ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ &
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ Ε.Μ.Π.

ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΤΗΣ ΕΛΛΗΝΙΚΗΣ ΝΟΗΜΑΤΙΚΗΣ ΓΛΩΣΣΑΣ ΓΙΑ ΣΥΣΤΗΜΑΤΑ ΣΤΑΤΙΣΤΙΚΗΣ ΜΗΧΑΝΙΚΗΣ ΜΕΤΑΦΡΑΣΗΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΔΗΜΗΤΡΙΟΥ ΚΟΥΡΕΜΕΝΟΥ

Διπλωματούχου Ηλεκτρολόγου Μηχανικού & Μηχανικού Υπολογιστών Ε.Μ.Π.

ΕΠΙΒΛΕΠΩΝ:

Στεφ. Κόλλιας

Καθηγητής Ε.Μ.Π.

ΑΘΗΝΑ, Ιούλιος 2020

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ Ε.Μ.Π.

ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΤΗΣ ΕΛΛΗΝΙΚΗΣ ΝΟΗΜΑΤΙΚΗΣ ΓΛΩΣΣΑΣ ΓΙΑ ΣΥΣΤΗΜΑΤΑ ΣΤΑΤΙΣΤΙΚΗΣ ΜΗΧΑΝΙΚΗΣ ΜΕΤΑΦΡΑΣΗΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΔΗΜΗΤΡΙΟΥ ΚΟΥΡΕΜΕΝΟΥ

Διπλωματούχου Ηλεκτρολόγου Μηχανικού & Μηχανικού Υπολογιστών Ε.Μ.Π.

Συμβουλευτική Επιτροπή : Στεφ. Κόλλιας, Καθ. Ε.Μ.Π. (Επιβλέπων)
Π. Μαραγκός, Καθ. Ε.Μ.Π.
Α.-Γ. Σταφυλοπάτης, Καθ. Ε.Μ.Π.

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 8η Ιουλίου 2020

.....	.....	.....
Στέφανος Κόλλιας, Καθηγητής Ε.Μ.Π. (Επιβλέπων)	Π. Μαραγκός, Καθ. Ε.Μ.Π.	Α.-Γ. Σταφυλοπάτης, Καθ. Ε.Μ.Π.
.....	.....	.....
Γ. Στάμου, Αναπλ. Καθ. Ε.Μ.Π.	Γ. Καρυδάκης, Επικ. Καθ. Παν. Αιγαίου	Κ. Καρπούζης, Ερευνητής Α' Ε.Π.Ι.Σ.Ε.Υ Ε.Μ.Π.
	.....	
	Κ. Νταλιάνης, Καθ. Παν. Δυτ. Αττικής	

ΑΘΗΝΑ, Ιούλιος 2020

Στους γονείς μου και στη μνήμη του Καθηγητή ΕΜΠ Γιώργου Καραγιάννη

I'm committed to sign in everything I communicate, but I also speak.
I still I believe that reach more people when I do that.
I bridge two different cultures and two different worlds,
and I think that bridge still needs work.
I. King Jordan

.....
Δημήτριος Κουρεμένος
Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Πρόλογος

Η παρούσα διατριβή διεξήχθη στη σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου της Αθήνας, στον Τομέα Τεχνολογίας Πληροφορικής και Υπολογιστών, στο εργαστήριο Ευφυών Συστημάτων (ISLAB), υπό την επίβλεψη του Καθηγητή ΕΜΠ κ. Στέφανου Κόλλια.

Για την πραγματοποίηση της παρούσας διατριβής οφείλω να εκφράσω τις ειλικρινείς ευχαριστίες μου στον αποθανόντα πρώην επιβλέποντα καθηγητή ΕΜΠ κ. Γ. Καραγιάννη. Σε όλη την πολυετή επαφή μας υπήρξε ένας από τους κυριότερους γνώμονες και οδηγούς μου τόσο σε επιστημονικά θέματα όσο και σε θέματα αρχής και άποψης.

Θα ήθελα να ευχαριστήσω ειλικρινά για την συμβολή τους τα μέλη της τριμελούς Συμβουλευτικής Επιτροπής της παρούσας διατριβής, που προσέφεραν πρόθυμα τη βοήθειά τους όπου την χρειάστηκα. Ειδικότερα θα ήθελα να ευχαριστήσω θερμά τον Καθηγητή κ. Κλήμη Νταλιάνη για την αμέριστη υποστήριξή του και ειδικά τον κ. Δρ. Γιώργο Σιόλα για την πολύτιμη βοήθειά του στην ολοκλήρωση της διατριβής μου.

Ειδική μνεία οφείλω, στις ερευνήτριες κα. Ευθυμίου Ελένη και κα. Εβίτα Φωτινέα με τις οποίες συνεργαστήκαμε σε αρκετά έργα του Ινστιτούτου Επεξεργασίας Λόγου και συνεισέφεραν στην ερευνητική μου δραστηριότητα.

Καθ' όλη τη διάρκεια της ερευνητικής προσπάθειας για την παρούσα διατριβή, θα ήθελα να ευχαριστήσω τον αγαπημένο μου αδελφό Δρ. Στέλιο Κουρεμένο για την αμέριστη υποστήριξή του, καθώς και τους συναδέλφους μου και ερευνητές στο Ινστιτούτο Πληροφορικής και Τηλεπικοινωνιών του ΕΚΕΦΕ Δημόκριτου για τη συνεχή υποστήριξή τους.

Σε όλη τη διάρκεια της ερευνητικής μου προσπάθειας, ο ρόλος της οικογένειάς μου ήταν καθοριστικός. Οφείλω σε καθέναν ξεχωριστά την ειλικρινή μου ευγνωμοσύνη για την καθολική υποστήριξή τους σε κάθε στάδιο της εκπόνησης της παρούσας διατριβής. Η συνεχής στήριξη και προτροπή τους υπήρξαν καθοριστικοί παράγοντες για την ολοκλήρωσή της.

ΔΗΜΗΤΡΙΟΥ ΚΟΥΡΕΜΕΝΟΥ

Ιούλιος 2020

Περίληψη

Η παρούσα διατριβή τοποθετείται στο πεδίο της αυτόματης Μηχανικής Μετάφρασης και της διαπροσωπείας ανθρώπου-μηχανής στην περίπτωση των ατόμων με προβλήματα ακοής, κάνοντας χρήση της γλώσσας των κωφών και της Ελληνικής Νοηματικής Γλώσσας. Αξιοποιεί, μελετά, δοκιμάζει και προτείνει νέες μεθοδολογίες Μηχανικής Μετάφρασης μεταξύ της Ελληνικής Γλώσσας και της Ελληνικής Νοηματικής Γλώσσας.

