Άρση της αμφισημίας με τον Mnemosyne Tagger

Παναγιώτης Γάκης
Neurolingo Company
gakis@sch.gr

Abstract

This paper presents the Mnemosyne tagger, a tool for dealing with ambiguity in computing environments. When the POS of a word is ambiguous, the computer must consider all its possible syntactic roles and, if more than one rule is activated, to produce all the structures they dictate. The Mnemosyne tagger is aimed at removing the lexical ambiguity in Modern Greek. It is based not on statistics but on the corresponding linguistic environment of the words. It supports the complete removal of lexical ambiguity not only to remove ambiguity in the POS but also in the gender and in the case of the ambiguous word. The tagger characterizes words that do not exist in the computational lexicon.

Keywords: tagger, ambiguity, computational lexicon

1 Ασάφεια[1]

Η ασάφεια παρατηρείται σε κάθε φυσική γλώσσα και είναι ιδιαίτερα εμφανής σε μια γλώσσα όπως η Ελληνική με τις πολλές ιδιαιτερότητες και την πολυπλοκότητα στην κλιτική παραγωγή.

Γενικά αναγνωρίζονται πέντε επίπεδα ανάλυσης του γλωσσικού συστήματος (Lyons 1981, Grishman 1986, Allen 1987, Filippaki-Warburton 1992,): **1. Φωνητική-Φωνολογία, 2. Μορφολογία, 3. Σύνταξη, 4. Σημασιολογία** και **5. Πραγματολογία**.

Ο όρος ασάφεια μπορεί να λάβει πολλούς προσδιορισμούς. Η ασάφεια μπορεί να είναι λεξική, σημασιολογική, συντακτική, αναφορική κ.ά. (Burgess and Simpson 1988).

Σύμφωνα με ορισμένες απόψεις, η λεξική ασάφεια παρουσιάζεται όταν μία λέξη αντιστοιχεί σε περισσότερα λήμματα του λεξικού (lexical entries) ή όταν χρησιμοποιείται με διαφορετικό νόημα σε μεταφορικό λόγο. Ο Jan van Eijck (Eijck and Jaspars 1996) ορίζει τη λεξική ασάφεια ως έλλειψη πληροφορίας για τη σημασία των λέξεων.

Πέραν των ερμηνειών αυτών που αποδίδονται στον όρο λεξική ασάφεια, για τον υπολογιστή η λεξική ασάφεια έχει άμεση σχέση με τον τρόπο αναπαράστασης και οργάνωσης των λεξιλογικών δεδομένων (Boguraev and Pustejovsky 1990), από τον οποίο προκύπτει ότι λεξικά ασαφείς είναι δύο ή περισσότερες λέξεις με κοινό ορθογραφικό τύπο που ανήκουν σε διαφορετικά λήμματα ή/και διαφέρουν σε ένα ή περισσότερα μορφοσυντακτικά χαρακτηριστικά (κυρίως όσον αφορά το μέρος του λόγου). Σύμφωνα με τον παραπάνω ορισμό ο τύπος *απαντήσεις* έχει λεξική ασάφεια, γιατί μπορεί να είναι ρήμα (<[**απαντώ**]) ή ουσιαστικό (<[**απάντηση**]). Επίσης ο τύπος *ματιών* έχει λεξική ασάφεια γιατί προέρχεται από διαφορετικά λήμματα (<[**μάτι**], [**ματιά**]). Επιπλέον, ο τύπος *κόρη* παρουσιάζει λεξική ασάφεια, επειδή είναι

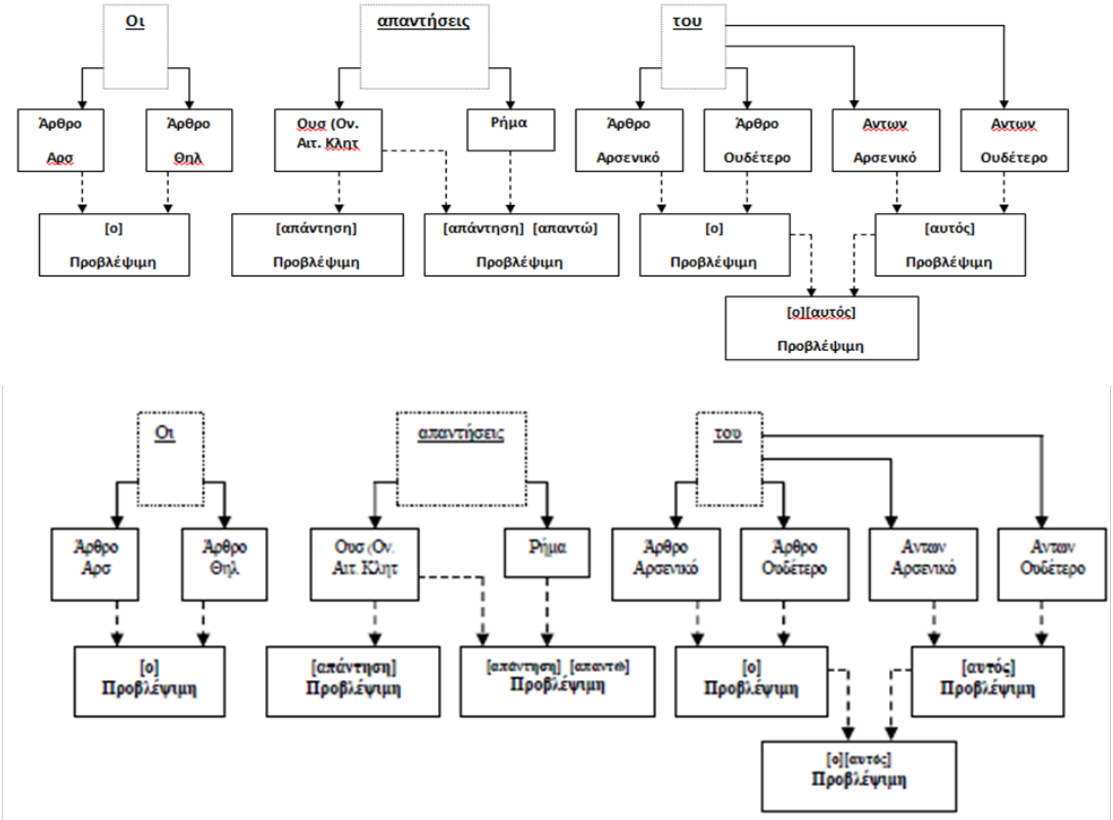
[1] Ο όρος *ασάφεια* ορίζεται και ως *αμφισημία*.

