Αγγελική Ράλλη

ΘΕΩΡΗΤΙΚΗ ΓΛΩΣΣΟΛΟΓΙΑ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ*

Εισαγωγή

Η επιστήμη της υπολογιστικής γλωσσολογίας είναι σχετικά καινούρια και η έρευνα σ' αυτό τον χώρο βρίσκεται σε εξέλιξη. Στη διαμόρφωση των αρχών και μεθόδων της υπολογιστικής γλωσσολογίας έχουν συμβάλει κυρίως οι επιστήμες της πληροφορικής, της θεωρητικής γλωσσολογίας, των μαθηματικών και της λογικής. Σ' αυτό το άρθρο, θα αναφερθώ στις αρχές της θεωρητικής γλωσσολογίας που χρησιμοποιούνται από την υπολογιστική γλωσσολογία και θα προσπαθήσω να δείξω πόσο σημαντική είναι η συμβολή της γνώσης που προσφέρεται από την πρώτη στην ανάπτυξη υπολογιστικών συστημάτων ικανών να αναλύουν και να συνθέτουν επαρκώς δεδομένα της φυσικής γλώσσας.

1. Ορισμός και σύγκριση υπολογιστικής και θεωρητικής γλωσσολογίας

Ως υπολογιστική γλωσσολογία (στο εξής ΥΓ) ορίζεται η επιστήμη που μελετά και δημιουργεί συστήματα ικανά να κατανοούν, να αναλύουν και να συνθέτουν τη φυσική γλώσσα. Τα συστήματα αυτά χρησιμοποιούνται όχι μόνο σε τομείς εφαρμογών (π.χ. μηχανική μετάφραση, ανάκτηση πληροφοριών από το περιεχόμενο ενός κειμένου κ.λπ.) αλλά και σε καθαρά επιστημονικό επίπεδο, ανεξάρτητα από συγκεκριμένες εφαρμογές, όπως για παράδειγμα, στην αναπαράσταση γνώσης, ή ως μέσο επιβεβαίωσης των κανόνων και των μοντέλων που προτείνονται από τη θεωρητική γλωσσολογία για την ανάλυση και τη σύνθεση γραμματικών προτάσεων.

Η θεωρητική γλωσσολογία (στο εξής ΘΓ) ορίζεται συνήθως ως η επιστήμη της γλώσσας. Ειδικότερα, η σύγχρονη θεωρητική γλωσσολογία ενδιαφέρεται όχι μόνο για την εξαντλητική περιγραφή των φαινομένων της γλώσσας αλλά και για την αναπαράσταση της γλωσσικής ικανότητας των ομιλητών. Δηλαδή, με τον καθορισμό καθολικών σχημάτων, προσπαθεί να ορίσει τον μηχανισμό της γλώσσας γενικά, ανεξάρτητα από συγκεκριμένες γλώσσες, καθώς και να εντοπίσει τις παραμέτρους που διέπουν τις διαφορές των γλωσσών.

Οι δύο επιστημονικοί τομείς συγκλίνουν στο αντικείμενο μελέτης: την

* Μία πρώτη μορφή αυτής της εργασίας παρουσιάσθηκε στον κύκλο σεμιναρίων για τις χρήσεις των υπολογιστών του Τμήματος Πληροφορικής του Πανεπιστημίου Αθηνών, τον Δεκέμβριο του 1992.

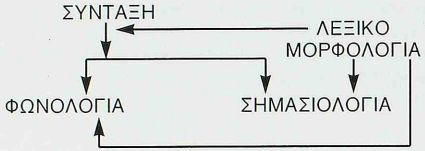
κατανόηση της γλωσσικής διαδικασίας. Η ΥΓ ενδιαφέρεται για τη δημιουργία έξυπνων υπολογιστικών συστημάτων ούτως ώστε, αναλύοντας όσο το δυνατόν περισσότερες προτάσεις και κείμενα, να καταστεί δυνατή η κατανόηση της διαδικασίας της παραγωγής της γνώσης. Τη ΘΓ ενδιαφέρει η κατανόηση και η αναπαράσταση της γλωσσικής ικανότητας σύμφωνα με την οποία, οι ομιλητές μιας γλωσσικής κοινότητας παράγουν απεριόριστο αριθμό προτάσεων από ένα μικρό αριθμό στοιχείων. Για τον σκοπό αυτό, η ΘΓ προσπαθεί να δημιουργήσει μια θεωρία που να αποδεικνύεται αληθινή για τις διάφορες γλώσσες. Στα πλαίσια της θεωρίας, κάνει χρήση φορμαλισμών και μοντέλων τυποποίησης· αναπτύσσει γραμματικές οι οποίες πρέπει να είναι απλές και γενικές για να καλύπτουν με τον πιο οικονομικό τρόπο τις γραμματικές φράσεις και να απορρίπτουν τις μη-γραμματικές.