Η Μηχανική Μετάφραση αποτελεί μέρος του ευρύτερου πεδίου της Γλωσσικής Τεχνολογίας που μελετά την επεξεργασία της ανθρώπινης γλώσσας από υπολογιστή, με σκοπό την επικοινωνία μεταξύ ανθρώπου και μηχανής, αλλά και ως βοηθητικό εργαλείο για την επικοινωνία μεταξύ ανθρώπων που μιλούν διαφορετικές γλώσσες.

Με τη βοήθεια της τεχνολογίας και της επιστήμης της πληροφορικής ανοίγονται νέοι ορίζοντες που θα αλλάξουν τον τρόπο με τον οποίο λειτουργούμε καθημερινά. Όταν θα μπορεί μια μηχανή, ένας υπολογιστής, να αναγνωρίζει τις εκφωνούμενες γλώσσες (προφορικό και γραπτό λόγο) και τις Νοηματικές γλώσσες (φυσικές γλώσσες των κοινοτήτων των κωφών), τότε θα έχουμε πιο εύκολη πρόσβαση στην πληροφορία και κατ' επέκταση στα οφέλη που θα προκύψουν από ένα ευρύτερο πλαίσιο νέων υπηρεσιών, αλλά και τη δυνατότητα να εκτελούμε επαγγελματικές συναλλαγές από απόσταση, με μεγάλη ευκολία και ταχύτητα.

Όταν θα μπορεί μια μηχανή να κατανοεί τις φυσικές (ανθρώπινες) γλώσσες (εκφωνούμενες γλώσσες και νοηματικές γλώσσες), να τις μεταφράζει σε άλλες γλώσσες και να τις αναπαράγει, τότε θα υπάρχει διαθέσιμο ένα ισχυρό μέσο επικοινωνίας ανθρώπου-μηχανής, χρήσιμο όχι μόνο για τους ακούοντες, αλλά και για τα άτομα με προβλήματα ακοής, σε πολλούς τομείς της ανθρώπινης δραστηριότητας.

Η παρούσα διατριβή προτείνει ένα πρωτότυπο σύστημα βασισμένο σε κανόνες μηχανικής μετάφρασης που έχει ως στόχο τη δημιουργία μεγάλων παράλληλων εύρωστων γραπτών σωμάτων κειμένων της ελληνικής και της Ελληνικής Νοηματικής Γλώσσας, με χρήση της Σύντομης Μεταγραφής της Ελληνικής Νοηματικής Γλώσσας (ΣΜΕΝΓ) (text glosses) που αναπτύχθηκε για τις ανάγκες της διατριβής.

Τα σώματα κειμένων χρησιμοποιούνται ως δεδομένα κατάρτισης για τη δημιουργία γλωσσικών μοντέλων ν-γραμμάτων (n-gram Language Model). Επιπλέον, χρησιμοποιού-

νται και ως δεδομένα εκπαίδευσης για το σύστημα MOSES Στατιστικής Μηχανικής Μετάφρασης. Πρέπει να σημειωθεί ότι όλη η διαδικασία είναι ισχυρή και ευέλικτη, καθώς δεν απαιτεί βαθιά γνώση γραμματικής της ΕΝΓ.

Τέλος, πρέπει να τονιστεί η έλλειψη γλωσσικών πόρων στην ΕΝΓ. Για τον λόγο αυτό, η παρούσα διατριβή προτείνει μια καινοτόμο μεθοδολογία για τη δημιουργία γλωσσικών πόρων που εκλείπουν από την επιστημονική βιβλιογραφία, με σκοπό τη μοντελοποίηση της ΕΝΓ και την εφαρμογή της σε συστήματα στατιστικής μηχανικής μετάφρασης. Η αξιολόγηση του προτεινόμενου συστήματος μετάφρασης πραγματοποιείται στο πεδίο των καιρικών προγνώσεων, από όπου έχουν παραχθεί 20.284 λέξεις (tokens) και 1.000 προτάσεις. Παρουσιάζονται μετρήσεις και χρονικές εκτιμήσεις για τη δημιουργία γλωσσικών πόρων και αξιολογούνται τα γλωσσικά μοντέλα της ΕΝΓ μέσω της περιπλοκής. Τέλος, το πρωτότυπο σύστημα ΜΜ που παρουσιάζεται επιτυγχάνει ελπιδοφόρες επιδόσεις, χρησιμοποιώντας τη μετρική βαθμολογία BiLingual Understudy Assessment (BLEU) για την αξιολόγηση της μετάφρασης.

Λέξεις-κλειδιά- machine translation, Greek, Greek Sign Language, GSL, Deaf people communication, SMT, Moses, Phrase model, BLUE, glosses, language models.

Summary

One of the objectives of Assistive Technologies is to help people with disabilities communicate with others and provide means of access to information. One of the aims of Assistive Technologies is to help people with disabilities to communicate with others and to provide means of access to information. As an aid to Deaf people, in this work we present a novel prototype Rule Based Machine Translation (RBMT) system for the creation of large quality written Greek Sign Language (GSL) glossed corpora. In particular, the proposed RBMT system supports the professional translator of GSL to produce different kinds of GSL glossed corpus.

Then the glossed corpus is used as training data for the production/creation of Language Model (LM) n-gram. With the GSL glossed corpus and for any domain, we can build, test and evaluate different kinds of Language Models for different kinds of glossed GSL corpus, even if there is no real written GSL large corpus. Here it should be noted that the whole process is robust and flexible since it does not demand deep grammar knowledge of GSL.

These GSL parallel corpus and languages models also will be used as training data by the Statistical Machine Translation (SMT) MOSES application system. It should be noted that the whole process is robust and flexible, since it does not demand deep grammar knowledge of GSL. Furthermore, it should be stressed that Language Models for written GSL Gloss are missing from the scientific literature, thus this work is pioneer in this field. Evaluation of the proposed scheme is carried out in the weather reports domain, where 20,284 tokens and 1,000 sentences have been produced. By using the BiLingual Evaluation Understudy (BLEU) metric score, our prototyped MT system achieves a very promising performance.

Keywords— machine translation, Greek, Greek Sign Language, GSL, Deaf people communication, SMT, Moses, Phrase model.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Με τον όρο Μηχανική Μετάφραση (ΜΜ) αναφερόμαστε σε μια αυτοματοποιημένη διαδικασία κατά την οποία μεταφέρεται ο γραπτός λόγος από μια γλώσσα-πηγή (ΓΠ) σε μια γλώσσα-στόχο (ΓΣ). Εδώ και δεκαετίες η ΜΜ αποτελεί ένα από τα δυσκολότερα επιστημονικά ζητήματα της εφαρμοσμένης γλωσσολογίας, συγκεντρώνοντας το ενδιαφέρον επιστημόνων από διάφορους χώρους, όπως της αυτόν της τεχνητής νοημοσύνης, της θεωρητικής και υπολογιστικής γλωσσολογίας, της λογικής και της ψυχογλωσσολογίας.