ασαφής ως προς την πτώση (ονομαστική ή αιτιατική ή κλητική)[2] μέσα στο ίδιο λήμμα [**κόρη**] και όχι επειδή έχει πολυσημία (*κορίτσι* || *κόρη ματιού*).

Ασάφεια είναι δυνατόν να προκύψει και για μια ολόκληρη πρόταση στο ευρύ πλαίσιο της επικοινωνίας (πραγματολογικό στάδιο). Και αυτό επειδή η πρόταση είναι ακόμα μια αφηρημένη μονάδα. Έτσι, μια πρόταση όπως: *Θα έρθω το βράδυ* μπορεί, ανάλογα με το γλωσσικό και εξωγλωσσικό πλαίσιο στο οποίο θα ενταχθεί και τη γλωσσική πράξη στην οποία εγγράφεται, να λειτουργήσει είτε ως απλή δήλωση είτε ως υπόσχεση είτε ως απειλή κτλ.

Η λεξική ασάφεια είναι προϊόν της πλούσιας μορφολογίας της ελληνικής γλώσσας. Είναι δυνατόν, για παράδειγμα, το λήμμα ενός ρήματος να περιλαμβάνει μέχρι και 250-300 τύπους, συμπεριλαμβανομένων των τύπων των δύο φωνών (ενεργητική, παθητική), καθώς και των προφορικών τύπων που αυτό έχει στην κλιτική του παραγωγή. Αντίστοιχα το λήμμα ενός επιθέτου μπορεί να περιλαμβάνει μέχρι και 100 τύπους, συμπεριλαμβανομένων και των τύπων που παράγονται στα παραθετικά του.

Στον παρακάτω σχήμα (Σχήμα 1) αποτυπώνονται οι μορφές λεξικής ασάφειας που παρατηρούνται στη φράση: *Οι απαντήσεις του έχουν καλά στοιχεία*.



Σχήμα 1 | Λεξική ασάφεια (συμπτωματική - προβλέψιμη)

Στα λήμματα του λεξικού δηλώνονται και κάποιοι λόγιοι τύποι της ελληνικής, οι οποίοι συναντώνται συχνά σε κείμενα του σύγχρονου γραπτού ελληνικού λόγου. Αυτό έχει ως συνέπεια να αυξάνονται οι ασαφείς τύποι και η υπολογιστική επεξεργασία του κειμένου, μέσω ενός τόσου ογκώδους λεξικού, να καθίσταται πιο πολύπλοκη. Και στις δύο περιπτώσεις (προβλέψιμη ή συμπτωματική) η άρση της

[2] *Συγκρητισμός* (συγχώνευση καταλήξεων διαφορετικών κλίσεων σ' έναν τύπο).

ασάφειας είναι εφικτή μόνο με την εξέταση του γλωσσικού περιβάλλοντος (context) εμφάνισης των ασαφών λέξεων.

1.1 Ασάφεια και υπολογιστικά συστήματα

Όταν η επεξεργασία της φυσικής γλώσσας γίνεται από υπολογιστικά συστήματα, υπάρχουν πολλές περιπτώσεις στις οποίες παρουσιάζεται ασάφεια τόσο ως προς το μέρος του λόγου -κάτι που είναι κύριο χαρακτηριστικό (major feature) (Pollard and Sag 1987)- όσο και ως προς τα επιμέρους μορφολογικά χαρακτηριστικά (attributes) που μπορεί να έχει ένας λεκτικός τύπος. Αυτό συμβαίνει, επειδή η επεξεργασία φυσικής γλώσσας στηρίζεται στη μορφολογική ανάλυση και, δεδομένης της λεξικής ασάφειας, δεν είναι αρκετή η απλή αναζήτηση των λέξεων σε μια βάση δεδομένων αλλά πρέπει να ληφθεί υπόψη και η πληροφορία των συμφραζομένων και/ή να γίνει μορφολογική ανάλυση.

Η ασάφεια (Gakis et al. 2012), ως εγγενές στοιχείο της φυσικής γλώσσας που απαντάται σε όλα τα επίπεδα μελέτης της φυσικής γλώσσας, αποτελεί το σημαντικότερο πρόβλημα μηχανικής επεξεργασίας φυσικής γλώσσας. Έρχεται σε αντίθεση με τον ντετερμινισμό των υπολογιστικών μηχανών και τις ντετερμινιστικές φιλοσοφικές δοξασίες.

Η άρση της ασάφειας αποτέλεσε και αποτελεί πρόκληση στην έρευνα της Γλωσσολογίας και της Υπολογιστικής Γλωσσολογίας.

Όταν το μέρος του λόγου μιας λέξης είναι ασαφές, ο υπολογιστής πρέπει να εξετάσει όλους τους πιθανούς συντακτικούς της ρόλους και, στην περίπτωση που ενεργοποιούνται περισσότεροι από έναν κανόνες, να παράγει όλες τις φραστικές δομές που αυτοί υπαγορεύουν, με την ελπίδα ότι μόνο μία ανάλυση τελικά θα επιτύχει. Αυτό όμως προϋποθέτει την επιτυχία στην αναγνώριση γειτονικών δομών, γεγονός που είναι αμφίβολο, αν θεωρήσουμε ότι και αυτές μπορεί να περιέχουν ασαφή συστατικά (Orphanos 2000).

