

τητα ισοδυναμεί με την ανάγκη να καλύπτεται ένα ευρύ φάσμα διαφορετικών γλωσσικών ποικιλιών και κειμενικών ειδών, διατηρώντας ταυτόχρονα την ισορροπία μεταξύ τους με βάση την πραγματική τους παρουσία στη σχετική γλωσσική κοινότητα.

Τέλος, για τον σχεδιασμό ενός σώματος κειμένων είναι απαραίτητο να προσδιορίζεται ο βαθμός στον οποίο τα κειμενικά δεδομένα θα είναι διαθέσιμα στην ακαδημαϊκή ή στην ευρύτερη κοινότητα, σύμφωνα με τους περιορισμούς των πνευματικών δικαιωμάτων αλλά και τις αρχές της ελεύθερης πρόσβασης στα ερευνητικά δεδομένα, καθώς και ο βαθμός και ρυθμός ανανέωσης των δεδομένων.

7.2. ΔΙΑΔΙΚΑΣΙΕΣ ΣΥΓΚΡΟΤΗΣΗΣ ΕΝΟΣ ΣΩΜΑΤΟΣ ΚΕΙΜΕΝΩΝ

Σύμφωνα με τους Γούτσο & Φραγκάκη (2015: 42-25), τα στάδια που ακολουθούν τον σχεδιασμό ενός σώματος κειμένων είναι τα ακόλουθα:

- ➞ διερεύνηση πηγών και ανάπτυξη
- ➞ εισαγωγή των δεδομένων
- ➞ μεταγραφή (για προφορικά δεδομένα)
- ➞ καθαρισμός και αποθήκευση
- ➞ τυποποίηση
- ➞ κωδικοποίηση
- ➞ γλωσσική επισημείωση.

Έχουμε ήδη αναφερθεί διεξοδικά σε προηγούμενα κεφάλαια στις πηγές κειμενικών δεδομένων και στα κύρια ζητήματα που συνδέονται με την αξιοποίησή τους. Εδώ μπορούμε να σημειώσουμε απλώς τη σημασία που έχουν οι αρχές προστασίας πνευματικών δικαιωμάτων για τη συγκρότηση ενός σώματος κειμένων, είτε τα κειμενικά δεδομένα προέρχονται από γραπτές (έντυπες ή ηλεκτρονικές) πηγές, είτε από προφορικές. Στη δεύτερη περίπτωση είναι απαραίτητη η ενημέρωση των πληροφοροπληροφορητών/πληροφορητριών για τους σκοπούς του ψηφιακού έργου και η γραπτή συναίνεσή τους για τη χρήση των δεδομένων. Γενικότερα, ζητήματα πνευματικών δικαιωμάτων μπορεί να εμποδίσουν τη δημιουργία ενός σώματος κειμένων με διάφορους τρόπους. Πάντως, θα πρέπει να υπογραμμιστεί ότι ζητήματα πνευματικών δικαιωμάτων υπερβαίνονται με την προϋπόθεση ότι τα κειμενικά δεδομένα που συγκροτούν ένα σώμα κειμένων χρησιμοποιούνται μόνο για ερευνητικούς σκοπούς και δεν είναι στην ολότητά τους διαθέσιμα στο ευρύ κοινό, παρά μόνο μέσω διεπαφών που προσφέρουν ένα μικρό μέρος τους με τη μορφή λ.χ. ενός συμφραστικού πίνακα με περιορισμένο εύρος λέξεων κ.ά.

Έχουμε επίσης αναφερθεί στις μεθόδους εισαγωγής των δεδομένων στο σώμα κειμένων, που περιλαμβάνουν τη χρήση ήδη ψηφιοποιημένου υλικού ή την ψηφι-

οποίησή του μέσω οπτικής αναγνώρισης χαρακτήρων (OCR) ή –σπανιότερα και σε συνάρτηση με το είδος των πηγών– πληκτρολόγηση (βλ. και Δημητρούλια & Τικτοπούλου 2015: 42-53). Ιδιαίτερη μνεία πρέπει να γίνει για τα προφορικά δεδομένα, στα οποία μεσολαβεί το στάδιο της μαγνητοφώνησης/ηχογράφησης και απομαγνητοφώνησης/μεταγραφής. Στην απομαγνητοφώνηση, στόχος μας είναι η πιστή απόδοση τού κειμένου έτσι ώστε, διαβάζοντας το κείμενο, να μπορούμε να φανταστούμε σε γενικές γραμμές πώς ειπώθηκε. Για τον σκοπό αυτό πρέπει να χρησιμοποιηθούν ειδικά σύμβολα απόδοσης του κειμένου, ενώ τα καθιερωμένα σημεία στίξης (π.χ. τελεία, κόμμα, θαυμαστικό, αποσιωπητικά) δεν πρέπει να χρησιμοποιούνται, γιατί έχουν πολλαπλές σημασίες που δεν μπορούν να αποκρυπτογραφηθούν μονοσήματα κάθε φορά.