Σύμφωνα με τους Porta et al. (2014) σχετικά με τα θεμελιώδη προβλήματα των Νοηματικών Γλωσσών (ΝΓ), τα περισσότερα νέα έργα της ΝΓ υιοθετούν τις γλωσσικές θεωρίες που αναπτύχθηκαν για τις προφορικές γλώσσες, αντί να δοκιμάσουν νέες θεωρίες. Από την άποψη της υπολογιστικής επεξεργασίας των γλωσσών, οι ΝΓ εξακολουθούν να μην διαθέτουν επαρκείς γλωσσικούς πόρους, ενώ επίσης μεγάλο πρόβλημα αποτελεί η έλλειψη επίσημης γραμματικής της Ελληνικής Νοηματικής Γλώσσας (ΕΝΓ), καθώς και επίσημου συστήματος γραπτής μορφής της. Η στατιστική μηχανική μετάφραση είναι η κυρίαρχη τάση στον τομέα της ΜΜ αυτή τη στιγμή και χρησιμοποιείται από τα ηλεκτρονικά συστήματα μετάφρασης που αναπτύσσονται από εταιρείες όπως η Google και η Microsoft. Στη στατιστική μηχανική μετάφραση (ΣΜΜ), τα μεταφραστικά συστήματα εκπαιδεύονται από μεγάλες ποσότητες παράλληλων δεδομένων (σώματα/corpora).

1.1 Σκοπός της διατριβής

Στην παρούσα διατριβή, λαμβάνοντας υπόψη τα σημαντικότερα εμπόδια της ΝΓ, εξετάζουμε μεθόδους και εργαλεία για την παραγωγή μεγάλων, εύρωστων και ποιοτικών σωμάτων κειμένων σε γραπτή μεταγραφή της ΕΝΓ, με σκοπό τη μοντελοποίηση της ΕΝΓ. Στη συνέχεια τα σώματα της ΕΝΓ χρησιμοποιούνται για την εκπαίδευση στατιστικών συστημάτων μηχανικής μετάφρασης της ΕΝΓ.

Τα ελληνικά σώματα κειμένων συλλέχθηκαν από το ευρύτερο διαδίκτυο και κυρίως από ιστοσελίδες μετεωρολογικών δελτίων, με στόχο τόσο τη βελτίωση της προσαρμοστικότητας των συστημάτων ΜΜ στην παραπάνω θεματική ενότητα αλλά και σε άλλες θεματικές περιοχές, όσο και τη βελτίωση της απόδοσή τους, χρησιμοποιώντας όσο το δυνατό λιγότερους πόρους (γλωσσικούς και ανθρώπινους).

Στόχος της διατριβής ήταν η εύρεση ενός αποδοτικού τρόπου ανάλυσης της γλώσσας και η δημιουργία ενός δίγλωσσου σώματος κειμένων ελληνικής γλώσσας-ΕΝΓ, ώστε να χρησιμοποιηθεί στο σύστημα στατιστικής μηχανικής μετάφρασης MOSES (Koehn et al., 2007).

Η στατιστική μηχανική μετάφραση (ΣΜΜ) είναι η κυρίαρχη τάση στον τομέα της ΜΜ και χρησιμοποιείται από τα ηλεκτρονικά συστήματα μετάφρασης που αναπτύσσονται από εταιρείες-κολοσσούς, όπως η Google και η Microsoft. Στη στατιστική μηχανική μετάφραση, μεγάλες ποσότητες παράλληλων δεδομένων (σώματα/corpora) εκπαιδεύουν τα μεταφραστικά συστήματα έτσι ώστε να μεταφράζουν μικρά τμήματα κειμένου/προτάσεις, παρακάμπτοντας την ανάγκη για γνώση της γραμματικής των γλωσσών του συστήματος.

Οι βασικές επιδιώξεις της παρούσας διατριβής συνοψίζονται στα παρακάτω σημεία:

- Πρόταση και θεμελίωση ενός συστήματος γραπτής μεταγραφής γλώσσας της Ελληνικής Νοηματικής Γλώσσας (ΣΜΕΝΓ), το οποίο αναπτύχθηκε για τις ανάγκες της παρούσας διατριβής.

- Δημιουργία ενός νέου πρότυπου συστήματος μηχανικής μετάφρασης βασισμένου σε κανόνες, με σκοπό την υποβοήθηση της δημιουργίας μεγάλων, εύρωστων και ποιοτικών σωμάτων κειμένου στην ΕΝΓ.

- Εκτενής χρήση ανοιχτού λογισμικού και δεδομένων σε όλα τα στάδια της παρούσας διατριβής.

- Δημιουργία εύρωστων κειμένων της ΕΝΓ διαφόρων τύπων, με σκοπό τη σύσταση στατιστικών γλωσσικών μοντέλων. Τα γλωσσικά μοντέλα μπορούν να χρησιμοποιηθούν σε αρκετές εφαρμογές, όπως σε συστήματα στατιστικής μηχανικής μετάφρασης. Για τη δημιουργία στατιστικών γλωσσικών μοντέλων και τη μέτρηση απόδοσης χρησιμοποιούμε τη σουίτα εφαρμογών ανοιχτής χρήσης SRILM.

- Ανάπτυξη στατιστικής μηχανικής μετάφρασης της πλατφόρμας MOSES, με χρήση παράλληλων σωμάτων κειμένων και γλωσσικών μοντέλων.

- Εφαρμογή μετρικών συστημάτων αυτόματης αξιολόγησης, στο πλαίσιο της βελτιστοποίησης, ώστε μέσα από τον συνδυασμό τους να αξιολογούνται τα αποτελέσματα της μεταφραστικής διαδικασίας πιο αντικειμενικά.

1.1.1 Δομή της διατριβής

Στο πρώτο κεφάλαιο παρουσιάζεται μια γενική περιγραφή της δουλειάς που έχει γίνει στην παρούσα διατριβή και γίνεται μια εισαγωγή στην ιστορία της μηχανικής μετάφρασης.

Στο δεύτερο κεφάλαιο παρουσιάζονται τα προβλήματα και τα χαρακτηριστικά της ΝΓ και γίνεται μια επισκόπηση της χρήσης των γλωσσικών τεχνολογιών που έχουν αναπτυχθεί για τη ΝΓ.