Οι δύο τομείς παρουσιάζουν αρκετές ομοιότητες στα εργαλεία της έρευνας και στη χρησιμοποιουμένη μεθοδολογία. Για παράδειγμα, η χρήση μιας γραμματικής και ενός μοντέλου τυποποίησης (π.χ. γραμματικές ανεξάρτητες από το περιβάλλον, γραμματικές εξαρτημένες από το περιβάλλον κ.λπ.) της ΘΓ διευκολύνουν τον σχεδιασμό και την υλοποίηση των υπολογιστικών συστημάτων. Τόσο στην υπολογιστική όσο και στη θεωρητική γλωσσολογία καταβάλλεται προσπάθεια να αναλυθεί ένας μεγάλος αριθμός γραμματικών προτάσεων ενώ απορρίπτονται οι μη-γραμματικές. Η απλότητα των εκφράσεων στη διατύπωση των κανόνων, η επεκτασιμότητα του συστήματος και η αξιολόγηση των μοντέλων βασισμένη στην επαλήθευση με καινούριες προτάσεις και κείμενα αποτελούν βασικές επιδιώξεις και των δύο τομέων. Τέλος, ο καταμερισμός της επεξεργασίας των γλωσσικών φαινομένων σε διάφορα επίπεδα (modularity), ανάλογα με τη φύση των φαινομένων, ελαχιστοποιεί τα προβλήματα και διαχωρίζει τις δυσκολίες αντιμετωπίζοντας τη μία ανεξάρτητα από την άλλη.

2. Εφαρμογές της γλωσσολογικής ανάλυσης σε συστήματα επεξεργασίας φυσικής γλώσσας

Στις περισσότερες σύγχρονες σχολές της γλωσσολογίας, η γλωσσολογική ανάλυση των γλωσσικών φαινομένων γίνεται με βάση μία συγκεκριμένη θεωρία και διακρίνεται σε τέσσερα επίπεδα: το λεξικό που στις περισσότερες περιπτώσεις περιέχει και τη μορφολογία, το συντακτικό, το φωνολογικό και το σημασιολογικό. Για κάθε επίπεδο υπάρχουν οι βασικές μονάδες ανάλυσης, οι κανόνες και οι αρχές, χωρίς να αποκλείεται χρησιμοποίηση των ίδιων αρχών σε περισσότερα του ενός επίπεδα. Ανάμεσα στα επίπεδα υπάρχει αλληλεπίδραση και η μετάβαση από το ένα στο άλλο γίνεται σύμφωνα με το σχήμα που δίνεται στο (1):

(1)



Για την Ελληνική γλώσσα, υπάρχουν αρκετές γλωσσολογικές μελέτες για τη σύνταξη, ενώ ο αριθμός δημοσιεύσεων για τα άλλα τρία επίπεδα είναι αρκετά περιορισμένος. Το αντίθετο συμβαίνει με τις δημοσιεύσεις στον χώρο της ΥΓ όπου μόνο στη φωνολογία και στη μορφολογία, υπάρχουν τα πρώτα και αντιπροσωπευτικότερα δείγματα υπολογιστικής επεξεργασίας με αντικείμενο μελέτης την Ελληνική γλώσσα[1]. Αναφερόμενη λοιπόν σε κάποιες από αυτές τις δημοσιεύσεις, θα προσπαθήσω να δείξω τη σημαντική βοήθεια που μπορεί να προσφέρει η ΘΓ στην υλοποίηση, πληρότητα και επεκτασιμότητα ενός συστήματος επεξεργασίας φυσικής γλώσσας.