Οι αμφίσημοι τύποι θα χαρακτηριστούν μορφοσυντακτικά από έναν σχολιαστή (tagger). Επιπλέον ο σχολιαστής (tagger) θα προσπαθήσει να «μαντέψει» τα μορφοσυντακτικά χαρακτηριστικά της λέξης (τουλάχιστον το μέρος του λόγου) ακόμη και για τους λεξικούς τύπους που δεν χαρακτηρίζονται από κανένα μορφοσυντακτικό χαρακτηρισμό (attribute). Αυτό θα γίνει αν εξετάσει κυρίως το γλωσσικό περιβάλλον της (τις λέξεις που προηγούνται ή/και έπονται). Κατ' αυτό τον τρόπο θα ολοκληρωθεί η λειτουργικότητα του λεξικού και η μετέπειτα ανάλυση και εξαγωγή της μορφοσυντακτικής πληροφορίας θα στηρίζεται σε αληθή δεδομένα.

Η ελληνική είναι γλώσσα με πολλές ιδιαιτερότητες, στοιχείο που καθιστά ακόμη πιο δύσκολη και πολύπλοκη την επεξεργασία της από τα συστήματα επεξεργασίας φυσικής γλώσσας. Για παράδειγμα η λέξη *απαντήσεις* <[**απάντηση**] μπορεί: α) να παίξει τον ρόλο της κεφαλής σε μια ονοματική φράση, β) να παίξει τον ρόλο κεφαλής σε μια ρηματική φράση: *απαντήσεις* <[**απαντώ**]. Επιπλέον ως ουσιαστικό έχει επιπλέον μορφολογική ασάφεια, καθώς μπορεί να είναι ονομαστική ή αιτιατική ή κλητική πληθυντικού (*συγκρητισμός*).

Μετά από στατιστική επεξεργασία των παραγόμενων κλιτικών τύπων του ηλεκτρονικού μορφολογικού λεξικού της Neurolingo, που αποτελεί και το ηλεκτρονικό λεξικό του γραμματικού διορθωτή (Neurolingo Lexicon), αναδείχτηκαν τα παρακάτω στατιστικά στοιχεία για τους κλιτικούς τύπους που έχουν μοναδικά μορφολογικά χαρακτηριστικά όπως και για τους κλιτικούς τύπους που είναι λεξικά ασαφείς (Πίνακας 1).

Οι ασαφείς τύποι προήλθαν μετά από ανάκληση όλων των λεξικών τύπων με κοινή ορθογραφική αναπαράσταση που επαναλαμβάνονται στο λεξικό (Πίνακας 1).

Μέρος του Λόγου	Αριθμός λέξεων
Αριθμός μοναδικών κλιτικών τύπων	**873,701**
Ασαφείς κλιτικοί τύποι (από διαφορετικά λήμματα)	**39,119**
Ασαφείς κλιτικοί τύποι (από το ίδιο λήμμα	
Σύνολο ασαφών τύπων	**4,758**
	917,578

Πίνακας 1 | Στατιστικά στοιχεία λεξικής ασάφειας

Η διερεύνηση των περιεχομένων του μορφολογικού ηλεκτρονικού λεξικού της Neurolingo (Πίνακας 2) έδειξε ότι η πλούσια μορφολογία της ελληνικής γλώσσας δημιουργεί ένα εκτεταμένο πεδίο λεξικής ασάφειας (lexical ambiguity), η οποία μπορεί να χαρακτηριστεί προβλέψιμη ή συμπτωματική.

Μέρος του Λόγου	Αριθμός κλιτικών τύπων
Ουσιαστικά	60,511
Επίθετα	22,844
Ρήματα	9,245
Μετοχές	865
Επιρρήματα	7,830
Άλλα μέρη του λόγου	420

Πίνακας 2 | Στατιστικά στοιχεία ηλεκτρονικού λεξικού

Οι ασαφείς παραγόμενοι τύποι του μορφολογικού ηλεκτρονικού λεξικού κατηγοριοποιήθηκαν σε ομάδες και στην ομάδα της προβλέψιμης ασάφειας (Πίνακας 3) περιγράφονται οι λέξεις που ταυτίζονται ορθογραφικά με άλλες κοινής ετυμολογικής προέλευσης. Η προβλέψιμη ασάφεια περιλαμβάνει δύο κατηγορίες:

α) η ασάφεια που παρατηρείται μέσα στο ίδιο το λήμμα. Ο τύπος *κρίνω*, για παράδειγμα, μπορεί να είναι οριστική ενεστώτα, υποτακτική ενεστώτα, οριστική εξακολουθητικού μέλλοντα, οριστική συνοπτικού μέλλοντα, καθώς και υποτακτική αορίστου. Σε αυτό το επίπεδο ασάφειας ανήκει και η ασάφεια που παρατηρείται μεταξύ της γενικής ενικού του αρσενικού και της γενικής ουδετέρου, καθώς και της γενικής πληθυντικού αρσενικού, θηλυκού και ουδετέρου των αντωνυμιών, επιθέτων και μετοχών. Η ύπαρξη κοινών τύπων στο ίδιο λήμμα (*συγκρητισμός*) θεωρείται εγγενές χαρακτηριστικό του κλιτικού μας συστήματος και δε σχολιάζεται στις επόμενες παραγράφους, αν και είναι επιβαρυντική για τον υπολογιστή.

β) η ασάφεια που παρατηρείται μεταξύ λεξικών τύπων διαφορετικών λημμάτων με το ίδιο ή διαφορετικό μέρους του λόγου. Υπάρχουν πολλές υποομάδες σε αυτή την κατηγορία (Gakis, Panagiotakopoulos, Sgarbas and Tsalidis 2013). Ενδεικτικά αναφέρεται η κατηγορία στα παροξύτονα ισοσύλλαβα αρσενικά σε *-ας* και στα παροξύτονα ισοσύλλαβα θηλυκά *-α*, που σχηματίζουν κοινούς όλους τους τύπους της κλιτικής τους παραγωγής με μόνη διαφορά τον μορφολογικό χαρακτηρισμό του γένους και της πτώσης. Το μόνο χαρακτηριστικό που παραμένει σταθερό είναι ο αριθμός. Έτσι, έχουμε τους τύπους: *κεφάλας*, *κεφάλα*, *κεφάλες*, *κεφαλών* που προέρχονται από τα ουσιαστικά [**κεφάλας**] και [**κεφάλα**]

και ανάλογα έχουν τον χαρακτηρισμό (γενική, ενικός, θηλυκό: *κεφάλας* <[**κεφάλα**]), άλλοτε τον χαρακτηρισμό (ονομαστική, ενικός, αρσενικό: *κεφάλας* <[**κεφάλας**]).