Στο επόμενο πλαίσιο παρουσιάζονται ορισμένες συμβάσεις που χρησιμοποιούνται στη μεταγραφή προφορικών δεδομένων στα ελληνικά (βλ. Γεωργακοπούλου & Γούτσος 2011: 66-67) και ακολουθεί ένα παράδειγμα μεταγραφής αποσπάσματος από προφορική συνομιλία:

(.)	μικρή παύση
(_)	μεγάλη παύση
[	ταυτόχρονη ομιλία (στην αρχή εκφωνημάτων που επικαλύπτονται)
-	διακοπή από τον ίδιο τον ομιλητή
(())	σχόλια αναλυτή για τις περιστάσεις της επικοινωνίας π.χ. γέλια, χειροκροτήματα, χειρονομίες κ.λπ.
:::	παράταση/επίσυρση της κατάληξης λέξεων
@@@	φθόγγος ακαθόριστης φωνητικής αξίας
ΚΕΦΑΛΑΙΑ	αύξηση έντασης στη λέξη, συλλαβή κ.λπ.
;	χρήση ερωτηματικού επιτονισμού

<T>	άσχετο ήμουν στο με- στο μετρό μέσα
<M>	ναι
<Λ>	ναι::
<T>	ακούς ΜΑνο;
<M>	ΝΑΙ
<T>	ήτανε ένας μικρούλης μάλλον με τον παππού του τόσος δα τόσος δα ((δείχνει)) και τον είχε πιάσει ένα κλάμα::: (_) μέσα στο μετρό απέναντι και να λέει φοβάμαι φοβάμαι και να κλαίει με λυγμούς [και το'χε πιάσει
<Λ>	[πω::: στο τρένο ε::;
<T>	στο ΜΕΤΡΟ τώρα το βρίσκω λογικό κάπως και::: να μην

	μπορούμε να το σταματήσουμε με τίποτα το πιτσιρίκι (.) έτσι μου'ρθε να του πω κατέβα βρε άνθρωπε κι ανέβα πάνω θα πεθάνει το παιδί από το κλάμα (.) μιλάμε για πολύ κλάμα
<M>	πόσω χρονώ 'τανε;
<T>	τόσο:: ((δείχνει)) δεν ξέρω (.) ήτανε δυόμισι; δύο; μικρούλι κάπου κει (.) δύο χρονών () και τελοσπάντων άρχισα εκεί με το κινητό και του 'βαζα διάφορα να:: ξεχαστεί () κάποια στιγμή σταμάτησε τους λυγμούς αλλά ό,τι κι αν του έλεγες έλεγε (.) φοβάμαι ((γέλια))
<Λ>	((γέλια))
<T>	είχε πολύ γέλιο () ναι αλλά φοβάμαι
<M>	α:::x
<T>	ναι ήτανε ωραίος
<Λ>	σώπα Τάκη σωπάστε @@@@ για ν' ακούσουμε ((δυναμώνει τη μουσική στο κασετόφωνο))

Στο παραπάνω παράδειγμα διακρίνονται μικρές και μεγάλες παύσεις, ανάλογα με τη σχετική διάρκεια των διακοπών που κάνουν οι ομιλητές. (Όπως μπορούμε να εικάσουμε, οι παύσεις αυτές έχουν πραγματικά διαφορετική λειτουργία). Επίσης, στο ακόλουθο σημείο οι T και Λ μιλούν ταυτόχρονα:

<T>	φοβάμαι φοβάμαι και να κλαίει με λυγμούς με λυγμούς [και το'χε πιάσει
<Λ>	[πω:: στο τρένο ε::;

Τα σύμβολα που χρησιμοποιήθηκαν μας δείχνουν ότι η Λ αρχίζει να μιλά μαζί με τον T όταν αυτός λέει *και το'χε πιάσει*. Το τέλος της επικάλυψης δεν είναι ανάγκη να δηλωθεί.