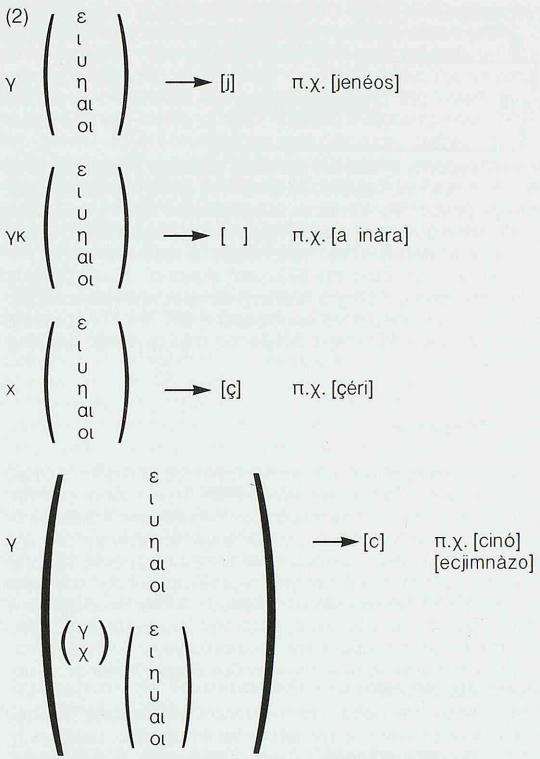
2.1 Φωνολογική επεξεργασία

Στα πλαίσια ενός προγράμματος για σύνθεση φωνής από ηλεκτρονικό υπολογιστή, οι Μπακαμίδης και Καραγιάννης (1987) δημοσίευσαν ένα άρθρο (βλ. Βιβλιογραφία) στο οποίο περιγράφουν ένα σύνολο κανόνων αντιστοιχίας μεταξύ των χαρακτήρων της γραφής και των φωνημάτων της φωνολογίας της Νέας Ελληνικής. Σύμφωνα με τους συγγραφείς, αυτοί οι κανόνες μπορούν να χρησιμοποιηθούν για τη μετατροπή ενός κειμένου από τη γραπτή μορφή στην προφορική του εκφορά. Οι ίδιοι υποστηρίζουν επίσης ότι το περιεχόμενο του άρθρου εκφράζει την άποψη του μηχανικού και αμφισβητώντας έμμεσα τη συμβολή της γλωσσολογικής ανάλυσης, παραδέχονται ότι για τη διατύπωση των κανόνων δεν στηρίχθηκαν σε γλωσσολογικά δεδομένα (βλ. σελ. 168).

Ας εξετάσουμε, λοιπόν, την ορθότητα της παραπάνω άποψης. Μεταξύ των φωνολογικών φαινομένων που αναφέρονται στο άρθρο, υπάρχει το φαινόμενο της ουρανικοποίησης των υπερωικών συμφώνων. Περιγράφεται με τα ακόλουθα τέσσερα γενικά σχήματα τα οποία περικλείουν έναν εξαιρετικά υψηλό αριθμό υποπεριπτώσεων:

---

1. Με εξαίρεση το σύστημα αυτόματης μηχανικής μετάφρασης EUROTRA, στο οποίο υπάρχει υπολογιστική κάλυψη της σύνταξης της Νέας Ελληνικής με βάση τον συγκεκριμένο φορμαλισμό του προγράμματος.



Στο άρθρο, τα παραπάνω σχήματα εμφανίζονται ως νόμοι εξαρτημένοι από το περιβάλλον (context-dependent rules), υπεύθυνοι για την ουρανικοποιημένη προφορά των υπερωικών χαρακτήρων 'γ', 'χ', 'γκ', 'κ' σε κατάλληλο γλωσσικό περιβάλλον[2]. Ως πρώτη παρατήρηση για το είδος αυτών των νόμων, θα μπορούσα να αναφέρω τη μη-γλωσσολογική τους

2. Όπως γίνεται αντιληπτό, τα δίψηφα φωνήεντα 'υι' και 'ει' δεν αποτελούν μέρος της δομικής περιγραφής των κανόνων. Η εξήγηση πρέπει να αναζητηθεί στο ότι τα συγκεκριμένα φωνήεντα έχουν ως πρώτο χαρακτήρα τα 'υ' και 'ε' αντιστοίχως, τα οποία ήδη αποτελούν μέρος των κανόνων. Πρόκειται για μια ιδιαίτερα μηχανιστική αποτύπωση της γραφής της Νέας

υπόσταση. Η περιγραφή τους περικλείει στο αριστερό μέρος γραφήματα και στο δεξί φωνήματα. Γραφή και φωνολογία όμως δεν πρέπει να αναμειγνύονται συγχρονικά αφού αποτελούν διαφορετικούς κώδικες αναπαράστασης των ήχων της γλώσσας. Η μόνη αλληλεπίδρασή τους αφορά σε κάποιες απλές αντιστοιχίες γραφημάτων και φωνημάτων όπως είναι, για παράδειγμα, οι ακόλουθες:

(3) Γραφήματα Φωνήματα

{ο, ω} = /o/

{ε, αι} = /e/[3]

π = /p/

ψ = /ps/

κ.λπ.

Θα μπορούσε βέβαια να ισχυρισθεί κανείς ότι οι νόμοι στο (2) χαρακτηρίζονται από περιγραφική επάρκεια αφού στο άρθρο τους, οι Μπακαμίδης, Καραγιάννης ενδιαφέρονται μόνο για την αντιστοιχία γραφημάτων και ουρανικοποιημένων υπερωικών ήχων και όχι να περιγράψουν γλωσσολογικά το φαινόμενο της ουρανικοποίησης. Ο ισχυρισμός όμως αυτός καταρρίπτεται από το ίδιο το αντικείμενο μελέτης, τη γλώσσα, η οποία δεν είναι ένας απλός κατάλογος πραγμάτων που μπορεί να μελετηθεί διεξοδικά με μια απλή περιγραφή. Αλλά και αν ακόμα η μέλετη εξαντλείται στην περιγραφή, η περιγραφική επάρκεια πρέπει να συνοδεύεται από απλότητα στις εκφράσεις και από καταγραφή των γενικεύσεων όπου αυτό είναι δυνατό. Αυτά τα τελευταία χαρακτηριστικά λείπουν από τους νόμους του (2).

