagnostopoulos, E. Carayiannis, T. Varvarigou & M. Nikolaidou (Eds.), *Proceedings of the 18th Panhellenic Conference on Informatics* (pp. 1-6). New York: Association for Computing Machinery.

Varlamis, I., Vazirgiannis, M., Halkidi, M. & Nguyen, B. (2004). THESUS, a closer view on web content management enhanced with link semantics. *IEEE Transactions on Knowledge and Data Engineering (TKDE), 16*(6), 685-700.

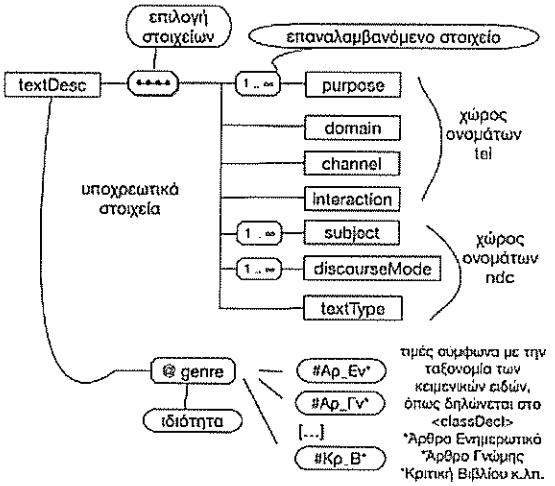
ΠΑΡΑΡΤΗΜΑ

Χρησιμοποιήθηκε το πρότυπο σχήμα *TEI for Linguistic Corpora* και στοιχεία από 10 ενότητες (για δικαιολόγηση βλ. Αφεντουλίδου 2016, *Τροποποίηση σχήματος TEI για τη Νεοδημία – ΣΕΝ*). Περιορίστηκαν τα στοιχεία λαμβάνοντας υπόψη το TEI LITE (Burnard & Sperberg-McQueen 2012) και τις ανάγκες του ερευνητικού προγράμματος. Τα σχήματα παρήχθησαν αρχικά μέσω της διαδικτυακής πλατφόρμας ROMA[55] (Εικόνες 10, 13-15) και τροποποιήθηκαν περαιτέρω σύμφωνα με τον σχεδιασμό του ΣΕΝ μέσω του oXygen XML editor, περιβάλλον επισημείωσης (προσεχώς διαθέσιμα στον δικτυότοπο του ΚΕΕΟΝ): Α) DTD Β) RELAX NG (REgular Language for XML Next Generation) Γ) Αρχείο ODD.



Εικόνα 10. Δημιουργία πρότυπου σχήματος επισημείωσης με το διαδικτυακό εργαλείο ROMA και επικύρωση δεδομένων

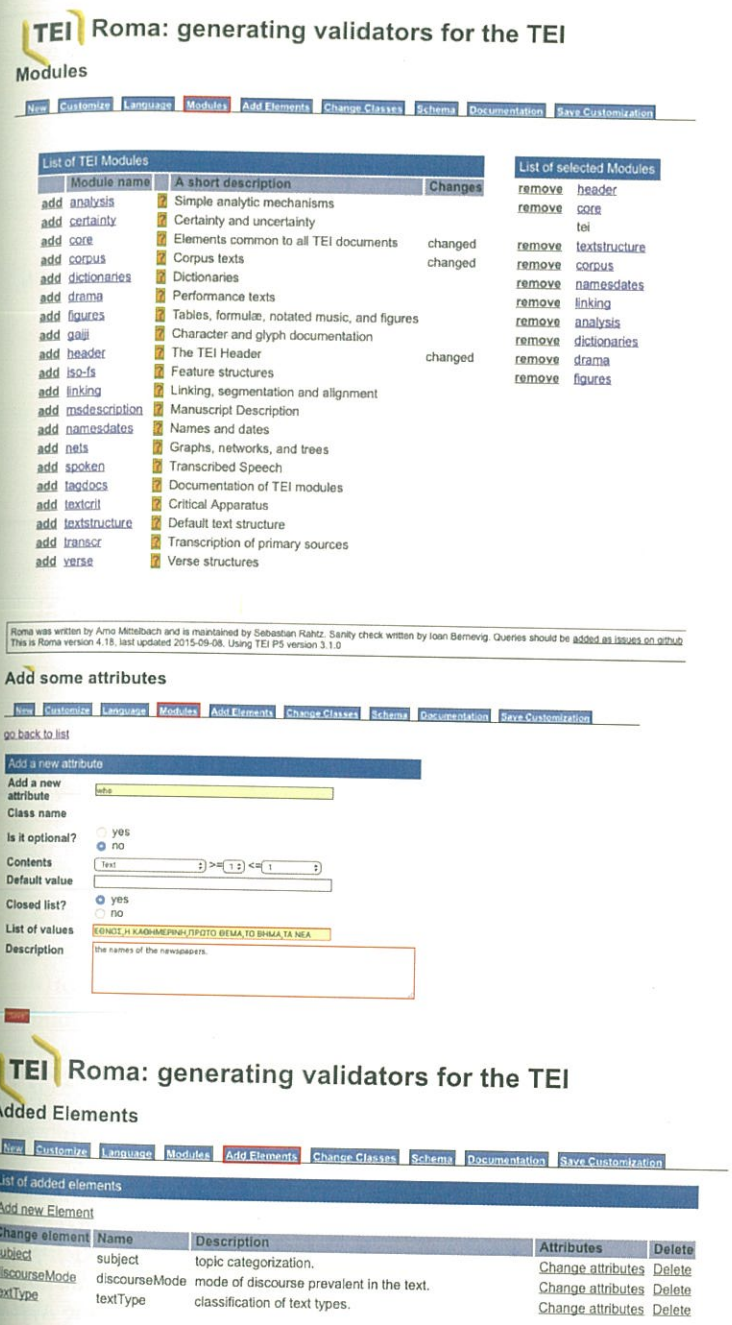
55 Διαθέσιμης στο http://www.tei-c.org/Roma/