Στο τρίτο κεφάλαιο εξετάζονται τα βασικά θέματα της ΜΜ και των ΝΓ, καταγράφονται οι θεμελιώδεις προσεγγίσεις που αναφέρονται στη βιβλιογραφία και εξετάζονται ειδικότερα τα θέματα εφαρμογής τους στη ΝΓ.

Στο τέταρτο κεφάλαιο παρουσιάζεται το σύστημα μηχανικής μετάφρασης της ελληνικής γλώσσας προς την ΕΝΓ, καθώς και η σύντομη μεταγραφή της Ελληνικής Νοηματικής Γλώσσας (ΣΜΕΝΓ) που χρησιμοποιείται στην παρούσα διατριβή και η αρχιτεκτονική του συστήματος.

Στο πέμπτο κεφάλαιο παρουσιάζονται τα στάδια ανάλυσης, μεταφοράς και εξαγωγής των σωμάτων κειμένου. Διερευνώνται οι βασικές αρχές που διέπουν τη διαδικασία της ΜΜ με κανόνες και στη συνέχεια αναλύονται οι διάφορες μέθοδοι και τα προβλήματα της μηχανικής μετάφρασης του ελληνικού γραπτού λόγου στη φυσική γλώσσα της κοινότητας των κωφών, τη ΝΓ.

Στο έκτο κεφάλαιο γίνεται εφαρμογή της τεχνικής του τεμαχισμού μέσω της χρήσης κανονικών εκφράσεων (Chunking with Regular Expressions), η οποία αποτελεί μια σημαντική διεργασία της επεξεργασίας για τη δημιουργία των εύρωστων σωμάτων κειμένων για τις ανάγκες της παρούσας διατριβής.

Στο έβδομο κεφάλαιο γίνεται εφαρμογή της τεχνικής της μεταφοράς τεμαχισμένων προτάσεων (transforming chunks) και συντακτικών δέντρων. Η μεταφορά τεμαχισμένων προτάσεων αποτελεί και αυτή σημαντική διεργασία στην επεξεργασία της δημιουργίας των εύρωστων σωμάτων κειμένων για τις ανάγκες της παρούσας διατριβής.

Στο όγδοο κεφάλαιο διερευνώνται οι βασικές αρχές που διέπουν τα στατιστικά γλωσσικά μοντέλα με διάφορες μεθόδους αξιολόγησης. Αναλύονται τα προβλήματα μοντελοποίησης της ΝΓ και πραγματοποιούνται δοκιμές διαφόρων παραλλαγών σωμάτων κειμένων της ΕΝΓ. Τέλος, παρουσιάζονται τα πειραματικά αποτελέσματα και οι αξιολογήσεις των γλωσσικών μοντέλων των γλωσσικών σωμάτων κειμένων που παράγονται από το υπο-

βοηθούμενο RBMT σύστημά μας. Επιπλέον, παρουσιάζονται αποτελέσματα γλωσσικών μοντέλων από άλλες ΝΓ, καθώς και αναλυτικά αποτελέσματα των αξιολογήσεων που εκτελέστηκαν.

Στο ένατο κεφάλαιο γίνεται μελέτη της εφαρμογής σωμάτων κειμένων στη στατιστική μηχανική μετάφραση της πλατφόρμας MOSES και των παραμετροποιήσεων που χρησιμοποιήθηκαν για το στήσιμο της πλατφόρμας για τις ανάγκες της παρούσας διατριβής.

Τέλος, στο δέκατο κεφάλαιο παρουσιάζονται τα συμπεράσματα της μελέτης, καθώς και μια σειρά από προτάσεις για τη συνέχιση της έρευνας και τη βελτίωση των αποτελεσμάτων της.

Κεφάλαιο 2

Ελληνική Νοηματική Γλώσσα και Γλωσσικές Τεχνολογίες

Στο Κεφάλαιο 2 γίνεται αναφορά στα βασικά χαρακτηριστικά των Νοηματικών Γλωσσών (ΝΓ) και μια πρώτη γνωριμία με την Ελληνική Νοηματική Γλώσσα (ΕΝΓ). Παρουσιάζονται συνοπτικά τα βασικά θέματα των γλωσσικών τεχνολογιών Νοηματικής Γλώσσας και εξετάζονται ειδικότερα θέματα εφαρμογής των γλωσσικών τεχνολογιών ως προς την Ελληνική Νοηματική Γλώσσα, που είναι η φυσική γλώσσα της κοινότητας των κωφών στην Ελλάδα.

2.1 Γλωσσικές Τεχνολογίες και Νοηματική Γλώσσα

2.1.1 Ενεργές περιοχές έρευνας στις Νοηματικές Γλώσσες

Κατά την έρευνά μας σχετικά με τις γλωσσικές τεχνολογίες ΝΓ, εντοπίσαμε αρκετό υλικό σε πολλά έργα και έρευνες που έχουν γίνει σε ακαδημαϊκά και ερευνητικά ιδρύματα ανά την υφήλιο. Η ΝΓ ως βασική γλώσσα επικοινωνίας των κωφών περιέχει αρκετά κοινά χαρακτηριστικά με τις ομιλούμενες γλώσσες, ως προς την επιστημονική αντιμετώπιση και μεθοδολογία μελέτης. Εντοπίσαμε τις παρακάτω βασικές περιοχές έρευνας σχετικά με τη ΝΓ:

- Γραμματική δομή της Νοηματικής Γλώσσας
- Μεθοδολογίες και Τεχνολογίες Λεξικογράφησης της Νοηματικής Γλώσσας
- Συστήματα μεταγραφής της Νοηματικής Γλώσσας
- Συστήματα Μηχανικής Μετάφρασης της Νοηματικής Γλώσσας

- Ψηφιακή Αναπαράσταση και Σύνθεση της Νοηματικής Γλώσσας
- Τεχνολογίες Αναγνώρισης της Νοηματικής Γλώσσας

2.1.2 Νοηματικές Γλώσσες – Γενικά Χαρακτηριστικά

Οι ΝΓ είναι ανθρώπινες οπτικές-κινησιακές γλώσσες, οι οποίες χρησιμοποιούν τον τρισδιάστατο χώρο για να εκφράσουν το γλωσσικό μήνυμα. Αντί του ήχου, χρησιμοποιούν συνδυασμούς χειρομορφών της παλάμης, κατεύθυνση και κίνηση παλάμης, θέση χεριών, ώμων, θέση άνω κορμού και εκφράσεις προσώπου για την έκφραση γλωσσικών μηνυμάτων. Οι ΝΓ είναι συστήματα φυσικής γλώσσας τα οποία ικανοποιούν όλα τα κριτήρια της ανθρώπινης γλώσσας (Hickok et al., 1998).