Ας δούμε τώρα πώς θα μπορούσε να περιγραφεί το φαινόμενο της ουρανικοποίησης των υπερωικών συμφώνων και να καταγραφούν οι αντιστοιχίες γραφημάτων και φωνημάτων μέσα στα πλάισια μιας γλωσσολογικής ανάλυσης. Στη σύγχρονη γλωσσολογική ανάλυση, τα φωνήματα μιας γλώσσας δεν αποτελούν ατομικά στοιχεία ανεξάρτητα το ένα από το άλλο. Σύμφωνα με κάποια προκαθορισμένα χαρακτηριστικά, τα οποία ορίζο-

---

Ελληνικής, δεδομένου ότι τουλάχιστον στην περίπτωση των 'ε' και 'ει' οι χαρακτήρες αντιστοιχούν σε εντελώς διαφορετικά φωνήματα (/e/ και /i/ αντιστοίχως).

3. Η αντιστοιχία 'αι' = /e/ αναιρείται από την παρουσία του τόνου και των διαλυτικών. Στα δίψηφα φωνήεντα, τόσο ο τόνος όσο και τα διαλυτικά αποτελούν στοιχεία που υπολογιστικά λαμβάνονται υπ' όψιν ανεξάρτητα από τους χαρακτήρες τους οποίους συνοδεύουν.

νται κυρίως με βάση τις ιδιότητες της άρθρωσης των φωνημάτων μέσα στη στοματική κοιλότητα, τα φωνήματα περιγράφονται ως σύνολα χαρακτηριστικών (βλ. Chomsky & Halle, 1968)[4]. Τα κοινά χαρακτηριστικά αναπαριστάνουν τις ομοιότητες μεταξύ των φωνημάτων ενώ τα διαφορετικά τις διαφορές τους. Για παράδειγμα, τα υπερωικά σύμφωνα μοιάζουν ως προς τα χαρακτηριστικά [+ σύμφωνο, - πρόσθιο, - ουρανικό] ενώ μεταξύ τους διαφέρουν ως προς το χαρακτηριστικό [± συνεχές] ([+ συνεχές] για τα /x/, /γ/ και [- συνεχές] για τα /k/ και /g/). Τα φωνήεντα /e/ και /i/ συγκλίνουν ως προς τα χαρακτηριστικά [+ φωνήεν, + πρόσθιο, - στρογγυλό, - χαμηλό] και διαφέρουν μεταξύ τους ως προς το χαρακτηριστικό [± υψηλό] ([+ υψηλό] για το /i/ και [- υψηλό] για το /e/).[5] Σημειωτέον ότι η περιγραφή των φωνολογικών νόμων που είναι υπεύθυνοι για τις φωνολογικές αλλαγές και τις αλλοφωνικές πραγματώσεις των φωνημάτων γίνεται με τη βοήθεια των χαρακτηριστικών. Μ' αυτόν τον τρόπο, επιτυγχάνονται οι στόχοι όχι μόνο της περιγραφικής επάρκειας αλλά και της γενίκευσης χωρίς οι εκφράσεις να χάνουν σε απλότητα. Γλωσσολογικά, το φαινόμενο της ουρανικοποίησης των υπερωικών συμφώνων δίνεται από ένα μόνο νόμο εξαρτημένο από το περιβάλλον (context-dependent) στην αριστερή πλευρά του οποίου περιγράφονται τα υπερωικά σύμφωνα, ενώ στο περιβάλλον περιγράφονται τα φωνήεντα /e/ και /i/:

(4)

$$\begin{bmatrix} +\text{ σύμφωνο} \\ -\text{ πρόσθιο} \\ -\text{ ουρανικό} \end{bmatrix} \rightarrow [+\text{ ουρανικό}] \;/\; \_\_ \; \begin{bmatrix} +\text{ φωνήεν} \\ -\text{ χαμηλό} \\ -\text{ στρογγυλό} \end{bmatrix}$$

Πώς θα ήταν, λοιπόν, και τι θα προσέφερε μία υπολογιστική επεξεργασία της πραγμάτωσης των ουρανικοποιημένων υπερωικών, στο γλωσσικό περιβάλλον που περιγράφεται από τους Μπακαμίδη, Καραγιάννη, με βάση τη γνώση που μας παρέχεται από τη γλωσσολογική ανάλυση; Κατ' αρχήν, σε δύο ξεχωριστά αρχεία του συστήματος θα έπρεπε να είχαν καταχωρηθεί: α) οι απλές αντιστοιχίες γραφημάτων - φωνημάτων και β) οι αναπαραστάσεις φωνημάτων με τη μορφή χαρακτηριστικών, έτσι όπως προσφέρονται από τη γλωσσολογική ανάλυση της Ελληνικής. Για παράδειγμα, απλές αντιστοιχίες του πρώτου αρχείου δίνονται στο (3), ενώ οι πλήρεις φωνολογικές αναπαραστάσεις για τα /e/ και /i/ είναι οι ακόλουθες:

---

4. Μια φωνολογική ανάλυση στα πλαίσια της σύγχρονης γλωσσολογίας δεν περιορίζεται, σήμερα, στην ανάλυση των φωνημάτων ως σύνολα χαρακτηριστικών. Χρησιμοποιεί επίσης στοιχεία συλλαβισμού και προσωδίας (για τα Ελληνικά, βλ. Nespor & Vogel, 1986 και Μαλικούτη - Drachman & Drachman, 1988). Παρ' όλα αυτά, για τις θέσεις που υποστηρίζονται στην παρούσα εργασία, είναι αρκετή η παραπομπή στα φωνολογικά χαρακτηριστικά.

