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 διαμόρφωσή 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 αναφερθούμε στις αρχές της θεωρητικής γλωσσολογίας που χρησιμοποιούνται από την υπολογιστική γλωσσολογία και θα προσπαθήσουμε να δείξουμε 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 κατανοηθούν, να αναλυθούν 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 επιβεβαιώσης 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 ορίζει τον μηχανισμό της γλώσσας γενικά, ανεξάρτητα από συγκεκριμένες γλώσσες, καθώς και να εντοπίζει τις παραμέτρους που διέπουν τις διαφορές των γλώσσων.

Οι δύο επιστημονικοί τομείς συγκλίνουν στο αντικείμενο μελέτης: την

* Μια 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 εξυπνών υπολογιστικών συστημάτων ούτως ώστε, αναλύοντας όσο το δυνατόν περισσότερες προτάσεις και κείμενα, να καταστεί δυνατή η κατανόηση της διαδικασίας της παραγωγής της γνώσης. Τη ΘΕ 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 ΘΕ προσπαθεί να δημιουργήσει μια θεωρία που να αποδεικνύει 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 το πιο οικονομικό τρόπο τις γραμματικές φράσεις και να απορρίπτουν τις μη-γραμματικές.

Οι δύο τομείς παρουσιάζουν αρκετές ομοιότητες στα εργαλεία της έρευνας και στη χρησιμοποιημένη μεθοδολογία. Απο την παράδειγμα, η χρήση μιας γραμματικής και ενός μοντέλου τυποποίησης (π.χ. γραμματικές ανα-ξάρτησες από το περιβάλλον, γραμματικές εξαρτήσεις από το περιβάλλον κ.λπ.) της ΘΕ 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 ε-πεκταομότητα 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 αντιμετώπιζοντας τη μια 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 σημαιολογικό. Για κάθε επίπεδο υπάρχουν οι βασικές μονάδες ανάλυσης, οι κανόνες και οι αρχές, χωρίς να αποκλείεται χρησιμοποίηση των ιδίων 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) ΣΥΝΤΑΞΗ

ΛΕΞΙΚΟ
MORFΟΛΟΓΙΑ

ΦΩΝΟΛΟΓΙΑ

ΣΗΜΑΣΙΟΛΟΓΙΑ

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 στη φωνολογία και στη μορφολογία, υπάρχουν τα πρώτα και αντιπροσωπευτικότερα δείγματα υπολογιστικής επεξεργασίας με αντικείμενο μελέτης την Ελληνική γλώσσα. Αναφέρομενη λοιπόν σε κάποιες από αυτές τις δημοσιεύσεις, θα προσπαθήσουμε να δείξουμε η σημαντική βοήθεια που μπορεί να προσφέρει η ΘΕ 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 χαρακτηρών της γραφής και των φωνημάτων της φωνολογίας της Νέας Ελληνικής. Σύμφωνα με τους συγγραφείς, οι νοημοσύνες μπορούν να χρησιμοποιηθούν για τη μετατροπή ενός κειμένου από τη γραπτή μορφή στην προφορική του εκφράση. Οι ιδιότητες υποστηρίζουν επίσης ότι το περιεχόμενο του άρθρου εκφράζει την απόψη του μηχανικού και αμφισβητώντας έμμεσα τη συμβολή της γλωσσολογικής ανάλυσης, παραδείχθηκε ότι για τη διατύπωση των κανόνων δεν στηρίχτηκαν 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 υφαννακοποίησης των υπερψηφικών συμφωνών. Περιγράφεται με τα ακόλουθα τέσσερα γενικά σχέματα 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 αρθρό, τα παραπάνω σχήματα εμφανίζονται ως νόμοι εξαρτημένοι από το περιβάλλον (context-dependent rules), υπεύθυνοι για την ουρανοκοποιημένη προφορά των υπερωικών χαρακτήρων 'γ', 'χ', 'γκ', 'κ' σε κατάλληλο γλωσσικό περιβάλλον. Οι πρώτες παρατηρήσεις για το είδος αυτών των νόμων, θα μπορούσα να αναφέρω τη μη-γλωσσολογική τους

2. Όπως γίνεται αντιληπτό, τα δίμηφα φωνήντα 'ι' και 'ε' δεν αποτελούν μέρος της δομικής περιγραφής των κανόνων. Η εξηγητική πρέπει να αναφέρθει στο ότι τα συγκεκριμένα φωνήντα έχουν ως πρώτο χαρακτήρα τα 'ι' και 'ε' αντιστοίχως, τα οποία ηδή 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

\[
\begin{align*}
\{0\} &= /o/ \\
\{e\} &= /e/ \\
\{ai\} &= /i/ \\
\pi &= /p/ \\
\psi &= /ps/ \\
\text{k.λπ.}
\end{align*}
\]