Κάθε ΝΓ είναι ένα οπτικό-κινησιακό σύστημα επικοινωνίας που έχει αναπτυχθεί και χρησιμοποιείται από μια συγκεκριμένη κοινότητα κωφών και βαρήκοων ατόμων. Πολύ περισσότερο από μια απλή συλλογή τυχαίων χειρονομιών, οι ΝΓ ταξινομούνται ως ξεχωριστές γλώσσες μειονοτήτων και είναι τόσο σύνθετες και αναγκαίες όσο και εκείνες που χρησιμοποιούνται από την πλειοψηφία των ακουόντων ανθρώπων. Δυστυχώς, η ενσωμάτωση των κωφών στη σημερινή κοινωνία έχει περιοριστεί καθώς ελάχιστοι ακούοντες ξέρουν πώς να νοηματίζουν. Το ίδιο φαίνεται να ισχύει και στο διαδίκτυο και τον παγκόσμιο ιστό, όπου η πληροφορία είναι διαμορφωμένη με τρόπο δύσχρηστο για τα άτομα που έχουν ως μητρική γλώσσα τη νοηματική.

2.1.3 Η Ελληνική Νοηματική Γλώσσα (ΕΝΓ) και ο Ψηφιακός Αποκλεισμός

Η ΕΝΓ είναι η φυσική γλώσσα της κοινότητας των κωφών στην Ελλάδα. Όπως συμβαίνει και με τις υπόλοιπες νοηματικές, η ιδιαιτερότητά της σε σχέση με αυτό που ο περισσότερος κόσμος έχει συνηθίσει να ονομάζει "γλώσσα" είναι ότι η γραμματική της, δηλαδή το σύστημα των γραμματικών κανόνων βάσει των οποίων διαρθρώνεται ο λόγος και επιτυγχάνεται η επικοινωνία, δεν είναι προφορικό αλλά οπτικο-κινησιακό. Η ΕΝΓ λέγεται "ελληνική" γιατί χρησιμοποιείται στην Ελλάδα από Έλληνες νοηματιστές. Αυτό όμως δεν σημαίνει σε καμία περίπτωση ότι απεικονίζει την ελληνική γλώσσα ή ότι προέρχεται από αυτήν. Αντίθετα, πρόκειται για ένα αυτόνομο γλωσσικό σύστημα που μπορεί να μελετηθεί και να αναλυθεί όπως και κάθε άλλη φυσική γλώσσα.

Η ΕΝΓ έχει αναπτυχθεί σε ένα κοινωνικό και γλωσσολογικό πλαίσιο παρόμοιο με εκείνο των περισσοτέρων νοηματικών γλωσσών. Χρησιμοποιείται ευρέως στην ελληνική κοινότητα κωφών και οι εκ γενετής χρήστες της υπολογίζονται σε 40.600. Υπάρχει επίσης

ένας μεγάλος αριθμός ακουόντων νοηματιστών της ΕΝΓ, κυρίως σπουδαστές της ΕΝΓ και συγγενείς κωφών. Αν και ο ακριβής αριθμός ακουόντων σπουδαστών της ΕΝΓ στην Ελλάδα είναι άγνωστος, η πιο πρόσφατη απογραφή της Ομοσπονδίας Κωφών Ελλάδας[1] δείχνει ότι το 2002 περίπου 300 άνθρωποι εγγράφηκαν σε μαθήματα ΕΝΓ ως δεύτερης γλώσσας, ενώ η τάση στα χρόνια που ακολούθησαν ήταν αυξητική. Η πρόσφατη αύξηση των κωφών σπουδαστών στη βασική εκπαίδευση, καθώς επίσης και ο αριθμός των κωφών σπουδαστών που φοιτούν σε άλλα ιδρύματα, εκπαιδευτικές μονάδες απομακρυσμένων πόλεων και στην ιδιωτική εκπαίδευση, ενδεχομένως να οδηγήσουν σε διπλασιασμό του συνολικού αριθμού χρηστών της ΕΝΓ στην Ελλάδα. Επίσημα ιδρύματα όπου χρησιμοποιείται η ΕΝΓ περιλαμβάνουν 11 συλλόγους κωφών στα ελληνικά αστικά κέντρα και 14 εκπαιδευτικά ιδρύματα κωφών.

Το 2000 η ΕΝΓ αναγνωρίστηκε από το Ελληνικό Κοινοβούλιο ως η επίσημη γλώσσα των κωφών που ζουν στην Ελλάδα (Νόμος 2817/2000). Από το 2001 το λεξικό ΝΟΗΜΑ, ένα ηλεκτρονικό λεξικό της ΕΝΓ με 3.000 βιντεο-λήμματα συνδεδεμένα με τις αποδόσεις τους στα ελληνικά και με μια ποικιλία τρόπων αναζήτησης των λημμάτων, είναι το μόνο επιστημονικά έγκυρο λεξικογραφικό βοήθημα στη δίγλωσση εκπαίδευση των κωφών (Ευθυμίου et al., 2000).

Η έλλειψη ενός ευρέως αποδεκτού συστήματος γραφής των νοηματικών γλωσσών αποτελεί επίσης εμπόδιο για την εκπαίδευση των ομιλητών της ΕΝΓ, εδώ και δεκαετίες. Παραδοσιακά, η γνώση της ΕΝΓ μεταφέρεται από γενιά σε γενιά ζωντανά, μέσα από τους κοινωνικούς ιστούς της κοινότητας των κωφών (οικογένειες-σωματεία). Η αδυναμία μεταγραφής της ΕΝΓ, καθώς και η αποκλειστική χρήση του βίντεο στην εκπαίδευση και την επικοινωνία των κωφών, έχουν δημιουργήσει συνθήκες αποκλεισμού που μόνο η επόμενη γενιά τεχνολογιών θα μπορέσει να αντιμετωπίσει με ικανοποιητικό τρόπο.

Υπό αυτές τις συνθήκες, οι κωφοί μαθητές αναγκάζονται να μαθαίνουν τον ελληνικό γραπτό λόγο ως δεύτερη γλώσσα, εφόσον εκ γενετής δεν έχουν την εμπειρία του προφορικού λόγου. Επίσης, είναι ευρέως γνωστό ότι οι κωφοί μαθητές σπάνια αναπτύσσουν την ίδια ικανότητα ανάγνωσης με τους συνομηλίκους τους ακούοντες μαθητές (Mayberry, 1993).