5. Μία ανάλυση των φωνημάτων της Νέας Ελληνικής συγκεκριμένα του Αιτωλικού ιδιώματος, με βάση τα διακριτικά χαρακτηριστικά, προσφέρεται στη διατριβή του Π. Κοντού, 1990 (βλ. Βιβλιογραφία).

(5) /e/ : [+ φωνήεν, - σύμφωνο, - υψηλό, - χαμηλό, + πρόσθιο, - στρογγυλό] /i/ : [+ φωνήεν, - σύμφωνο, + υψηλό, + πρόσθιο, - στρογγυλό]

Για την ενεργοποίηση του φαινομένου της ουρανικοποίησης, υπεύθυνος είναι ο νόμος της ουρανικοποίησης που δίνεται στο (4) και ο οποίος θα έπρεπε να βρίσκεται καταχωρημένος σε ένα τρίτο αρχείο μαζί με τους άλλους φωνολογικούς νόμους της Ελληνικής. Αξίζει να σημειωθεί ότι η υπολογιστική κάλυψη των τριών αυτών αρχείων είναι αρκετά εύκολη να υλοποιηθεί[6]. Επιπλέον, παρατηρούμε ότι ένας και μόνον νόμος καλύπτει όλες τις ουρανικοποιήσεις των υπερωικών συμφώνων, ενώ στην ανάλυση των Μπακαμίδη, Καραγιάννη χρειάζονται τέσσερις νόμοι με πληθώρα υποπεριπτώσεων. Επομένως, ένα υπολογιστικό σύστημα που βασίζεται σε γλωσσολογική ανάλυση μπορεί να είναι οικονομικό και κερδίζει όχι μόνο σε περιγραφική επάρκεια αλλά και σε επεξηγηματική πληρότητα.

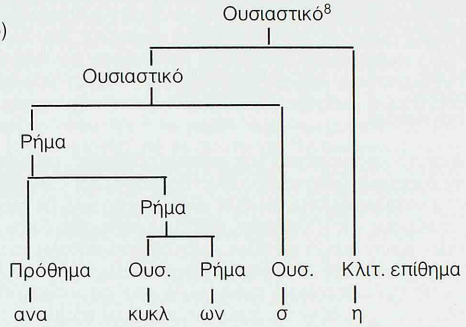
2.2 Μορφολογική επεξεργασία

Η ανάλυση μιας μορφολογικά σύνθετης λέξης συνίσταται στην αναγνώριση και επεξεργασία των επιμέρους συστατικών της λέξης. Συνήθως, η ανάλυση γίνεται μετά από την εφαρμογή μορφολογικών νόμων επαναγραφής ανεξάρτητων από το περιβάλλον (context-free rewriting rules) και λαμβάνει τη μορφή δένδρου που αναπαριστάνει την εσωτερική δομή της λέξης. Π.χ. η λέξη 'ανακύκλωση' αναλύεται ως εξής[7]:

6. Σ' ό,τι αφορά στην υπολογιστική επεξεργασία των φωνημάτων με τη μορφή χαρακτηριστικών και την υπολογιστική κάλυψη των φωνολογικών νόμων, βλ. Galiotou & Ralli, 1991.

7. Για την υπολογιστική επεξεργασία του τονισμού των κλιτών λέξεων, βλ. Touratzidis & Ralli, 1992.

(6)



Οι ίδιοι νόμοι της ανάλυσης χρησιμοποιούνται και στη σύνθεση για τον σχηματισμό νέων λέξεων αφού έχουν «γενετικές» δυνατότητες (generative power). Η σειρά με την οποία συνδυάζονται τα διάφορα συστατικά καθορίζεται από βασικές αρχές της μορφολογίας. Για παράδειγμα, σε ένα μονολεκτικό σχηματισμό, η κλίση ακολουθεί συνήθως την παραγωγή. Επίσης, οι πληροφορίες των λημμάτων του γλωσσολογικού λεξικού (καταχωρημένα με τη μορφή μορφημάτων) και οι νόμοι σχηματισμού λέξεων καθορίζουν εάν ένας συνδυασμός μορφημάτων είναι επιτρεπτός στη γλώσσα, σε αντίθεση με κάποιον άλλο που καταλήγει σε μη-γραμματική δομή. Έτσι, το επίθημα '-σ(η)' προστίθεται πάντοτε σε ρηματική βάση για να δώσει ουσιαστικό, ενώ το επίθημα '-ών(ω)' συνδυάζεται με ουσιαστικό για να δώσει ρήμα[9].