ΛΕΞΙΚΗ ΑΜΦΙΣΗΜΙΑ		
Ουσιαστικό – ρήμα	**Ρήμα – ρήμα**	**Ουσιαστικό – ουσιαστικό**
8.73% (8084 κλ. τύποι)	3.35% (3103 τύποι)	10.79% (9992 κλ. τύποι)
Επίθετο - επίρρημα	**Επίθετο – ρήμα**	**επίθετο – ουσιαστικό**
25.5% (23659 κλ. τύποι)	9.85% (9127 κλ. τύποι)	35.55% (32926 κλ. τύποι)
Επίθετο - επίθετο 1.7% (1576 κλ. τύποι)	**Αντωνυμία – άρθρο** 4.46%	

Πίνακας 3 | Στοιχεία προβλέψιμης λεξικής ασάφειας

Στη συμπτωματική λεξική ασάφεια, η οποία παρατηρείται στους ασαφείς τύπους του ηλεκτρονικού μορφολογικού λεξικού (Πίνακας 4), οι λεξικοί τύποι που την προκαλούν έχουν διαφορετικά θέματα και είναι διαφορετικής ετυμολογικής προέλευσης. Για παράδειγμα, ο τύπος *βάλτε* μπορεί να είναι κλητική ενικού του ουσιαστικού [**βάλτος**] ή προστακτική αορίστου του ρήματος [**βάζω**]. Επίσης υπάρχουν τύποι που είναι αδύνατον να αποδοθούν υπολογιστικά με τη σωστή συντακτική ιδιότητα λόγω απουσίας φωνητικής πληροφορίας, όπως ο τύπος *ήπια*, που μπορεί να είναι α΄ ενικό οριστικής αορίστου του ρήματος [**πίνω**] ή ονομαστική, αιτιατική και κλητική πληθυντικού ουδετέρου του επιθέτου [**ήπιος**] ή το επίρρημα [**ήπια**] (αν και η ασάφεια μεταξύ επιθέτου-επιρρήματος είναι προβλέψιμη).

ΣΥΜΠΤΩΜΑΤΙΚΗ ΑΜΦΙΣΗΜΙΑ							
ΜτΛ 1 **Επιφώνημα**	**ΜτΛ 2** **Μετοχή**	**Τύποι** **3**	**Παράδ.** άι	**ΜτΛ 1** **Αντων.**	**ΜτΛ 2** **Άρθρο**	**Τύποι** **50**	**Παράδ.** τα
	Επίθετο	**12**	καλέ		**Ουσιαστ.**	**57**	σου
	Επίρρημα	**3**	ίσα		**Επίθετο**	**39**	ίδια
	Ρήμα	**3**	ορίστε		**Ρήμα**	**8**	εμείς
	Ουσιαστικό	**104**	γιούχα		**Επίρρημα**	**18**	κάμποσο
Πρόθεση	**Αντωνυμία**	**16**	με	**Σύνδεσμος**	**Σύνδεσμος** **Ουσιαστ.**	**3** **11**	να μόλο
	Ουσιαστικό	**29**	συν		**Ρήμα**	**4**	όντας
	Επίρρημα	**12**	υπό		**Επίρρημα**	**33**	πριν
					Αντωνυμία	**6**	όσον
Ουσιαστικό	**Ουσιαστικό**	**5**	η				

Πίνακας 4 | Στοιχεία συμπτωματικής Ασάφειας

2 Αποσαφήνιση ασάφειας από τον σχολιαστή (tagger)

Η αποσαφήνιση των χαρακτηριστικών γίνεται από τον σχολιαστή (tagger) που αποδίδει τα ορθά μορφολογικά χαρακτηριστικά. Η αποσαφήνιση της λεξικολογικής ασάφειας είναι από τα σημαντικότερα ζητήματα στην επεξεργασία του κειμένου. Για παράδειγμα, το να αποφασίσουμε αν το *απαντήσεις* είναι ρήμα ή ουσιαστικό μπορεί

να επιλυθεί με επισημείωση των μερών του λόγου (**p**art-**o**f-**s**peech tagging). Το να αποφασίσουμε αν το *κόρη* σημαίνει *κοπέλα* ή *κόρη του ματιού* μπορεί να επιλυθεί με αποσαφήνιση των εννοιών των λέξεων (word sense disambiguation), το οποίο όμως δεν είναι αντικείμενο έρευνας της παρούσας διατριβής.

Μια μεγάλη ποικιλία από εργασίες ανήκουν στα προβλήματα λεκτικής αποσαφήνισης. Για παράδειγμα το γλωσσικό περιβάλλον (context) είναι αυτό που θα καθορίσει εάν ο τύπος *το* είναι άρθρο ή αντωνυμία, γνώση απολύτως αναγκαία σε μετέπειτα επίπεδο ανάλυσης των γραμματικών λαθών (π.χ. στους κανόνες του τελικού -*ν*).

3 Υπάρχοντες μηχανισμοί άρσης ασάφειας

Η υπολογιστική άρση της μορφοσυντακτικής ασάφειας είναι εφικτή μόνο με την εξέταση του γλωσσικού περιβάλλοντος (context) μιας ασαφούς λέξης. Οι υπολογιστικές μέθοδοι που έχουν αναπτυχθεί με στόχο τη μορφοσυντακτική αποσαφήνιση χωρίζονται γενικά σε δύο κατηγορίες:

α) Σύμφωνα με τη γλωσσολογική προσέγγιση, οι ειδικοί κωδικοποιούν χειρωνακτικά κανόνες βασισμένους σε γενικεύσεις παραδειγμάτων αποσαφήνισης, τα οποία συνήθως συλλέγονται από σώμα κειμένων μορφοσυντακτικά χαρακτηρισμένων (Ορφανός et al. 1999).
β) Σύμφωνα με την προσέγγιση της μηχανικής εκμάθησης, ένα στατιστικό μοντέλο για την επίλυση του γλωσσικού προβλήματος επάγεται αυτόματα από σώμα χαρακτηρισμένων κειμένων.