Οι περισσότεροι άνθρωποι, εφόσον δεν έχουν προσωπικές ή κοινωνικές επαφές με κωφούς (Κατσογιάννου, 2002), έχουν την τάση να πιστεύουν ότι οι νοηματικές γλώσσες είναι ένα είδος παντομίμας ή αναπαράστασης κάποιας από τις φωνούμενες γλώσσες που μιλάνε οι ίδιοι. Η αλήθεια όμως είναι πολύ διαφορετική από αυτήν την ευρέως διαδεδομένη αντίληψη. Οι νοηματικές γλώσσες, οι οποίες σημειωτέον είναι πολλές και διαφορετικές μεταξύ τους, διαφέρουν από τις υπόλοιπες φυσικές γλώσσες μόνο ως προς το ότι

[1] https://www.omke.gr/

μας είναι λιγότερο γνωστές και όχι ως προς τις γλωσσολογικές αρχές που διέπουν την επικοινωνιακή τους λειτουργία.

Η συνεχής επέκταση του διαδικτύου, τόσο στο πλαίσιο της βιομηχανίας και του ακαδημαϊκού κόσμου όσο και στο σπίτι, έχει συντελέσει σημαντικά στο να μπορούν τα άτομα με προβλήματα ακοής να πληροφορούνται online στη νοηματική γλώσσα τους. Παρόλο που οι κωφοί ή βαρήκοοι μπορούν εύκολα να κατανοήσουν το μεγαλύτερο ποσοστό των δακτυλογραφημένων πληροφοριών που δημοσιεύονται στο διαδίκτυο, τα πολυμέσα και οι τηλεδιασκέψεις μέσω υπολογιστή αποτελούν πεδία όπου η νοηματική γλώσσα δεν υποστηρίζεται. Ένα απλό παράδειγμα αποκλεισμού των κωφών ή βαρήκοων ατόμων αποτελεί το σενάριο όπου ένας ακούων ομιλητής κάνει μια προφορική διάλεξη μέσω διαδικτύου σε ένα γεωγραφικά διασκορπισμένο ακροατήριο που περιλαμβάνει κωφούς και δεν υπάρχει δυνατότητα διερμηνείας στη νοηματική γλώσσα.

Για να δημιουργήσει κανείς ψηφιακή αναπαράσταση νοηματικής γλώσσας, θα πρέπει να λάβει υπόψη ότι μια νοηματική γλώσσα είναι οπτική και δημιουργείται από τους ομιλητές της μέσω των κινήσεων των χεριών, του σώματος και της κεφαλής, με σημαντική συμμετοχή των εκφράσεων του προσώπου και του βλέμματος. Μια πιθανή μέθοδος για την ψηφιακή αναπαράσταση της γλώσσας νοημάτων είναι η δημιουργία ενός τρισδιάστατου μοντέλου ενός εικονικού νοηματιστή. Για τον σκοπό αυτό, είναι ουσιαστική η κατανόηση των εννοιών και των αρχών του τρισδιάστατου προγραμματισμού.

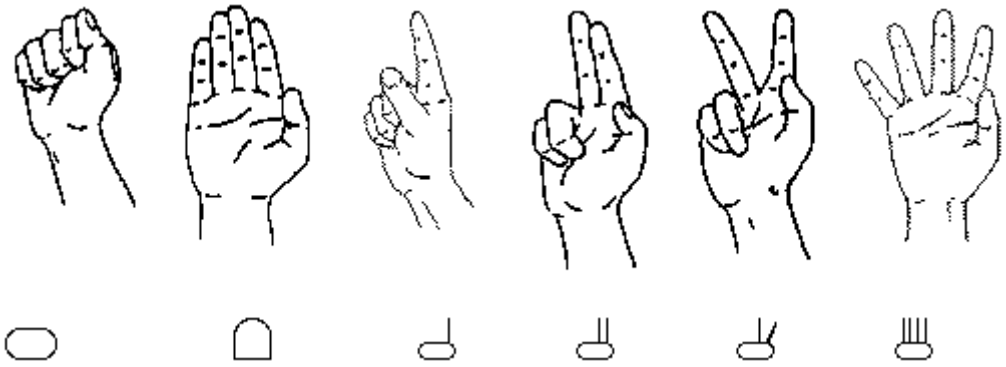
Έχουμε ήδη αναφέρει ότι υπάρχουν πολλές διαφορετικές νοηματικές γλώσσες ανά γεωγραφική περιοχή. Δύο από τις πιο χαρακτηριστικές νοηματικές γλώσσες για την ανάδειξη των μεταξύ τους διαφορών είναι η Αμερικανική Νοηματική Γλώσσα (ASL) και η Βρετανική Νοηματική Γλώσσα (BSL), οι οποίες αποτελούν δύο εντελώς διαφορετικά γλωσσικά συστήματα. Όπως συμβαίνει και με τις προφορικές γλώσσες, οι εθνικές νοηματικές γλώσσες μπορούν να υποδιαιρεθούν περαιτέρω σε τοπικές παραλλαγές, με νοήματα που διαφέρουν μορφολογικά για την ίδια λέξη.

2.1.4 Γλωσσικά μέσα της Νοηματικής Γλώσσας

Τα γλωσσικά μέσα που χρησιμοποιεί η ΕΝΓ, όπως και άλλες νοηματικές γλώσσες, για να διατυπώσει τις έννοιες και να δημιουργήσει μορφολογία και σύνταξη βασίζονται στην κίνηση των χεριών, στη στάση ή την κίνηση του σώματος και στην έκφραση του προσώπου. Οι βασικές μονάδες του λόγου της ΕΝΓ, τις οποίες η επιστήμη της γλωσσολογίας ονομάζει γλωσσικά σημεία, ονομάζονται νοήματα. Τα νοήματα μπορούν να έχουν λεξική ή γραμματική σημασία, ακριβώς όπως τα μορφήματα και οι λέξεις στις φυσικές γλώσσες.

Λόγω της ιδιαίτερης άρθρωσης των νοηματικών γλωσσών στον τρισδιάστατο χώρο, η ΕΝΓ δεν διαθέτει επίσημο σύστημα μεταγραφής ή σημειολογίας. Ωστόσο, έχουν γί-

νει αρκετές προσπάθειες, σε διεθνές επίπεδο, να δημιουργηθούν διάφορα συστήματα μεταγραφής ή σημειολογίας για τις ΝΓ, τα οποία χρησιμοποιούνται κυρίως σε πειραματικά περιβάλλοντα. Ενδεικτικά, δύο τέτοια συστήματα που παρουσιάζονται περισσότερο ως "φωνητικά" συστήματα ("phonetic" systems), είναι το HamNoSys: Hamburg Notation System (Prillwitz et al., 1989) (Σχήμα 2.1) και το Stokoe Notation System που αναπτύχθηκε για την Αμερικανική Νοηματική Γλώσσα (Stokoe et al., 1976). Τα συστήματα αυτά μπορούν να μεταγράψουν με ειδικά σύμβολα (transcription symbols) οποιαδήποτε ΝΓ στον κόσμο.