Σ' ό,τι αφορά στην υπολογιστική επεξεργασία της μορφολογίας της Ελληνικής, όλες οι μέχρι τώρα προσπάθειες περιορίζονται στην κλίση, κυρίως για τρεις λόγους:

α) Επικρατεί συνήθως η άποψη ότι στην υπολογιστική επεξεργασία των λέξεων χρήσιμη είναι μόνον η κλίση.

β) Το φαινόμενο της κλίσης εμφανίζει τις λιγότερες δυσκολίες σε σχέση με την παραγωγή και τη σύνθεση.

γ) Περιγράφεται διεξοδικά σε όλες τις παραδοσιακές σχολικές γραμματικές, σε αντίθεση με την παραγωγή και τη σύνθεση των οποίων η περιγραφή εμφανίζει σημαντικά κενά.

Μία υπολογιστική επεξεργασία της κλίσης σημαίνει ότι το σύστημα μπο-

8. Μπορούμε να θεωρήσουμε ότι το παραγωγικό επίθημα '-ων-' χάνει το σύμφωνο 'ν' πριν από το σίγμα με την εφαρμογή λεξικού φωνολογικού νόμου. Εναλλακτικά, είναι δυνατόν να υιοθετηθεί η άποψη των δύο αλλομόρφων '-ων-' και '-ω-', κατά την οποία, στη λέξη 'ανακύκλωση' εμφανίζεται η μορφή '-ω-'.

9. Εντός παρενθέσεως εμφανίζονται τα κλιτικά επιθήματα.

ρεί και διαχωρίζει την κατάληξη (κλιτικό επίθημα) από το θέμα αλλά δεν είναι σε θέση να προχωρήσει σε περαιτέρω ανάλυση του θέματος στα επιμέρους συστατικά του. Δηλαδή, στο παράδειγμα της λέξης 'ανακύκλωση' αναγνωρίζεται η κατάληξη '-η' αλλά το θέμα 'ανακύκλωσ-' λαμβάνεται ως αδιαίρετη ενότητα.

Θεωρώ ότι μία μορφολογική επεξεργασία από υπολογιστή που περιορίζεται στην ανάλυση και σύνθεση του φαινομένου της κλίσης είναι ελλιπής. Κατά την άποψή μου, εάν για την ανάπτυξη των περισσοτέρων μορφολογικών επεξεργαστών της Νέας Ελληνικής είχαν ληφθεί υπ' όψιν τα δεδομένα της γλωσσολογικής έρευνας στον τομέα της παραγωγής και της σύνθεσης, οι φερόμενοι ως μορφολογικοί επεξεργαστές θα υπερείχαν σε δυνατότητες και επιστημονική πληρότητα.

Στην προσπάθειά μου να υποστηρίξω και πρακτικά την παραπάνω θέση, με βοήθησε ο πληροφορικός Αντώνης Δραγγιώτης, ο οποίος ανέπτυξε έναν επεξεργαστή της κλίσης που δεν υστερεί σε υπολογιστική κάλυψη κλιτικών φαινομένων από τα ήδη υπάρχοντα συστήματα της μορφολογίας τα οποία επίσης καλύπτουν μόνο την κλίση[10]. Το υπολογιστικό σύστημα του Δραγγιώτη κάνει κατ' αρχήν γραμματική αναγνώριση της κλίσης σε 700.000 λέξεις, δηλαδή, αναλύει τις λέξεις σε θέμα και κατάληξη προχωρώντας, συγχρόνως, στον μορφοσυντακτικό χαρακτηρισμό τους (κατηγορία, πτώση, αριθμός, χρόνος, πρόσωπο κ.λπ., ανάλογα με την περίπτωση). Στη συνέχεια, επανασυνθέτει τις ήδη αναγνωρισμένες λέξεις, ενώ είναι σε θέση να δώσει και τα κλιτικά τους παραδείγματα, αφού το σύστημα σύνθεσης του επεξεργαστή μπορεί και δημιουργεί όλες τις κλιτές μορφές ενός συγκεκριμένου θέματος.

Με τη βοήθεια αυτού του συστήματος, υποβλήθηκε σε έλεγχο γραμματικής αναγνώρισης μία έκδοση της εφημερίδας «Το Βήμα» που περιείχε 106.034 λέξεις. Εκτός των λέξεων που δεν αναγνωρίσθηκαν είτε γιατί στην εφημερίδα περιείχαν ορθογραφικά σφάλματα, είτε γιατί τα λήμματά τους, που, αν και κοινά, δεν είχαν ακόμη καταχωρηθεί στο λεξικό του συστήματος, υπήρξε και ένας σημαντικός αριθμός λέξεων που θεωρήθηκαν άγνωστες γιατί ο επεξεργαστής δεν ελάμβανε υπ' όψιν τα φαινόμενα της παραγωγής και της σύνθεσης. Δεν αναγνωρίσθηκαν, δηλαδή, νεολογισμοί που αποτελούσαν σύνθετες και παράγωγες λέξεις. Για παράδειγμα, σύνθετες λέξεις που περιείχαν ως πρώτο συνθετικό θέματα όπως, 'ψευδο-' ('ψευδοδημοκράτες'), 'χρονο-' ('χρονομίσθωση', 'χρονοντούλαπο') 'ραδιο-' ('ραδιοδέκτης'), 'πρωτο-' ('πρωτοπαρουσίασε'), 'φιλο-' ('φιλοσοβιετικής'), 'κοινοτικο-' ('κοινοτικοποίηση'), 'κρατικο-' ('κρατικοδίαιτη'), 'νεο-' ('νεο-