Τα πρότυπα της λεξιλογικής ασάφειας για τα ελληνικά δεν κωδικοποιούνται πουθενά εντελώς. Πολλές μελέτες σχετικά με τη λεξιλογική ασάφεια προσδιορίζουν μερικά πρότυπά της. Κατά συνέπεια ένας μεγάλος αριθμός διαφορετικών προσεγγίσεων για την άρση της ασάφειας στο ΜτΛ έχει επιχειρηθεί όπως οι προσπάθειες που έγιναν από την ομάδα Δερματά – Κοκκινάκη (Dermatas and Kokkinakis 1995) και οι οποίες στηρίζονται στη χρήση στοχαστικών συστημάτων και χρησιμοποιούν μοντέλα Hidden Markov (HMM). Η πειραματική διαδικασία ξεκινά με μέγεθος εκπαίδευσης τις 10Κ λέξεις και αυξανόταν κατά 10Κ κάθε φορά. Με τη χρήση του μικρού συνόλου ετικετών, το ποσοστό ακρίβειας φτάνει σε αρκετά υψηλά επίπεδα. Ξεκινώντας με τις 10.000 λέξεις, είναι ιδιαίτερα μικρό στην αρχή και αυξάνεται κατακόρυφα μέχρι το πρώτο 25% του συνόλου εκπαίδευσης.

Άλλος τρόπος άρσης της ασάφειας επιχειρήθηκε από συστήματα που βασίζονται στην εκμάθηση δένδρων απόφασης και υλοποιήθηκε από την ομάδα των: Ορφανού-Χριστοδουλάκη (Orphanos and Christodoulakis 1999). Επίσης, χρησιμοποιήθηκε και ο αλγόριθμος IGTREE του συστήματος TiMBL, με σκοπό τη σύγκριση της απόδοσής του σε σχέση με τις άλλες παραλλαγές αλγορίθμων εκμάθησης δένδρων απόφασης που χρησιμοποιήθηκαν. Ένα σύνολο κειμένων με μέγεθος 137.765 λέξεων χρησιμοποιήθηκε για την προσέγγιση αυτή. Το σώμα (corpus) κειμένων αυτό είναι ποικίλο και ετερόκλητο και αποτελείται από γραπτά φοιτητών, τμήματα λογοτεχνικών κειμένων, άρθρα από εφημερίδες, τεχνικά, οικονομικά και αθλητικά περιοδικά. Έγινε διαχωρισμός των λέξεων και αφέθηκε σε ένα λεξικό η αυτόματη, πρώτη απόδοση των ετικετών. Το μέγεθος των ετικετών δεν καθορίστηκε αυστηρά, και αφέθηκε στην αρμοδιότητα ενός προγράμματος που συνεργαζόταν με ένα λεξικό να πραγματοποιήσει την αρχική επισημείωση και να επιλέξει το μέγεθος της ετικέτας

που θα αποδοθεί στις λέξεις που ανήκουν σε κάθε μέρος του λόγου. Ακολούθησε χειρωνακτική διόρθωση των κειμένων. Το μοντέλο αυτό, ακολουθώντας τη γλωσσολογική προσέγγιση, επέλυσε την ασάφεια πτώσης, γένους, αριθμού κτλ. μέσα από ένα επίπεδο ρηχής συντακτικής ανάλυσης. Έχοντας επιλύσει την ασάφεια του ΜτΛ, χρησιμοποίησε ένα ελάχιστο σύνολο κανόνων φραστικής δομής που περιγράφουν τον σχηματισμό απλών ονοματικών φράσεων (ΟΦ), προθετικών φράσεων (ΠΦ) και ρηματικών συνόλων (ΡΣ) και έτσι, επιλύθηκε η ασάφεια των λοιπών μορφοσυντακτικών χαρακτηριστικών σε επίπεδο φράσης (Ορφανός et al. 1999).

Άλλες δύο προσπάθειες που έγιναν για την ελληνική γλώσσα βασίστηκαν στο ίδιο σύστημα, τον Brill tagger (Petasis, et al. 1999). Για την επίτευξη μεγαλύτερης συνέπειας και ακρίβειας επελέγη ένα αρκετά περιορισμένο σύνολο ετικετών, μόλις 58, αν αναλογιστούμε το πλήθος των μορφολογικών χαρακτηριστικών της νέας ελληνικής. Και στο σύστημα αυτό στο τμήμα της προεργασίας ακολουθείται χειρωνακτική επισημείωση για δύο σώματα κειμένων αποτελούμενα από 65Κ περίπου λέξεις. Στη φάση των πειραμάτων χρησιμοποιήθηκε δεκαπλή επικυρωμένη επικύρωση προκειμένου να επιτευχθεί η εξαγωγή πιο αμερόληπτων συμπερασμάτων. Με το ίδιο σύστημα, τον Brill tagger, ασχολήθηκε ερευνητική ομάδα του ΙΕΛ η οποία εκτός από την αρχική μορφή του Brill tagger χρησιμοποιεί και άλλα ενισχυτικά προγράμματα, για να βελτιώσουν την αρχική διαδικασία.