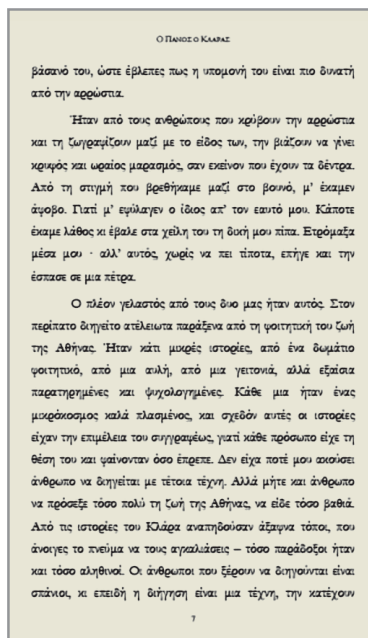
Στον διάλογο περιέχονται σε διπλές αγκύλες (()) τα σχόλια του αναλυτή για τις περιστάσεις της επικοινωνίας όπως π.χ. χειρονομίες με τις οποίες επεξηγεί τα λεγόμενά του ο ομιλητής. Επίσης, στην απομαγνητοφώνηση έχουν δηλωθεί τα σημεία στα οποία ο ομιλητής «τραβάει» τις καταλήξεις των λέξεων π.χ. *να:::*, *πω:::* (προφέρονται *ναααα*, *πωωωω*), σημεία που δηλώνουν δισταγμό, σχεδιασμό του ομιλητή κ.λπ. Στο εκφώνημα της Λ: *σωπάστε @@@@ για ν' ακούσουμε* υπάρχει ένα τμήμα που δεν ακούγεται καθαρά ή δεν μπορεί να γίνει κατανοητό στην απομαγνητοφώνηση και δηλώνεται με @ που επαναλαμβάνονται για όσο περίπου διαρκεί το τμήμα αυτό. Τέλος, το ερωτηματικό (;) χρησιμοποιείται όχι στα σημεία που θα το αναμέναμε από τη συμβατική χρήση των σημείων στίξης αλλά σε όλα τα σημεία που υπάρχει ερωτηματικός επιτονισμός π.χ.

ήτανε *δυόμισι; δύο;*, όπου εκφράζεται η αμφιβολία του ομιλητή με την τυπική ανοδική καμπύλη της φωνής που χαρακτηρίζει τον ερωτηματικό επιτονισμό. (Για περισσότερες λεπτομέρειες για τη μεταγραφή προφορικών κειμενικών δεδομένων, βλ. Meyer 2002: 56-61, 69-78, Zufferey 2020: 152-154 και στα ελληνικά Βασιλοπούλου 2014).

Ανάλογες δυσκολίες προκύπτουν και με τα γραπτά κειμενικά δεδομένα. Καταρχάς, θα πρέπει κανείς να εντοπίσει το κείμενο για το οποίο ενδιαφέρεται λ.χ. σε μια ιστοσελίδα ή σε ένα αρχείο με μορφότυπο pdf. Η χρήση τέτοιου υλικού σε μια μορφή που να μπορεί να υποστεί επεξεργασία και ανάλυση μέσω υπολογιστή δεν είναι πάντοτε απαλλαγμένη από προβλήματα και προϋποθέτει μια διαδικασία μετατροπής, καθαρισμού και τυποποίησης. Η διαδικασία οπτικής αναγνώρισης χαρακτήρων μπορεί επίσης να εισαγάγει λάθη στο κείμενο (βλ. Κεφάλαιο 6), οπότε είναι απαραίτητη η διόρθωση του κειμένου. Από την άλλη, δεν θα πρέπει να επεξεργαζόμαστε γραπτά κειμενικά δεδομένα με αυτόματους διορθωτές λαθών για ενσωμάτωσή τους σε σώματα κειμένων, καθώς πολύτιμες πληροφορίες για τις ορθογραφικές προτιμήσεις των συγγραφέων/ομιλητών μπορεί να χαθούν στη διαδικασία, όπως επίσης και να εισαχθούν καινούργια λάθη από τις επιλογές του αυτόματου διορθωτή. Σε ορισμένες μάλιστα περιπτώσεις, όπως λ.χ. σε σώματα μαθητικών κειμένων, η διατήρηση και επισήμανση των λαθών μπορεί να ανήκει σε έναν από τους κύριους στόχους της δημιουργίας του σώματος κειμένων. Σε αυτή την περίπτωση είναι χρήσιμο να υπάρχει και μια παράλληλη εκδοχή των ίδιων δεδομένων με διόρθωση των λαθών ή σύνδεση των λαθών με τις ορθογραφημένες εκδοχές τους ώστε να είναι εύκολη η ανάκληση των σχετικών τύπων.

Ο καθαρισμός είναι μια απαραίτητη διαδικασία για κάθε είδους μεταφορά υλικού, όπως εύκολα μπορούμε να διαπιστώσουμε. Για παράδειγμα, τα κειμενικά δεδομένα της Εικόνας 7.1 από ένα αρχείο pdf μετατρέπονται με απλή αντιγραφή και επικόλληση στο κείμενο της Εικόνας 7.2, που παρουσιάζεται με εμφάνιση όλων των σημαδιών παραγράφων, μορφοποίησης κ.λπ.

Είναι σαφές ότι, για να μπορεί να εισαχθεί το υλικό της Εικόνας 7.2 σε ένα σώμα κειμένων, πρέπει να καθαριστεί από διάφορα στοιχεία όπως η κεφαλίδα και το υποσέλιδο (που εδώ εμφανίζονται μαζί στην αρχή) και να διορθωθεί ο χαρακτήρας στην 8η σειρά, που για κάποιο λόγο δεν έχει μετατραπεί κατά τη διαδικασία. Ωστόσο, υπάρχουν τουλάχιστον δύο ακόμη σημαντικές διαδικασίες καθαρισμού, που είναι απαραίτητες. Στην 6η σειρά υπάρχουν δύο κενά