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 υπερανικοποιημένων υπερωικών ήχων και όχι να περιγράψουν γλωσσολογικά το φαινόμενο της ουρανικοποίησης. Ο ισχυρισμός όμως αυτός καταρρίπτεται από το ίδιο το αντίκειμενο 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 μέλετε 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 καταγράφουν οι αντιστοιχίες γραφημάτων και φωνημάτων μέσα στα πλαίσια 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 'ε' οι χαρακτήρες αντιστοιχούν σε εντελώς διαφορετικά φωνήματα (/e/ και // αντιστοιχίας).

3. Η αντιστοιχία 'ατ' = /e/ αναφέρεται από την παρουσία του τόνου και των διαλυτικών. Στα διήμητρα φωνητά, τόσο ο τόνος όσο και τα διαλυτικά αποτελούν στοιχεία που υπολογιστικά λαμβάνονται υπ' όμοι ανεξάρτητα από τους χαρακτήρες τους οποίους συνοδεύουν.
νταί κυρίως με βάση τις ιδιότητες της άρθρωσης των φωνημάτων μέσα στη στοματική κοιλότητα, τα φωνήματα περιγράφονται ως σύνολα χαρακτηριστικών (βλ. Chomsky & Halle, 1968). Τα κοινά χαρακτηριστικά αναπαριστάνουν τις ομοιότητες μεταξύ των φωνημάτων ενώ τα διαφορετικά τις διαφορές τους. Για παράδειγμα, τα υπερωκικά σύμφωνα μοιάζουν ως προς τα χαρακτηριστικά [+ σύμφωνο, - πρόσθιο, - ουρανικό] ενώ μεταξύ τους διαφέρουν ως προς το χαρακτηριστικό [+ συνεχές] για τα /x/, /y/ και [- συνεχές] για τα /k/ και /g/). Τα φωνήματα /e/ και /i/ συγκλίνουν ως προς τα χαρακτηριστικά [+ φωνήμα, + πρόσθιο, - στρογγυλό, - χαμηλό] και διαφέρουν μεταξύ τους ως προς το χαρακτηριστικό [+ υψηλό] για το /i/ και [- υψηλό] για το /e/). Σημειώθηκε 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 περιγραφής επάρκειας αλλά και της γενικευμένης χώρις οι εκφράσεις να χάνουν σε απλότητα. Χλωσολογικά, το φαινόμενο της ουρανικοποίησης των υπερωκικών συμφώνων δίνεται από ένα μόνο νόμο εξαρτημένο από το περιβάλλον (context-dependent) στην αριστερή πλευρά του οποίου περιγράφονται τα υπερωκικά σύμφωνα, ενώ στο περιβάλλον περιγράφονται τα φωνήματα /e/ και /i/:

(4)

\[
\begin{array}{c}
[+ \text{σύμφωνο}] \rightarrow [+ \text{ουρανικό}] / \quad \quad [+ \text{φωνήμα}, \quad - \text{πρόσθιο}, \quad - \text{στρογγυλό}] \\
- \text{πρόσθιο} \quad \quad - \text{ουρανικό} \quad \quad - \text{χαμηλό}
\end{array}
\]

Πώς θα ήταν, λοιπόν, και τι θα προσέφερε μία υπολογιστική επεξεργασία της πραγμάτωσης των ουρανικοποιημένων υπερωκικών, στο γλωσσικό περιβάλλον που περιγράφεται από τους Μπακαμίθ, 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 πληρες φωνολογικές αναπαραστάσεις για τα /e/ και /i/ είναι οι ακόλουθες:


5. Μια ανάλυση των φωνημάτων της Νέας Ελληνικής συγκεκριμένα του Αιτωλικού ιδίωματος, με βάση τα διακριτικά χαρακτηριστικά, προσφέρεται στη διατριβή του Π. Κοντού, 1990 (βλ. Βιβλιογραφία).
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. Επιπλέον, παρατηρούμε ότι ένας και μόνον νόμος καλύπτει ολές τις ουρανικοποιήσεις των υπερωϊκών συμφώνων, ενώ στην ανάλυση των Μπακάμιδη, 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 επεξεργαστική 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:

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 συ- 
οποικατά καθορίζεται από βασικές αρχές της μορφολογίας. Για παράδειγ-
μα, σε ένα μονολεκτικό σχηματισμό, η κλίση ακολουθεί συνήθως την πα-
ραγωγή. Επίσης, οι πληροφορίες των λημμάτων του γλωσσολογικού 
λεξικού (καταχωρημένες με τη μορφή μορφημάτων) και οι νόμοι σχηματι-
σμού λέξεων καθορίζουν εάν ένας συνδυασμός μορφημάτων είναι επι-
τρεπτός στη γλώσσα, σε αντίθεση με κάποιον άλλο που καταλήγει σε μη-
γραμματική δομή. Επίσης, αυτό το επίθημα `ο-'(ο)´ προστίθεται πάντοτε σε 
ρηματική βάση για να δώσει ουσιαστικό, ενώ το επίθημα `−ων(ω)´ συνδυά-
ζεται με ουσιαστικό για να δώσει ρήμα.9

Σ' ο, τι αφορά στην υπολογιστική επεξεργασία της μορφολογίας της 
Ελληνικής, όλες οι μέχρι τώρα προσπάθειες περιορίζονταν στην κλίση, κυρ-
ρώς 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 σχέ-

ση με την παραγωγή και τη σύνθεση.

γ) Περιγράφεται διεξοδικά σε όλες τις παραδοσιακές σχολικές γραμμα-

τικές, σε αντίθεση με την παραγωγή και τη σύνθεση των οποίων η περι-

γραφή εμφανίζει σημαντικά κενά.

Μια 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 σήμα με την εφαρμογή λεξικού φυσικολογικού νόμου. Εναλλακτικά, έχει δυνατόν να 

υιστηθεί η αποφή των δύο αλλομόρφων `−ων-` και `−ω-`, κατά την οποία στη 

λέξη 'ανακλώση' εμφανίζεται η μορφή `−ω-´.

9. Εντός παρενθέσεως εμφανίζονται τα κλιτικά επίθημα.
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 αδιάφορη ενότητα.