Σχήμα 2.1 Χειρομορφές σε σημειογραφία του συστήματος Hamnosys

Αν και οι χειρονομίες χρησιμοποιούνται σχεδόν ασυναίσθητα ως συμπλήρωμα της προφορικής γλώσσας για να τονίσουν τη σημασία και να προσθέσουν έμφαση στην ομιλούμενη φράση, διαφέρουν πολύ σε σχέση με την πολυπλοκότητα και την εκλέπτυνση των χειρομορφών των νοηματικών γλωσσών που χρησιμοποιούνται από τον κωφό ομιλητή. Η χειρονομία είναι η φυσική μέθοδος επικοινωνίας για τους κωφούς ανθρώπους, καθώς χωρίς την ικανότητα της ακοής, η εκμάθηση της ομιλίας είναι πολύ δύσκολη. Επομένως, αντί αυτής, οι κωφοί άνθρωποι χρησιμοποιούν ολόκληρο το σώμα τους για να επικοινωνούν. Όπως όλες οι φυσικές γλώσσες, έτσι και οι νοηματικές γλώσσες γεννήθηκαν μέσω της αλληλεπίδρασης στις τοπικές κοινωνίες κωφών.

Στις νοηματικές γλώσσες, η μικρότερη γλωσσική μονάδα είναι το μόρφημα. Τα νοήματα, δηλαδή τα αντίστοιχα των λέξεων στις φωνούμενες γλώσσες, αποτελούνται από ένα ή περισσότερα μορφήματα. Κάθε νοηματικό μόρφημα/νόημα αποτελείται από συνδυασμό χαρακτηριστικών γνωρισμάτων ως προς τη μορφή των χεριών και τη θέση στην οποία αρθρώνεται (πάνω ή κοντά στο σώμα), τον προσανατολισμό της παλάμης και την κίνηση που το χαρακτηρίζει, καθώς και κάποιων μη χειροκινησιακών χαρακτηριστικών που πιθανώς να εμπλέκονται στον σχηματισμό του νοήματος (π.χ. κίνηση κεφαλής, παρακολούθηση βλέμματος, κίνηση φρυδιών, μορφή των χειλιών κλπ.) (Ευθυμίου and Φωτεινέα, 2006).

Ακολουθεί μια σύντομη παρουσίαση των σημαντικότερων χαρακτηριστικών γνωρισμάτων της ΕΝΓ που καταγράφονται και στη μελέτη άλλων νοηματικών γλωσσών και τα οποία θα μπορούσαν να χαρακτηριστούν ως καθολικά χαρακτηριστικά των γλωσσών που αρθρώνονται στον τρισδιάτατο χώρο (Sutton-Spence and Woll, 1999).

Η σειρά των νοημάτων στον νοηματικό λόγο (Sign Order): Η ΕΝΓ παρουσιάζει τη δομή θέμα-σχόλιο (topic-comment structure), στην οποία νοηματίζεται πρώτα η κύρια πληροφορία ή το κεντρικό θέμα. Το θέμα είναι το πλαίσιο μέσα στο οποίο πραγματοποιείται η απόφανση (predication). Αφότου έχει προσδιοριστεί το θέμα, νοηματίζεται το υπόλοιπο της πρότασης που είναι το σχόλιο, καθώς και οι νέες πληροφορίες πάνω σε αυτό. Επιπλέον, η ΕΝΓ δεν έχει καμία σταθερή διάταξη των βασικών στοιχείων της πρότασης. Αυτή η ευελιξία οφείλεται στις πρόσθετες μορφολογικές πληροφορίες που φέρουν τα κατευθυντικά ρήματα, καθώς και στην παρακολούθηση με το βλέμμα (eye-gaze).

Χώρος νοηματισμού (Signing Space), Τοποθέτηση (Placement) και Αντωνυμικό σύστημα (Pronouns): Στην ΕΝΓ και σε πολλές άλλες ΝΓ, οι νοηματιστές εκμεταλλεύονται τον αποκαλούμενο χώρο νοηματισμού μπροστά από το σώμα τους. Κατά τη διαδικασία της ομιλίας τα συστατικά μιας περιγραφής μπορούν να τοποθετηθούν στον χώρο νοηματισμού, δηλαδή πρώτα καθορίζεται μια συγκεκριμένη περιοχή και έπειτα όλα τα στοιχεία ή οι ενέργειες που συσχετίζονται με εκείνη την περιοχή. Κατά συνέπεια, η ΕΝΓ έχει περισσότερες αντωνυμίες από γλώσσες όπως η ελληνική, οι οποίες αρθρώνονται με την υπόδειξη μιας θέσης που έχει συνδεθεί προηγουμένως με ένα ουσιαστικό. Αυτό σημαίνει επίσης ότι, για παράδειγμα τα ελληνικά είναι (υπο)προσδιορισμένα όταν χρησιμοποιούν έναν μορφολογικό τύπο για τη δήλωση του πληθυντικού, ενώ αντίθετα η ΕΝΓ μπορεί να εκφράσει σχέσεις όπως εμείς-ΔΥΟ, εμείς-ΤΡΕΙΣ, και ούτω καθεξής, που βοηθούν στη διάκριση μεταξύ του συνυπολογισμού ή του αποκλεισμού του/των "ακροατή/ών" (μη-νοηματιστών).

Κατευθυντικά ρήματα ή ρήματα συμφωνίας (Directional or Agreement Verbs): Τα κατευθυντικά ρήματα ή ρήματα συμφωνίας περιλαμβάνουν τις πληροφορίες για το πρόσωπο και τον αριθμό του θέματος και του αντικειμένου. Πραγματοποιούνται με την κίνηση του ρήματος στον συντακτικό χώρο, στον οποίο το θέμα και το αντικείμενο τοποθετούνται γύρω από τον νοηματιστή. Το νόημα του ρήματος αρχίζει από τη θέση του θέματος και τελειώνει στη θέση του αντικειμένου (π.χ. ΔΙΝΩ, ΛΕΩ, κλπ.), ενώ μερικά ρήματα αρχίζουν στο αντικείμενο και τελειώνουν στο θέμα (π.χ. ΔΑΝΕΙΖΩ).