Εικόνα 7.1: Έντυπη μορφή κειμένου

Ο ΠΑΝΟ □ Ο ΚΛΑΡΑ □ 7·βάσανό·του, ώστε έβλεπες πως η υπομονή του είναι πιο δυνατή από την αρρώστια. ¶
 Ήταν από τους ανθρώπους που κρύβουν την αρρώστια και τη ζωγραφίζουν μαζί με το είδος των, την βιάζουν να γίνει κρυφός και ωραίος μαρσμός, σαν εκείνον που έχουν τα δέντρα. Από τη στιγμή που βρεθήκαμε μαζί στο βουνό, μ'έκαμεν άφοβο. Γιατί μ'ερώτασεν ο ίδιος απ'τον εαυτό μου. Κάποτε έκαμε λάθος κι έβαλε στα χείλη του τη δική μου πίπα. Ετομάξα μέσα μου... αλλ' αυτός, χωρίς να πει τίποτα, επήγε και την έσπασε σε μια πέτρα. ¶
 Ο πλέον γελαστός από τους δυο μας ήταν αυτός. □ τον περίπατο διηγείτο ατέλειωτα παράξενα από τη φοιτητική του ζωή της Αθήνας. Ήταν κάτι μικρές ιστορίες, από ένα δωμάτιο φοιτητικό, από μια αυλή, από μια γειτονιά, αλλά εξαίσια παρατηρημένες και ψυχολογημένες. Κάθε μια ήταν ένας μικρόκοσμος καλά πλάσμενος, και σχεδόν αυτές οι ιστορίες είχαν την επιμέλεια του συγγραφέως, γιατί κάθε πρόσωπο είχε τη θέση του και φαίνονταν όσο έπρεπε. Δεν είχα ποτέ μου ακούσει άνθρωπο να διηγείται με τέτοια τέχνη. Αλλά μήτε και άνθρωπο να πρόσεξε τόσο πολύ τη ζωή της Αθήνας, να είδε τόσο βαθιά. Από τις ιστορίες του Κλάρα αναπηδούσαν άξαφνα τόποι, που άνοιγες το πνεύμα να τους αγκαλιάσεις—τόσο παραδοξοί ήταν και τόσο αληθινοί. Οι άνθρωποι που ξέρουν να διηγούνται είναι σπάνιοι, κι επειδή η διήγηση είναι μια τέχνη, την κατέχουν. ¶

Εικόνα 7.2: Μετατροπή εικόνας 1 σε κείμενο

διαστήματα πριν την άνω τελεία ανάμεσα στο *μέσα μου* και το *αλλ' αυτός*. Τα κενά διαστήματα, αλλά και χαρακτήρες όπως η ελληνική άνω τελεία, μπορούν να δημιουργήσουν προβλήματα στην εμφάνιση του κειμένου λ.χ. σε ένα λογισμικό δημιουργίας συμφραστικών πινάκων και χρειάζονται ειδικό χειρισμό. Γι' αυτό μια χρήσιμη διαδικασία είναι να αφαιρεθούν τα κενά για να διευκολύνεται η αναζήτηση αυτόματα λ.χ. με *εύρεση και αντικατάσταση* σε ένα πρόγραμμα όπως το Word, όπου στην εύρεση δίνονται δύο κενά διαστήματα και στην Αντικατάσταση ένα κενό διάστημα. Η λειτουργία αυτή επαναλαμβάνεται έως ότου να μην υπάρχουν άλλα κενά. Με παρόμοιο τρόπο αφαιρούμε αυτόματα τυχόν κενά πριν από σημεία στίξης.

Αυτό ισχύει, ωστόσο, και για το κενό που ακολουθεί το σημάδι της παραγράφου ¶. Σε αυτή την περίπτωση θα πρέπει να αποφασίσουμε αν ενδιαφερόμαστε να κρατήσουμε την πληροφορία της παραγραφοποίησης του κειμένου ή όχι. Στην πρώτη περίπτωση, που είναι και η πιο συνηθισμένη, μπορούμε να αντικαταστήσουμε αυτόματα τα σημάδια της παραγράφου με μια επισημείωση που θα δηλώνει το τέλος της προηγούμενης παραγράφου και την αρχή της επόμενης γράφοντας στην *εύρεση* `^p` και στην *αντικατάσταση* `<p>` (με κενά πριν, ανάμεσα και μετά). Αντίθετα, θα πρέπει να αφαιρέσουμε αλλαγές γραμμής ή παραγράφους που έχουν εισαχθεί κατά λάθος από τη διαδικασία μετατροπής όπως το τελευταίο σημάδι παραγράφου στο παραπάνω κείμενο, που έχει εισαχθεί από τη σελιδοποίηση. Με παρόμοιο τρόπο μπορούμε να επισημειώσουμε πληροφορίες για την αλλαγή στροφής ή στίχου σε ένα ποιητικό κείμενο, υποσημειώσεις, λεζάντες εικόνων κ.λπ.¹

1. Οπωσδήποτε, θα πρέπει να έχουμε υπόψη μας ότι η εισαγωγή ετικετών θα εμφανίσει υλικό στα διάφορα λογισμικά κειμενικής ανάλυσης που δεν σχετίζεται με το κείμενο.