Άλλες προσεγγίσεις αποσαφήνισης της ασάφειας βασίζονται σε συστήματα που αναπτύχθηκαν για να υποστηρίξουν τη μορφολογία φτωχότερων κλιτικά γλωσσών. Η πλειονότητα αυτών των εργαλείων -μορφολογικοί επεξεργαστές, ετικετοποιητές, ανιχνευτές θεμάτων (θέμα κατάληξη)- χρησιμοποιούν πρακτικές διεθνώς καθιερωμένες όπως: το μορφολογικό μοντέλο των δύο επιπέδων του Koskenniemi (Sgarbas et al. 1995), το άμεσο ακυκλικό γράφημα λέξης (Sgarbas et al. 2000), οι τεχνικές μηχανικής εκμάθησης (Papageorgiou et al. 2000), οι στατιστικές μέθοδοι (Tambouratzis and Carayiannis 2001) κ.ά.

3.1 Σχολιαστής (tagger) Mnemosyne

Ο σχολιαστής (tagger) που υλοποιήθηκε είναι προσανατολισμένος στην άρση της λεξικής ασάφειας στα νέα ελληνικά. Είναι βασισμένος όχι σε στατιστικά στοιχεία αλλά στο ανάλογο γλωσσικό περιβάλλον των λέξεων. Η υλοποίηση του σχολιαστή (tagger) έχει προσανατολιστεί στις ανάγκες των κανόνων των μετέπειτα επιπέδων. Αυτό σημαίνει ότι το σύστημα αντιμετωπίζει τον σχολιαστή (tagger) ως υποστηρικτικό αλλά απολύτως απαραίτητο υλικό και δεν αναλύει όλες τις μορφές ασάφειας. Το Mnemosyne όμως υποστηρίζει την πλήρη άρση της λεξικής ασάφειας μόνο με γλωσσολογική πληροφορία και στοχεύει στην ανάδειξή του ως το μοναδικό εργαλείο που διαχειρίζεται τις ασαφείς λέξεις μόνο βάσει του περιβάλλοντός τους.

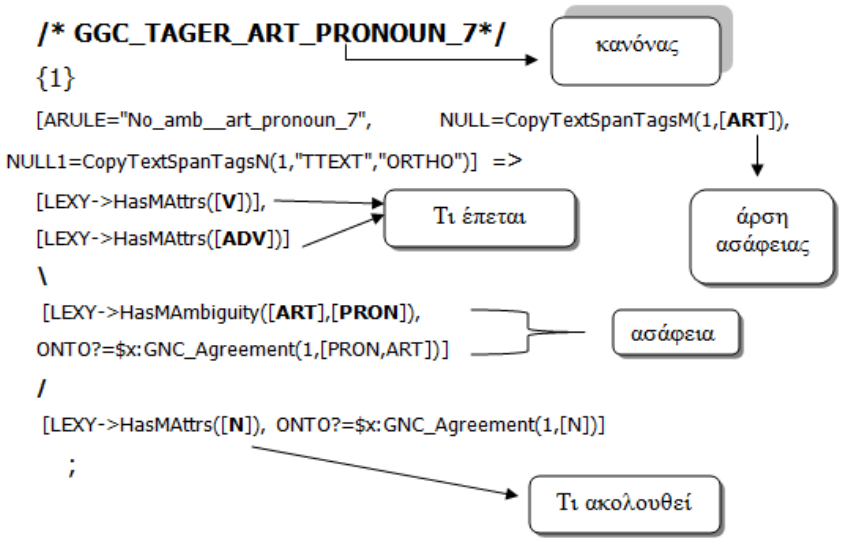
Αποτελείται από 70 κανόνες (rules) και η πλειονότητα των κανόνων αφορά στη άρση της λεξικής ασάφειας μεταξύ άρθρου και αντωνυμίας (64 rules). Από αυτούς τους κανόνες το γλωσσικό περιβάλλον αποδίδει τον μορφολογικό χαρακτηρισμό: άρθρο ως ΜτΛ σε 20 περιπτώσεις (20 rules) και τον μορφολογικό χαρακτηρισμό: αντωνυμία ως ΜτΛ σε 14 rules. Ο σχολιαστής (tagger) επεκτείνεται και χαρακτηρίζει λέξεις που δεν υπάρχουν στο μορφολογικό λεξικό (άτονες λέξεις, λέξεις με ανορθογραφία) η γνώση των μορφολογικών των οποίων είναι απαραίτητη στο επίπεδο της συντακτικής ανάλυσης. Οι επιπλέον κατηγορίες άρσης λεξικής ασάφειας είναι οι ακόλουθες:

- Ουσιαστικό και επίρρημα
- Ουσιαστικό και πρόθεση
- Ουσιαστικό και αντωνυμία
- Ουσιαστικό και άρθρο
- Ουσιαστικό και σύνδεσμος
- Ουσιαστικό και ρήμα
- Ουσιαστικό - επιφώνημα
- Ουσιαστικό και μετοχή
- Ρήμα και ουσιαστικό
- Πρόθεση και αντωνυμία
- Επίρρημα και σύνδεσμος
- Επίθετο και ουσιαστικό

Ο σχολιαστής (tagger) του Mnemosyne δε στοχεύει μόνο στην άρση της ασάφειας ως προς το ΜτΛ αλλά και ως προς το γένος και την πτώση της ασαφούς λέξης. Η απόδοση των σωστών μορφολογικών χαρακτηριστικών είναι απαραίτητη σε περιπτώσεις συμφωνίας γένους και αριθμού που εξετάζονται στο επίπεδο της συντακτικής ανάλυσης. Έτσι, αίρει την ασάφεια μεταξύ:

- Αρσενικού και ουδετέρου
- Αρσενικού και θηλυκού

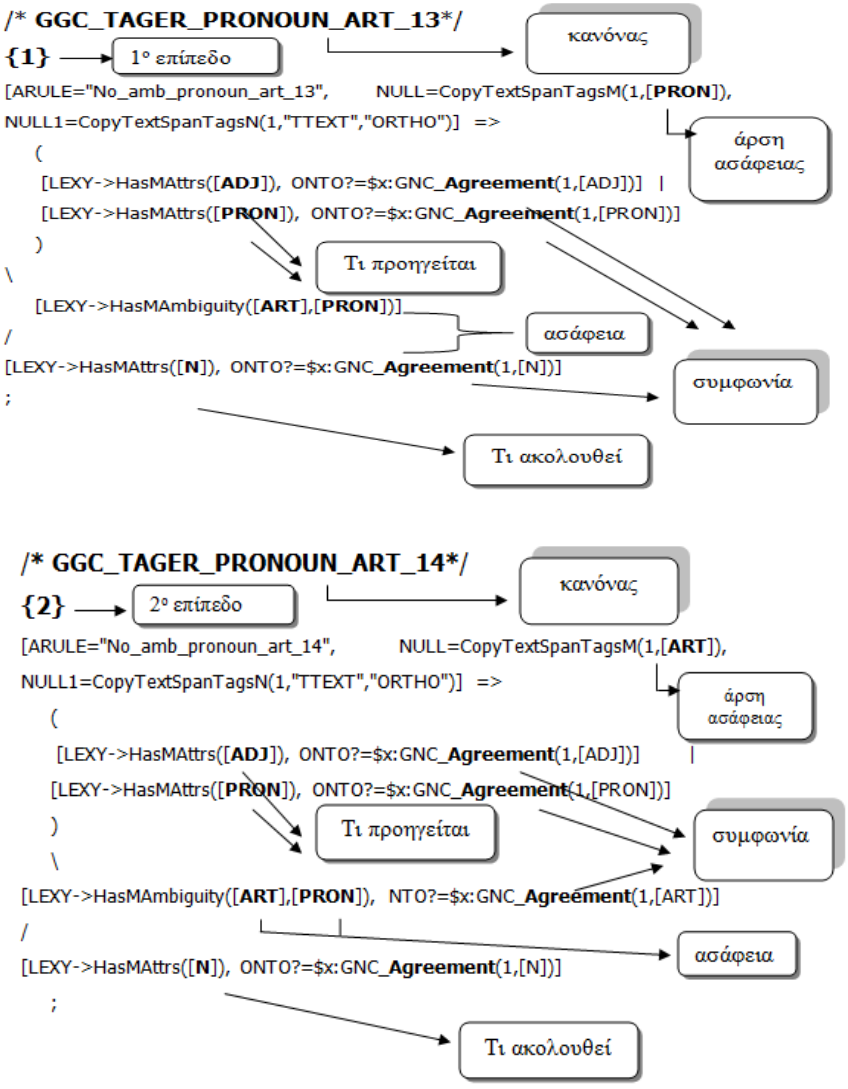
Για την άρση της λεξικής ασάφειας εξετάζονται τόσο οι προηγούμενες λέξεις -σε αριθμό μέχρι και 4 οντότητες (tokens)- όσο και/ή οι επόμενες -σε αριθμό μέχρι και 4 οντότητες (tokens). Η μορφή που μπορεί να έχει ένας κανόνας είναι η ακόλουθη (Σχήμα 1, 2):



Σχήμα 2 | Κανόνας Mnemosyne tagger

Το γεγονός ότι η γλωσσολογική πληροφορία είναι αυτή που καθορίζει τη λεξιλογική ασάφεια αποδεικνύεται και από το γεγονός ότι ο σχολιαστής (tagger) δεν αίρει την ασάφεια με μοντέλα γενικευμένης ισχύος (π.χ. όταν η ασαφής λέξη ακολουθεί άρθρο είναι ουσιαστικό) αλλά επιτρέπει μέσω των επιπέδων να αντιμετωπίσει και πιο ειδικές, σπάνιες περιπτώσεις που όμως επιτρέπουν την καλύτερη διαχείριση των συντακτικών κανόνων. Έτσι, στη φράση: *τις μυστικές τους δυνάμεις* η ακολουθία των

ΜτΛ έχει ως εξής: άρθρο, επίθετο, **ασαφής λέξη**, ουσιαστικό. Ομοίως και στη φράση: *την καλή την κοπέλα*, η ακολουθία των ΜτΛ έχει την ίδια μορφή: άρθρο, επίθετο, **ασαφής λέξη**, ουσιαστικό. Στην πρώτη περίπτωση όμως η ασαφής λέξη είναι αντωνυμία ενώ στη δεύτερη άρθρο. Μια γενική και καθολική ισχύς του κανόνα θα δημιουργούσε προβλήματα, ενώ και γλωσσολογικά είναι μη αποδεκτή μια άρση της ασάφειας με καθαρά στατιστικά κριτήρια. Έτσι, με δύο επίπεδα -στο πρώτο επίπεδο επισημαίνεται η συμφωνία σε γένος, αριθμό και πτώση των οντοτήτων (tokens) που προηγούνται και έπονται, ενώ σε δεύτερο η συμφωνία και της ασαφούς λέξης- έχουμε ορθή απόδοση της λεξιλογικής ασάφειας. Η μορφή που έχουν οι πιο πάνω περιπτώσεις είναι η ακόλουθη:



Σχήμα 3 | Κανόνες Mnemosyne tagger

4 Συμπεράσματα

Η αποσαφήνιση των χαρακτηριστικών γίνεται από τον σχολιαστή (tagger) που αποδίδει τα ορθά μορφολογικά χαρακτηριστικά. Η άρση της γλωσσικής ασάφειας αποτέλεσε ένα μεγάλο πρόβλημα στα Νέα Ελληνικά, καθώς στην ελληνική

παρατηρείται μεγάλου εύρους λεξική ασάφεια. Ο γραμματικός διορθωτής στηρίζεται στη μορφολογική ανάλυση και ως εκ τούτου η λεξική ασάφεια αποτελεί το μεγαλύτερο πρόβλημα στην επεξεργασία φυσικής γλώσσας. Για να αρθεί η ασάφεια, έγινε ανάκληση όλων των τύπων με κοινή ορθογραφία και ακολούθως κατηγοριοποιήθηκαν οι κατηγορίες της λεξικής ασάφειας (κυρίως της προβλέψιμης) και με καθαρά γλωσσικούς κανόνες – το ανάλογο γλωσσικό περιβάλλον των λέξεων. Η ανάλυση όλων των μορφών λεξικής ασάφειας που επηρέαζαν τους κανόνες (rules) του γραμματικού διορθωτή γίνεται από το ίδιο περιβάλλον του λογισμικού (Mnemosyne) και συγκεκριμένα μέσα από 68 κανόνες. Για την άρση της λεξικής ασάφειας εξετάζονται τόσο οι προηγούμενες λέξεις -σε αριθμό μέχρι και 4 οντότητες (tokens)- όσο και/ή οι επόμενες -σε αριθμό μέχρι και 4 οντότητες (tokens).