10. Ο συγκεκριμένος επεξεργαστής της κλίσης αποτελεί μέρος ενός συντακτικού επεξεργαστή των Ελληνικών που επεξεργάζεται πληροφορίες σε σχέση με τη δομή και τις λειτουργίες των συστατικών μιας πρότασης. Για τους σκοπούς ενός συντακτικού επεξεργαστή, η μορφολογική ανάλυση της παραγωγής και της σύνθεσης δεν αποτελεί άμεση προτεραιότητα. Η επεξεργασία της κλίσης είναι αρκετή για να προσφέρει τις μορφολογικές πληροφορίες που χρειάζεται η σύνταξη.

κομμουνιστών'), 'ξανα-' ('ξαναπλησιάσει'), 'ανωτατο-' ('ανωτατοποίηση') κ.λπ., καταχωρήθηκαν ως άγνωστες λέξεις για το σύστημα.[11] Επίσης, δεν αναγνωρίσθηκαν παράγωγες λέξεις όπως 'εκβαρβαριστής', 'ευτροφισμός', 'υποτίτληση', 'κρατικίστικο', 'παραγοντίσκος', 'σοφόκλεια', 'λαϊκίζει', κ.λπ. Αξίζει να σημειωθεί ότι οι παραπάνω λέξεις δεν περιέχονται συνήθως στα κοινά λεξικά. Αποτελούν σχηματισμούς που παράγονται από τον μηχανισμό της γλώσσας με βάση τους κανόνες της και τις αρχές της. Σε ένα μορφολογικό επεξεργαστή που λαμβάνει υπ' όψιν μόνο την κλίση, είναι αδύνατον να ληφθούν υπ' όψιν όλοι οι πιθανοί σχηματισμοί και να καταχωρηθούν ως ανεξάρτητα λήμματα, ακόμα και αν το λεξικό του συστήματος εμπλουτίζεται σε καθημερινή βάση. Ας υποθέσουμε ότι στον επεξεργαστή καταχωρείται ως νέο λήμμα το θέμα 'λαϊκίστικ-' ('λαϊκίστικος' χωρίς την κατάληξη). Αν ο επεξεργαστής δεν είναι σε θέση να κάνει χρήση των μηχανισμών της παραγωγής και της σύνθεσης, τότε, το κέρδος είναι μηδαμινό γιατί συγχρόνως σε άλλα κείμενα είναι δυνατόν να εμφανισθούν οι λέξεις 'ψευδολαϊκίστικος', 'αντιλαϊκίστικος', 'υπερλαϊκίστικος' κ.λπ., καθώς και τα 'λαϊκίζω', 'λαϊκιστής', 'αντιλαϊκιστής', 'φιλολαϊκιστής', 'λαϊκούρα', 'υπερλαϊκισμός', 'εκλαϊκισμός' κ.λπ. Δεν είναι δυνατόν να προβλέψει κανείς τον σχηματισμό και τη χρήση όλων των πιθανών να παραχθούν λέξεων. Καταχώρηση νέων λημμάτων πρέπει να γίνεται βέβαια στην περίπτωση των δανείων (π.χ. η λέξη 'φιλμ') ή των ξένων κυρίων ονομάτων όπως 'Γκένσερ'[12]. Στη συνέχεια, ο μηχανισμός της παραγωγής μπορεί να αναγνωρίσει και να αναλύσει πιθανά παράγωγα αυτών των λέξεων (π.χ. 'φιλμάκι', 'γκενσερισμός', 'γκενσεριστής' κ.λπ.) χωρίς να χρειάζεται πρόσθετος εμπλουτισμός του λεξικού μ' αυτά. Βλέπουμε λοιπόν ότι μορφολογική επεξεργασία μιας λέξης δεν σημαίνει μόνο αναγνώριση της κλίσης. Αν ληφθούν υπ' όψιν η παραγωγή και η σύνθεση, τουλάχιστον το παραγωγικότερο τμήμα αυτών των διαδικασιών, το σύστημα κερδίζει σε επάρκεια, επιστημονική πληρότητα και το λεξικό αποκτά μικρότερο όγκο, αφού δεν είναι απαραίτητο να καταχωρηθούν όλα τα παράγωγα και τα σύνθετα μιας γλώσσας.