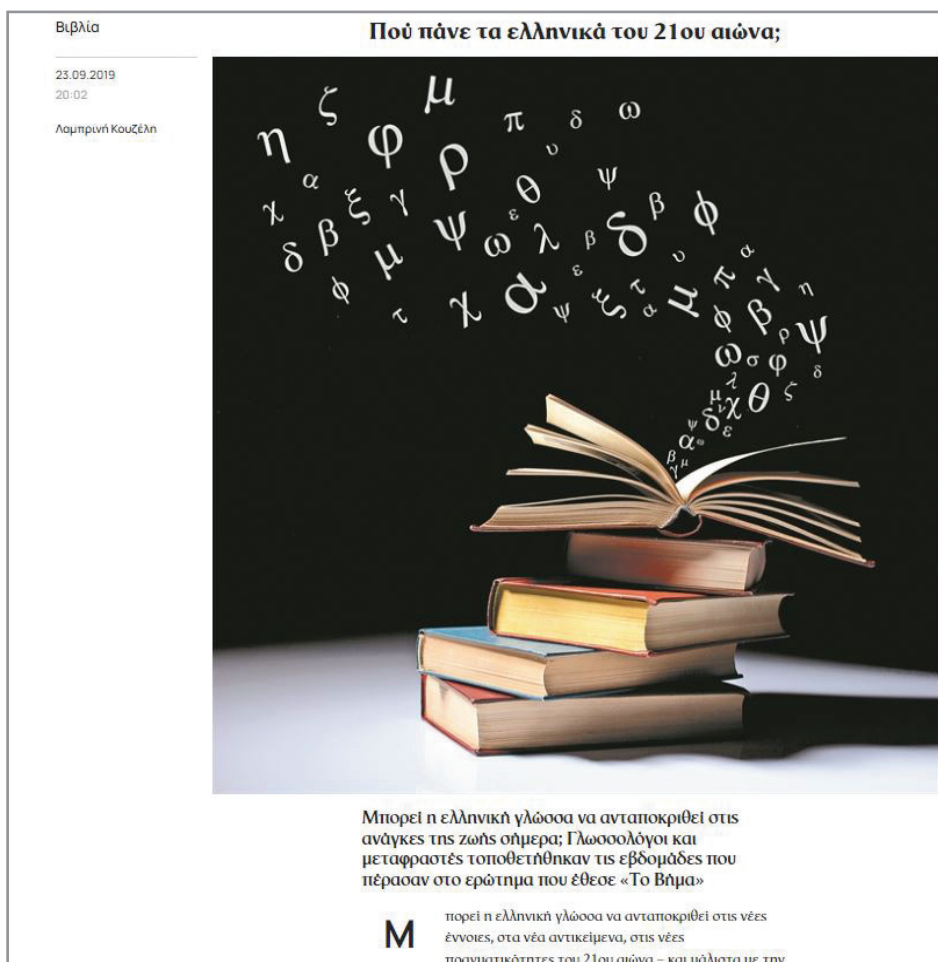
βάσανό του, ώστε έβλεπες πως η υπομονή του είναι πιο δυνατή από την αρρώστια. </p><p>Ήταν από τους ανθρώπους που κρύβουν την αρρώστια και τη ζωγραφίζουν μαζί με το είδος των, την βιάζουν να γίνει κρυφός και ωραίος μαρασμός, σαν εκείνον που έχουν τα δέντρα. Από τη στιγμή που βρεθήκαμε μαζί στο βουνό, μ' έκαμεν άφοβο. Γιατί μ' εφύλαγεν ο ίδιος απ' τον εαυτό μου. Κάποτε έκαμε λάθος κι έβαλε στα χείλη του τη δική μου πίπα. Ετρόμαξα μέσα μου· αλλ' αυτός, χωρίς να πει τίποτα, επήγε και την έσπασε σε μια πέτρα. </p><p>Ο πλέον γελαστός από τους δυο μας ήταν αυτός. Στον περίπατο διηγείτο ατέλειωτα παράξενα από τη φοιτητική του ζωή της Αθήνας. Ήταν κάτι μικρές ιστορίες, από ένα δωμάτιο φοιτητικό, από μια αυλή, από μια γειτονιά, αλλά εξάισια παρατηρημένες και ψυχολογημένες. Κάθε μια ήταν ένας μικρόκοσμος καλά πιασμένος, και σχεδόν αυτές οι ιστορίες είχαν την επιμέλεια του συγγραφέως, γιατί κάθε πρόσωπο είχε τη θέση του και φαίνονταν όσο έπρεπε. Δεν είχα ποτέ μου ακούσει άνθρωπο να διηγείται με τέτοια τέχνη. Αλλά μήτε και άνθρωπο να πρόσεξε τόσο πολύ τη ζωή της Αθήνας, να είδε τόσο βαθιά. Από τις ιστορίες του Κλάρα αναπηδούσαν άξαφνα τόποι, που άνοιγες το πνεύμα να τους αγκαλιάσεις – τόσο παράδοξοι ήταν και τόσο αληθινοί. Οι άνθρωποι που ξέρουν να διηγούνται είναι σπάνιοι, κι επειδή η διήγηση είναι μια τέχνη, την κατέχουν

Εικόνα 7.3: Καθαρισμός κειμένου εικόνας 7.2

Το αποτέλεσμα της διαδικασίας του καθαρισμού μπορούμε να το δούμε στην Εικόνα 7.3.