Η κωδικοποίηση των πληροφοριών της ελληνικής γλώσσας αποτελεί αδιάκοπη διαδικασία και απαιτεί διαρκή έλεγχο των δεδομένων της. Θεμέλιος λίθος της παρούσας έρευνας αποτελεί η περιγραφή του πρώτου σχολιαστή (tagger) για τα ελληνικά με αμιγώς γλωσσικά κριτήρια. Ο όγκος και η πληρότητα των δεδομένων του μορφολογικού ηλεκτρονικού λεξικού και ο γλωσσικός σχολιαστής (tagger) αποτελούν καινοτόμα εργαλεία για τα ελληνικά στον τομέα της υπολογιστικής γλωσσολογίας. Οι δύο αυτές εφαρμογές μπορούν να αποτελέσουν τη βάση για την υλοποίηση και άλλων υπολογιστικών εργαλείων (αυτόματη σύνοψη, μετάφραση κ.ά.).

Βιβλιογραφία

Allen, James. 1987. *Natural Language Understanding*. San Francisco: The Benjamin/Cummings.

Boguraev, Branlair, and James Pustejovsky. 1990. "The role of knowledge representation in Lexicon Design." Στο *Proceedings of the 13th conference on computation linguistics*, edited by Hans Karlgren, 36-41.USA.

Burgess, Curt, and Greg Simpson B. 1988. "Cerebral hemispheric mechanisms in the retrieval of ambiguous word meanings". *Brain and Language* 33:86-103.

Dermatas, Evangelos, and George Kokkinakis. 1995. "Automatic stochastic tagging of natural language texts". *Computational Linguistics* 21(2): 137-163.

Philippaki-Warburton, Eirini. 1992. *Εισαγωγή στη θεωρητική γλωσσολογία*. Αθήνα: Νεφέλη.

Gakis, Panagiotis, Christos Panagiotakopoulos, Kyriakos Sgarbas, and Christos Tsalidis. 2012. "Design and implementation of an electronic lexicon for Modern Greek". *Literary and Linguistic Computing* 27: 1-15.

Gakis, Panagiotis, Christos Panagiotakopoulos, Kyriakos Sgarbas, and Christos Tsalidis. 2013. "Analysis of lexical ambiguity in Modern Greek using a computational lexicon". *Literary and Linguistic Computing* 27(2):20-38.

Grishman, Ralph. 1986. *Computational Linguistics. An introduction*. Cambridge: University Press.

Lyons, James. 1981. *Language and Linguistics. An introduction*, Cambridge: University Press.

Orphanos, Georgios, and Dimitris Christodoulakis. 1999. "Part-of-speech Disambiguation and Unknown Word Guessing with Decision Trees." Στο *Proceedings of EACL*, edited by Thompson S. Henry and Lascarides, Alex. 134-141. Association for Computational Linguistics: USA.

Orphanos, Georgios. 2000. “Computational morphosyntactic analysis of Greek.” PhD diss., University of Patras.

Ορφανός, Γεώργιος, Παναγιώτης Γάκης, και Άννα Ιορδανίδου. 1999. “Μορφοσυντακτική ασάφεια στη νέα ελληνική: Η περίπτωση επιθέτου - ουσιαστικού – επιρρήματος.” Στο *Πρακτικά της 20ής Συνάντησης του Τομέα Γλωσσολογίας του Τμήματος Φιλολογίας*. Αριστοτέλειο Πανεπιστήμιο, 373-383. Θεσσαλονίκη.

Papageorgiou, Harris, Prokopis Prokopidis, Voula Giouli, and Stelios Piperidis. 2000. “A Unified POS Tagging Architecture and its Application to Greek”. Στο *Proceedings of the 2nd Language Resources and Evaluation Conference*, edited by Gavrilidou, M. and Carayannis, G. and Markantonatou, S. and Piperidis, S. and Stainhauer, G, European. 1455-1462. European Language Resources Association (ELRA): Athens.

Petasis, Georgios, Georgios Paliouras, Vangelis Karkaletsis, Constantine Spyropoulos, and Ion Androutsopoulos. 1999. “Resolving Part-Of-Speech Ambiguity in the Greek language Using Learning Techniques.” Στο *Proceedings of the ECCAI Advanced Course on Artificial Intelligence (ACAI '99)* (July 5 - 16 1999, In Fakotakis, N. et al. (Eds.), Machine Learning in Human Language Technology (Proceedings of the ACAI Workshop). 29-34. Chania. Greece.

Pollard, Carl, and Ivan Sag A. 1987. “Information-Based Syntax and Semantics.” *CSLI Publications*, Vol. 1, Fundamentals (Center for the Study of Language and Information Publication Lecture Notes, No. 13).

Sgarbas, Kyriakos, Nikolaos Fakotakis, and George Kokkinakis. 1995. “A PC-KIMMO-Based Morphological Description of Modern Greek.” *Literary and Linguistic Computing* 10(3):189-201.

Sgarbas, Kyriakos, Nikolaos Fakotakis, and George Kokkinakis. 2000. “Two Algorithms for Incremental Construction of Directed Acyclic Word Graphs.” *International Journal on Artificial Intelligence Tools* 4(3):369-381

Tambouratzis, George, and George Carayannis. 2001. Automatic Corpora – based Stemming in Greek. *Literary and Linguistic Computing* 16(4):445-466.

Van J. Eijck, and Jan Jaspars. 1996. Ambiguity and reasoning. *Technical Report CS-R9616*. Dutch national research institute for mathematics and computer science.