Θεωρώ ότι μία μορφολογική επεξεργασία από υπολογιστή που περιορίζεται στην ανάλυση και σύνθεση του φαινομένου της κλίσης είναι ελλιπής. Κατά την άποψή μου, εαν για την ανάπτυξη των περισσότερων μορφολογικών επεξεργασιών 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 προσπάθεια μου να υποστηρίξω και πρακτικά την παραπάνω θέση, με βοηθήσει ο πληροφορικός Αντώνης Δραγιώτης, ο οποίος ανέπτυξε ενα εντυπωσιακής κλίσης που δεν υποτελεί σε υπολογιστική κάλυψη κλιτικών φαινομένων από τη θέση υπάρχοντα συστήματα της μορφολογίας τα οποία επίσης καλύπτουν μόνο την κλίση. Το υπολογιστικό σύστημα του Δραγιώτη κάνει κατ' αρχήν γραμματική αναγνώριση της κλίσης σε 700.000 λέξεις, δηλαδή, ανάλυες τις λέξεις σε θέμα και κατάληξη προχωρώντας, συγχρόνως, στον μορφοσυντακτικό χαρακτηρισμό τους (κατηγορία, πτώση, αριθμός, χρόνος, πρόσωπο κ.λπ., ανάλογα με την περιπτώση). Στη συνέχεια, επανασυνθέτει τις θέση αναγνωρισμένες λέξεις, ενώ είναι σε θέση να δώσει και τα κλιτικά τους παραδείγματα, αφού το σύστημα σύνθεσης του επεξεργαστή μπορεί και δημιουργεί όλες τις κλίτες 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 περείχαν ορθογραφικά οφάλματα, είτε γιατί τα λήμματα 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 ελάβανε 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 θέμα όπως, 'ψευδο-' ('ψευδοδημοκράτες'), 'χρονο-' ('χρονομήχανη', 'χρονοντούλαπο'), 'ραδιο-', ('ραδιοδέκτης'), 'πρωτο-' ('πρωτοπαρουσίαση'), 'φιλο-' ('φιλοσοβιετική'), 'κοινοτικο-' ('κοινοτικοποιηση'), 'κρατικο-' ('κρατικοδιαίτη'), '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 μας πρότασης. Για τους σκοπούς αυτούς συντακτικού επεξεργαστής, η μορφολογική ανάλυση της παραγωγής και της σύνθεσης δεν αποτελεί άμεση προτεραιότητα. Η επεξεργασία της κλίσης είναι αρκετή για να προορίσει τις μορφολογικές πληροφορίες που χρειάζεται η σύνταξη.
κομμουνιστών"), 'ξανα-' ('ξαναπλησιάσει'), 'ανωτάτο-' ('ανωτατοποίηση')
k.λπ., καταχωρήθηκαν ως άγνωστες λέξεις για το σύστημα.11 Επίσης, δεν
αναγνωρίσθηκαν παράγωγες λέξεις όπως 'εκβαρβαριστής', 'ευτροφι-
σμός', 'υποτίτληση', 'κρατικιστικό', 'παραγωντισμός', 'σοφόκλεια', 'λαϊκί-
ζει', κ.λπ. Αξίζει να σημειωθεί ότι οι παραπάνω λέξεις δεν περιέχονται συ-
νήθως στα κοινά λεξικά. Αποτελούν σχηματισμούς που παράγονται από
tον μηχανισμό της γλώσσας με βάση τους κανόνες της και τις αρχές της.
Σε ένα μορφολογικό επεξεργαστή που λαμβάνει υπ' ομί, μόνο την κλίση,
eίναι αδύνατον να ληφθούν υπ' ομί, όλοι οι πιθανοί σχηματισμοί και να κα-
tαχωρηθούν ως ανεξάρτητα λήμματα, ακόμα και αν το λεξικό του συστή-
ματος εμπλουτίζεται σε καθημερινή βάση. Ας υποθέσουμε ότι στον επε-
ξεργαστή καταχωρείται ως νέο λήμμα το θέμα 'λαϊκιστικό' ('λαϊκιστικός'
χωρίς την κατάληψη). Αν ο επεξεργαστής δεν είναι σε θέση να κάνει χρή-
ση των μηχανισμών της παραγωγής και της σύνθεσης, τότε, το κέρδος εί-
ναι μηδαμινό γιατί συγχρόνως σε άλλα κείμενα είναι δυνατόν να εμφανι-
σθούν οι λέξεις 'ψευδολαϊκιστικός', 'αντιλαϊκιστικός', 'υπερλαϊκιστικός'
κ.λπ., καθώς και τα 'λαϊκιζω', 'λαϊκιστής', 'αντιλαϊκιστής', 'φιλολαϊκιστής',
'λαϊκούρα', 'υπερλαϊκισμός', 'έκλαϊκισμός' κ.λπ. Δεν είναι δυνατόν να προ-
βλέψει κανείς τον σχηματισμό και τη χρήση όλων των πιθανών να παρα-
χθούν λέξεων. Καταχώρηση νέων λημμάτων πρέπει να γίνεται βέβαια
στην περίπτωση των δανείων (π.χ. η λέξη 'φιλμ') ή των ξένων κυρίων ονο-
mάτων όπως 'Γκένσερ'.12 Στη συνέχεια, ο μηχανισμός της παραγωγής
μπορεί να αναγνωρίσει και να αναλύει πιθανά παράγωγα αυτών των λέ-
xεων (π.χ. 'φιλμάκι', 'γκένσερισμός', 'γκένσεριστής' κ.λπ.) χωρίς να χρειά-
ζεται πρόσθετος εμπλουτισμός του λεξικού μ' αυτά. Βλέπουμε λοιπόν ότι
μορφολογική επεξεργασία μιας λέξης δεν σημαίνει μόνο αναγνώριση της
κλίσης. Αν ληφθούν υπ' ομί, η παραγωγή και η σύνθεση, τουλάχιστον το
παραγωγικότερο τμήμα αυτών των διαδικασιών, το σύστημα κερδίζει σε ε-
πάρκεια, επιστημονική πληρότητα και το λεξικό αποκτά μικρότερο όγκο,
αφού δεν είναι απαραίτητο να καταχωρηθούν όλα τα παράγωγα και τα σύν-
θετα 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 υπολο-
γιστική επεξεργασία της χωρίς θεωρία υποστήριξης δεν παρέχει ικανο-
ποιητικά αποτελέσματα.

11. Το φωνήν /o/ που εμφανίζεται ανάμεσα στα συστατικά ενός συνθέτου δεν αποτελεί
μέρος του πρώτου θέματος. Η ανάλυση του όμως δεν ενδιαφέρει την παρούσα μελέτη γι' αυτό και το καταχώρισμα ως συστατικό του πρώτου θέματος των συνθέτων.
12. Χρησιμοποιούμε ως παράδειγμα το όνομα 'Γκένσερ' γιατί τη λέξη 'γκένσερισμός' ήταν με-
tαξί των λέξεων της εφημερίδας «Το Βήμα» που δεν αναγνωρίζηκαν από τον επεξεργα-
στή.
ΕΠΙΛΕΓΜΕΝΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

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

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.