Παρόμοια ζητήματα ανακύπτουν στον καθαρισμό υλικού από ιστοσελίδες, όπως μπορούμε να δούμε στις Εικόνες 7.4 και 7.5. Εδώ είναι σαφές ότι πρέπει να καθαριστούν πολύ περισσότερα στοιχεία και να ληφθούν αντίστοιχες αποφάσεις για τις πληροφορίες που κρίνεται ότι πρέπει να διατηρηθούν. Για παράδειγμα, οι διάφοροι τίτλοι της στήλης και του άρθρου, όπως και το όνομα της συντάκτριας ή οι πληροφορίες για την ημερομηνία και ώρα δημοσίευσης θα μπορούσαν να τηρηθούν εκτός κειμένου ως μέρος των μεταδεδομένων. Εναλλακτικά, οι τίτλοι μπορούν να παραμείνουν με επισημείωση (λ.χ. με τη μορφή <title> </title> ή <subtitle> </subtitle>).

Σε περίπτωση που θεωρήσουμε ότι η πληροφορία για την ύπαρξη εικόνας είναι απαραίτητη μπορούμε να επισημειώσουμε τη θέση της (λ.χ. με τη μορφή <image> </image>). Το ίδιο ισχύει για άλλα σημεία της εμφάνισης του κειμένου όπως π.χ. την έντονη γραμματοσειρά στον τίτλο πριν από το κυρίως κείμενο, που θα μπορούσε να επισημειωθεί ως εξής: Που πάνε τα ελληνικά του 21ου αιώνα; Γενικά, όπως σημειώνει ο Weisser, «αν και όλες αυτές είναι ενδιαφέρουσες πληροφορίες που συμβατικά αναπαρίστανται μέσα στο έγγραφο και επηρεάζουν τη σελιδοποίηση και την εμφάνισή του, δεν αποτελούν γενικά μέρος της νοηματοδότησης (*meaning potential*) του ίδιου του κειμένου, που αποτελεί αυτό για το οποίο συνήθως ενδιαφερόμαστε περισσότερο όταν αναλύουμε κείμενα από μια γλωσσική οπτική» (2015:



Εικόνα 7.4: Απόκομμα από ιστοσελίδα

34). Το τι χρειάζεται να αφαιρεθεί από τα κειμενικά δεδομένα εξαρτάται από τις συμβάσεις μορφοποίησης ή την κωδικοποίηση του υλικού, ενώ πάντοτε είναι χρήσιμο να τηρούμε αρχείο με τις διαδικασίες επεξεργασίας και καθαρισμού του υλικού τόσο για να μπορούμε να ανακαλέσουμε σε μια μεταγενέστερη φάση πώς ακριβώς χειριστήκαμε το κείμενο (και πιθανόν να αντιστρέψουμε μια λανθασμένη ή περιττή απόφαση) όσο και για να πληροφορήσουμε τους χρήστες του σώματος κειμένων για τις αποφάσεις μας.

Η διαδικασία του καθαρισμού ακολουθείται από την αποθήκευση σε αρχεία, που πρέπει να είναι οπωσδήποτε με τη μορφή Απλού κειμένου (txt) ώστε να διευκολύνονται τα λογισμικά ανάλυσης των δεδομένων. Αν και πολλά λογισμικά όπως το AntConc, που θα δούμε στο επόμενο Κεφάλαιο, ή η πλατφόρμα Sketch Engine,

Πού πάνε τα ελληνικά του 21ου αιώνα;

23.09.2019 20:02

Λαμπρινή Κουζέλη



Μπορεί η ελληνική γλώσσα να ανταποκριθεί στις ανάγκες της ζωής σήμερα; Γλωσσολόγοι και μεταφραστές τοποθετήθηκαν τις εβδομάδες που πέρασαν στο ερώτημα που έθεσε «Το Βήμα»

Μπορεί η ελληνική γλώσσα να ανταποκριθεί στις νέες έννοιες, στα νέα αντικείμενα, στις νέες πραγματικότητες του 21ου αιώνα – και μάλιστα με την αμεσότητα και την ταχύτητα που απαιτεί η ανάγκη επικοινωνίας σε έναν παγκοσμιοποιημένο κόσμο; Ήταν ένα ερώτημα που θέσαμε σε πεπειραμένους μεταφραστές και μεταφράστριες από τρεις μεγάλες

Εικόνα 7.5: Μετατροπή εικόνας 7.4 σε κείμενο

γενικά μπορούν να λειτουργήσουν με πολλά μορφότυπα, είναι καλή ιδέα να αποθηκεύονται τα αρχεία ενός σώματος κειμένων στο μορφότυπο txt με κωδικοποίηση Unicode Utf-8 ώστε να αποφεύγονται τυχόν προβλήματα, ειδικά για τα ελληνικά για τα οποία πολλές κωδικοποιήσεις δεν λειτουργούν.