Εικόνα 11. Οπτικοποίηση σχήματος XML για το τμήμα περιγραφής κειμένου <textDesc> βάσει καταστασιακών παραμέτρων

```
<elementSpec ident="discourseMode"
ns="http://www.academyofathens.gr/keeon/NeoDemiaCorpus/ns/1.0"
mode="add">
    <desc>mode of discourse prevalent in the text.</desc>
    <classes>
        <memberOf key="model.textDescPart"/>
    </classes>
    <content>
        <rng:ref name="macro.phraseSeq.limited"/>
    </content>
    <attList>
        <attDef ident="type" mode="add" usage="req">
            <desc>the traditional (3fold) model.</desc>
            <valList type="closed">
                <valItem ident="narration"/>
                <valItem ident="description"/>
                <valItem ident="argumentation"/>
                <valItem ident="hybrid"/>
            </valList>
        </attDef>
        <attDef ident="degree" mode="add">
            <desc>same as purpose.</desc>
            <valList type="closed">
                <valItem ident="primary"/>
                <valItem ident="secondary"/>
            </valList>
        </attDef>
    </attList>
</elementSpec>
```

Εικόνα 12. Τμήμα του TEI ODD (=One Document Does it all)



Εικόνες 13 -15. Παραμετροποίηση TEI μέσω της διαδικτυακής πλατφόρμας ROMA

Kathimerini.properties

```
#Mon Feb 02 14:53:10 EEST 2016
javascript=true
rss=http://www.kathimerini.gr/rss
charset=utf-8
name=Kathimerini
tagsToRemove=script,div
xpathbody=//div[@class="large-9 columns"]/div[@class="freetext"]
xpathsurtitle=-
xpathsubtitle=-
xpathauthor=//div[@class="large-8 columns"]/article[@id="item-article"]/header/span[@class="item-author"]/a
jquerymidtitle=strong
jquerycaption=//div[@class="large-9 columns"]/p[@id="item-photo-description"]
xpathlead=-
xpathedition=//div[@class="large-9 columns"]/span[@class="edition edition_ONLINE"]
dateFormat=EEE, dd MMM yyyy HH:mm:ss zzz
bodyremoveregex=(\/\*<!).*(>\*\/)
category=Unclassified
#Ρύθμιση υπεύθυνη για την αυτόματη θεματική κατηγοριοποίηση κειμένων βάσει URL
parquery=p
```

ToVima.properties

```
#Mon Feb 02 14:53:10 EEST 2016
javascript=true
rss=http://www.tovima.gr/feed/allnews/
charset=utf-8
name=ToVima
tagsToRemove=script
xpathbody=//div[@id="intext_content_tag"]
xpathsurtitle=//div[@class="article_ext"]//div[@class="article_suptitle"]
xpathsubtitle=//div[@class="article_ext"]//div[@class="article_subtitle"]
xpathauthor=//div[@class="page_wide"]//div[@class="container"]//div[@class="arthro_syntaktis heightfix"]//
div[@class="author_information"]//div[@class="author"]//span[@class="name"]/a
midtitle=strong
xpathcaption=//div[@class="article_ext"]//div[@class="article_photo"]//div[@class="article_photo_lead"]
xpathlead=
xpathedition=//div[@class="article_ext"]//div[@class="article_text"]//div[@class="article_source_cat
heightfix"]//div[@class="article_source"]
dateFormat=EEE, d MMM yyyy HH:mm:ss Z
bodyremoveregex=(\\/\\*<!).*?(>\\*\\/)
parQuery=(<br>)+
presstitle=TO BHMA
```

Εικόνες 16-17. Επέκταση του NDCrawler, ώστε να καλύπτει τα επίπεδα πληροφορίας του ΣΕΝ κατά το TEI P5 (αρχεία ρυθμίσεων για την εξαγωγή κειμένου από δύο εφημερίδες – NDCrawler έκδοση 5.0 και 5.5 αντιστοίχως)[56]

56 Ο *NDCrawler* TEI P5 άρχισε να υλοποιείται κατά τη διάρκεια της πρακτικής άσκησης φοιτητών Πληροφορικής στο ΚΕΕΟΝ (έκδοση 5.0: Αλεξίου Μιχάλης, *Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών, Πολυτεχνείο Κρήτης*) και εξελίχθηκε με ταυτόχρονη δημιουργία μετατροπέων για τη μετάπτωση των παλαιότερων αρχείων XML του ΣΕΝ σε TEI P5 XML (έκδοση 5.5: Βάσιος Γεώργιος, Λίχας Άγγελος, *Τμήμα Πληροφορικής & Τηλεματικής, Χαροκόπειο Πανεπιστήμιο*). Επιβλέποντες: Ηρακλής Βαρλάμης & Αναστασία Χριστοφίδου.