3. Επίλογος

Η φυσική γλώσσα είναι ένα σύστημα τόσο δύσκολο και πολύπλοκο που η εφαρμογή μιας ελλιπούς μεθόδου ή μιας τυχαίας μεθόδου στην υπολογιστική επεξεργασία της χωρίς θεωρία υποστήριξης δεν παρέχει ικανοποιητικά αποτελέσματα.

11. Το φωνήεν /o/ που εμφανίζεται ανάμεσα στα συστατικά ενός συνθέτου δεν αποτελεί μέρος του πρώτους θέματος. Η ανάλυσή του όμως δεν ενδιαφέρει την παρούσα μελέτη γι' αυτό και το καταχωρώ ως συστατικό του πρώτου θέματος των συνθέτων.

12. Χρησιμοποιώ ως παράδειγμα το όνομα 'Γκένσερ' γιατί η λέξη 'γκενσερισμός' ήταν μεταξύ των λέξεων της εφημερίδας «Το Βήμα» που δεν αναγνωρίσθηκαν από τον επεξεργαστή.

ΕΠΙΛΕΓΜΕΝΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

Allen, J. (1987): *Natural Language Understanding.* The Benjamin / Cummings.

Bakamidis, S. and G. Carayannis (1987): "«Phonemia». A Phoneme Transcription System for Speech Synthesis in Modern Greek». *Speech Communication*, 6:159-169.

Chomsky, N. and M. Halle (1968): *The Sound Pattern of English.* New York: Harper and Row.

Chomsky, N. (1992): «A Minimalist Program for Linguistic Theory», Ms. MIT.

Δραγγιώτης, Α. (1992): «Αποτελέσματα μετρήσεων από το σύστημα γραμματικής αναγνώρισης του Πανεπιστημίου Αθηνών». Τμήμα Πληροφορικής, Πανεπιστήμιο Αθηνών.

Galiotou, E. and A. Ralli (1991): «Affixation in Modern Greek: A Computational Treatment». EURISCON.

Θεοφανοπούλου-Κοντού, Δ. (1990): *Μετασχηματιστική Σύνταξη.* Αθήνα.

Κοντός, Π. (1990): *Φωνολογική Ανάλυση του Αιτωλικού Ιδιώματος. Συμβολή στη Ν.Ε. Διαλεκτολογία.* Διδακτορική Διατριβή. Τομέας Γλωσσολογίας. Πανεπιστήμιο Αθηνών.

Kotsanis, Y., and Y. Maistros (1985). «Lexi Fanis»: A Lexical Analyser of Modern Greek. ACL Proceedings.

Kyriakopoulou, T. (1990): *Les Dictionnaires Electroniques - La Flexion Verbale en Grec Moderne.* Thèse de doctorat. Paris 8.

Loumos, V., Touratzidis, L., G. Carayannis (1990): Structural and Statistical Analysis for an Improved Morphological Processor». Τομέας Πληροφορικής, Εθνικό Μετσόβειο Πολυτεχνείο.

Μαλικούτη-Drachman, Α. & G. Drachman (1988): «Ο Τονισμός στα Ελληνικά». *Μελέτες για την Ελληνική Γλώσσα.* Θεσσαλονίκη: Κυριακίδης.

Nespor, M. & I. Vogel (1986): *Prosodic Phonology.* Dordrecht: Foris.

Ralli, A. and E. Galiotou (1987): «A Morphological Processor for Modern Greek». ACL Proceedings.

Ralli, A. (1989): «Compounds in a Transfer-based Machine Translation System». *Glossologia*, 8:117-130.

Ralli, A. (1993): «Compounds in Modern Greek». *Rivista di Linguistica. Special Issue on Compounds.* Pisa.

Sproat, R. (1992): *Morphology and Computation.* MIT Press.

Touratzidis, L. and A. Ralli (1992): «A Computational Treatment of Stress in Greek Inflected Forms». *Language and Speech*, 35:435-453.

Τριαντοπούλου, Θ., Τσαλίδης, Χ., Χριστοδουλάκης, Δ. (1991): «InterLEX, Μία context-free προσέγγιση για τη μορφολογική περιγραφή της Νέας Ελληνικής». *Μελέτες για την Ελληνική Γλώσσα.* Θεσσαλονίκη: Κυριακίδης.

Φιλιππάκη-Warburton, E. (1992): *Εισαγωγή στη Θεωρητική Γλωσσολογία*. Αθήνα: Νεφέλη.

SUMMARY

Angela Ralli, *Theoretical linguistics and Natural language processing*

This paper deals with the issue of how studies in theoretical linguistics can contribute to the research and development in computational linguistics. It is argued that a computational natural language system, such as a parser or a generator, must incorporate a solid analysis of theoretical linguistics. A range of applications for programs with knowledge of phonology and morphology are discussed, which are not generally taken into consideration by computer scientists in Greece. It is shown that a number of basic technics which have been proposed for phonological and morphological processing are inadequate from both the descriptive and computational point of view.