Στο στάδιο αυτό περιλαμβάνεται και η ονομασία των σχετικών αρχείων. Βασική αρχή είναι ότι κάθε αρχείο θα πρέπει να διαθέτει μια μοναδική ονομασία, που θα μπορεί κατά προτίμηση να ανακαλέσει το είδος του περιεχομένου του. Έτσι, για παράδειγμα στο *Σώμα Κειμένων Ελληνικών Κινηματογραφικών Διαλόγων*, ένα σώμα κειμένων που έχει δημιουργηθεί για τη μελέτη του κινηματογραφικού λόγου στα ελληνικά και συλλέγει διαλόγους ταινιών που έχουν απομαγνητοφωνηθεί για αυτόν το σκοπό, τα αρχεία έχουν την ονομασία της μορφής SFF19_1940_0002, όπου το SFF19 αντιστοιχεί στην ταξινόμηση σε κειμενικά είδη που γίνεται στο *Σώμα Ελληνικών Κειμένων* (και ειδικότερα το S αντιστοιχεί στα προφορικά κείμενα, το F σε κείμενα λογοτεχνίας κ.λπ.), το 1940 αντιστοιχεί στη δεκαετία από την οποία προέρχονται τα δεδομένα και το 0002 στο μοναδικό αριθμό του αρχείου. (Προσέξτε ότι δεν υπάρχουν κενά ανάμεσα στα στοιχεία που συνθέτουν την ονομασία του αρχείου).

Συνδεδεμένο με την ονομασία είναι και το ερώτημα του αν θα πρέπει να δημιουργήσουμε ένα μοναδικό αρχείο για όλο το σώμα κειμένων. Όπως αναφέρει η Zufferey (2020: 149), είναι προτιμότερο να αποθηκεύουμε κάθε γλωσσικό δείγμα σε ένα χωριστό αρχείο, παρά να δημιουργούμε ένα αρχείο για όλο το σώμα κειμένων

ή ένα υπο-σώμα του, για να μπορούμε αργότερα να συνδυάζουμε τα αρχεία με διαφορετικό τρόπο και να δημιουργούμε υπο-σώματα. Για παράδειγμα, αν συλλέγουμε κειμενικά δεδομένα από προφορικές συνομιλίες, μπορούμε κάθε συνομιλία να την αποθηκεύουμε σε ξεχωριστό αρχείο και στην ονομασία του αρχείου να περιλαμβάνονται στοιχεία για τα χαρακτηριστικά των ομιλητών/ομιλητριών λ.χ. Μ για συνομιλίες μεταξύ ανδρών, W για συνομιλίες μεταξύ γυναικών, Χ για συνομιλίες μεταξύ ανδρών και γυναικών, Υ για νέους έως 30 ετών, Ο για ομιλητές/ομιλήτριες πάνω από 30 ετών κ.λπ. Έτσι, μπορούμε εύκολα να απομονώσουμε όλα τα αρχεία συνομιλιών μεταξύ λ.χ. νέων ομιλητριών (FY) και να μελετήσουμε χωριστά τα χαρακτηριστικά της γλωσσικής τους ποικιλίας ή να συγκρίνουμε με άλλες.

Μέρος της τυποποίησης των κειμενικών δεδομένων είναι και ο σχεδιασμός της επισημείωσης για τον χαρακτηρισμό των δομικών μερών του κειμένου (π.χ. παραγράφων, ενοτήτων, τίτλων, στίχων, ομιλητών κ.λπ.), ανάλογα με τις ανάγκες του ψηφιακού έργου, ενώ στην κωδικοποίηση περιλαμβάνεται η τήρηση μεταδεδομένων, στα οποία έχουμε αναφερθεί και σε άλλα κεφάλαια του βιβλίου. Σύμφωνα με τη Drucker (2021: 52-54), τα μεταδεδομένα μπορεί να περιγράφουν τόσο την αρχική πηγή των δεδομένων (π.χ. μια κινηματογραφική ταινία) όσο και το συγκεκριμένο ψηφιακό αρχείο στο οποίο είναι αποθηκευμένα τα δεδομένα (ποιος το δημιούργησε, πότε, σε ποιο μορφότυπο κ.λπ.) και εξυπηρετούν τρεις σκοπούς:

- α) την περιγραφή (descriptive metadata): τα μεταδεδομένα αυτά αφορούν την ονομασία, την ταύτιση και την περιγραφή αντικειμένων ή οντοτήτων όπως τα δεδομένα. Πρόκειται για πληροφορίες απαραίτητες για την ταξινόμηση των δεδομένων έτσι ώστε να ανακαλούνται εύκολα και να μπορούν να χρησιμοποιηθούν στην ανάλυση των δεδομένων.
- β) τη διαχείριση (administrative metadata): τα μεταδεδομένα αυτά αφορούν την ταξινόμηση δεδομένων και αρχείων σύμφωνα με τη χρήση και το είδος τους λ.χ. την προέλευση και το ιστορικό των δεδομένων, τον εμπιστευτικό ή απόρρητο χαρακτήρα τους ή άλλες πληροφορίες που αφορούν τον θεσμικό χειρισμό τους,
- γ) τη λειτουργικότητα (operational metadata): τα μεταδεδομένα αυτά παρέχουν πληροφορίες για τις λειτουργικές απαιτήσεις του αρχείου δεδομένων λ.χ. το λειτουργικό σύστημα που απαιτείται για ένα αρχείο ή ένα νηπτικό, την ανάλυση μιας εικόνας κ.ά.

Ο ευκολότερος τρόπος για να τηρηθούν μεταδεδομένα είναι μέσω ενός λογισμικού υπολογιστικών φύλλων όπως το Excel. Στον παρακάτω πίνακα παρουσιάζεται ένα απόσπασμα από το αρχείο μεταδεδομένων για το *Σώμα Κειμένων Ελληνικών Κινηματογραφικών Διαλόγων*, που αναφέρθηκε πιο πάνω:

ΟΝΟΜΑ ΑΡΧΕΙΟΥ	ΟΝΟΜΑ ΤΑΙΝΙΑΣ	ΕΤΟΣ ΚΥΚΛΟΦΟΡΙΑΣ	ΣΚΗΝΟΘΕΤΗΣ	ΣΕΝΑΡΙΟ	ΔΕΚΑΕΤΙΑ	ΕΙΔΟΣ	ΑΡΙΘΜΟΣ ΛΕΞΕΩΝ	ΟΝΟΜΑ ΜΕΤΑΓΡΑΦΕΑ	ΗΜΕΡΟΜΗΝΙΑ ΚΑΤΑΧΩΡΙΣΗΣ
SFF19_1930_0001	Η προσφυγοπούλα	1932	Mizrahi Togo	Mizrahi Togo Διασκευή: Δημήτρης Μπόγγρης	1930	Δραματική	8.728	ΣΓ	12/03/2021
SFF19_1940_0001	Αδούλωτοι σκλάβοι	1946	Βίων Παπαμικάλης	Βίων Παπαμικάλης	1940	Δραματική	1.780	ΜΛ	20/04/2021
SFF19_1940_0002	Οι Γερμανοί ξανάρχονται	1948	Αλέκος Σακελλάριος	Αλέκος Σακελλάριος, Χρήστος Γιαννακόπουλος	1940	Κωμωδία	8.671	ΜΛ	20/04/2021

Εναλλακτικά, σε πολλά σώματα κειμένων πληροφορίες για τα μεταδεδομένα τηρούνται στην αρχή κάθε αρχείου με επισημείωση που περιλαμβάνει τις γωνιώδεις αγκύλες, τις οποίες αναφέραμε πιο πάνω, έτσι ώστε να απομονώνονται εύκολα από το λογισμικό κειμενικής ανάλυσης (βλ. Γούτσος & Φραγκάκη 2015: 44-46 για σχετική συζήτηση).

Συνολικά, μπορούμε να ελέγξουμε τη διαδικασία συγκρότησης ενός σώματος κειμένων με βάση την ακόλουθη λίστα ελέγχου (βλ. Drucker 2021: 31):

- ➡ Είναι αποθηκευμένα όλα τα δεδομένα με το ίδιο μορφότυπο;
- ➡ Υπάρχει συνέπεια στον τρόπο με τον οποίο ακολουθούνται οι συμβάσεις;
- ➡ Είναι αρκετές οι πληροφορίες που αποθηκεύονται για κάθε αρχείο;
- ➡ Μήπως υπάρχουν περιττές πληροφορίες;
- ➡ Έχουν τηρηθεί τα απαραίτητα μεταδεδομένα;
- ➡ Υπάρχουν διαθέσιμες σαφείς οδηγίες για τις διαδικασίες που ακολουθήθηκαν;

7.3. ΕΙΔΗ ΣΩΜΑΤΩΝ ΚΕΙΜΕΝΩΝ

Έχουμε ήδη αναφερθεί σε ορισμένα από τα βασικότερα είδη σωμάτων κειμένων. Στην ενότητα αυτή συνοψίζουμε την κατηγοριοποίηση των σωμάτων κειμένων έτσι ώστε να διαπιστώσουμε την ποικιλία στον σχεδιασμό και τη συγκρότησή τους. Ακο-