Ταξινομητές (Classifiers): Οι ταξινομητές είναι οι χειρομορφές που μπορούν να δείξουν ένα αντικείμενο από μια ομάδα σημασιολογικά σχετικών αντικειμένων. Χρησιμοποιούνται με τα ρήματα που απαιτούν έναν ταξινομητή, έτσι ώστε όταν ο ταξινομητής συνδυάζεται με τη θέση, τον προσανατολισμό, την κίνηση και τα μη χειροκινησιακά χαρακτηριστικά γνωρίσματα, το σύνθετο να διαμορφώνει ένα κατηγόρημα. Η χειρομορφή χρησιμοποιείται για να δείξει αναφορά σε μια κατηγορία (κλάση/ομάδα) αντικειμένων που έχουν παρόμοια χαρακτηριστικά γνωρίσματα (π.χ. κυκλικό σωληνοειδές αντικείμενο, επίπεδη επιφάνεια, αντικείμενα σε κάθετη ή οριζόντια παράταξη κλπ.).

Γραμμές χρόνου (Time lines): Η ΕΝΓ δεν διαθέτει σύστημα χρόνου όπως οι φωνούμενες γλώσσες. Προκειμένου να εκφραστούν οι χρονικές πληροφορίες που περιέχουν τα μορφολογικά ή τα συντακτικά χαρακτηριστικά γνωρίσματα που συνδέονται με τα ρήματα, γίνεται χρήση τεσσάρων χρονικών γραμμών (time lines) στον χώρο νοηματισμού ή δηλώνονται με τη σειρά των προτάσεων μέσα στην ομιλία.

Σχηματισμός του νοήματος

Το χαρακτηριστικότερο συστατικό ενός νοήματος είναι η χειρομορφή. Η χειρομορφή είναι το σχήμα που παίρνει η παλάμη και η θέση στην οποία τοποθετούνται τα δάχτυλα τη στιγμή που αρχίζει να σχηματίζεται ένα νόημα. Η χειρομορφή όμως από μόνη της δεν είναι φορέας σημασίας. Για να αποκτήσει σημασία, για να δημιουργηθεί δηλαδή ένα νόημα, η χειρομορφή πρέπει να συνοδεύεται και από τα παρακάτω χαρακτηριστικά:

1. Τον "προσανατολισμό" της παλάμης, δηλαδή την κατεύθυνση προς την οποία στρέφεται η χειρομορφή κατά το σχηματισμό του νοήματος. Ο δείκτης που δείχνει προς τα πάνω ή στρέφεται προς τα δεξιά αποτελεί τμήμα διαφορετικών νοημάτων.

2. Τη θέση της χειρομορφής στον χώρο ή επάνω στο σώμα. Τα νοήματα παράγονται σε καθορισμένο χώρο που λέγεται χώρος νοηματισμού. Ο χώρος αυτός αντιστοιχεί περίπου σε ένα τετράγωνο που ορίζεται από την κορυφή της κεφαλής ως τον άνω κορμό και εκτείνεται περίπου 30 εκατοστά δεξιά και αριστερά από τα μπράτσα. Αν χρησιμοποιήσουμε μια χειρομορφή έξω από τον χώρο αυτό, π.χ. με τα μπράτσα κρεμασμένα δίπλα στο σώμα, το αποτέλεσμα δεν είναι αναγνωρίσιμο ως νόημα.

3. Την κίνηση του χεριού, χωρίς την οποία δεν μπορεί να ολοκληρωθεί ένα νόημα. Ο δείκτης που δείχνει προς τα πάνω ή στρέφεται προς τα δεξιά χωρίς να κινείται δεν είναι ολοκληρωμένο νόημα, δεν αντιστοιχεί δηλαδή σε ορισμένη σημασία. Εκτός από τη συμμετοχή της στο σχηματισμό του νοήματος, η κίνηση μπορεί να είναι

και φορέας άλλων εννοιών, για παράδειγμα να δηλώνει τον αριθμό (ενικό ή πληθυντικό), το μέγεθος ενός αντικειμένου (μικρότερο ή μεγαλύτερο), ακόμα και τη συχνότητα μιας ενέργειας.

4. Τη στάση (ή κίνηση) του σώματος και/ή την έκφραση του προσώπου, που αποτελούν επίσης συστατικά του νοήματος, με την έννοια ότι λειτουργούν για να μεταφέρουν πληροφορία όπως αυτή που δηλώνεται από τον τόνο της φωνής στις ομιλούμενες γλώσσες. Για παράδειγμα, η έννοια του μέλλοντος στην ΕΝΓ διατυπώνεται συνδυάζοντας το νόημα με μία ελαφρά κλίση του σώματος προς τα εμπρός (Κατσογιάννου, 2002).

Έχουν παρατηρηθεί πάνω από 50 βασικές χειρομορφές που μπορούν να γίνουν με το ένα ή και με τα δύο χέρια. Οι χειρομορφές τοποθετούνται είτε στον χώρο νοηματισμού που προαναφέραμε είτε σε επαφή με το σώμα ή το πρόσωπο. Υπάρχουν γύρω στις 25-30 κοινές θέσεις στον χώρο νοηματισμού για τις χειρομορφές, με καθεμία να έχει διαφορετική έννοια. Την έννοια ενός νοήματος αλλάζει ο προσανατολισμός της παλάμης και των δακτύλων, με αποτέλεσμα η ίδια χειρομορφή να δίνει διαφορετικές έννοιες ανάλογα με τη θέση της παλάμης. Οι μετακινήσεις μέσω του διαστήματος νοηματικής είναι επίσης ξεχωριστές σε κάθε νόημα. Η μεταβολή οποιωνδήποτε από αυτά τα τέσσερα στοιχεία αλλάζει και την έννοια του νοήματος.

Στην παρακάτω εικόνα παρουσιάζονται οι 53 χειρομορφές (Σχήμα 2.2) της ΕΝΓ όπως έχουν καταγραφεί στο λεξικό ΝΟΗΜΑ (Ευθυμίου et al., 2000):

Όπως αναφέρθηκε και προηγουμένως, η ΝΓ χρησιμοποιεί το πρόσωπο, το κεφάλι, και το σώμα, καθώς επίσης και τα χέρια και τους βραχίονες. Ένα νόημα που χρησιμοποιεί ένα άλλος μέρος του σώματος εκτός από τα χέρια καλείται πολυκαναλικό ή πολυτροπικό νόημα (multi-channel/multi-modal sign).

Δακτυλικό αλφάβητο

Ο δακτυλικός συλλαβισμός (Σχήμα 2.3) δεν αποτελεί μορφή της ΝΓ. Είναι μια κωδικοποίηση μεταγραφής ενός αλφαβήτου σε χειρομορφικά σύμβολα, όπου κάθε γράμμα του αλφάβητου αντιπροσωπεύεται από μια χειρομορφή. Ο δακτυλικός συλλαβισμός μπορεί να χρησιμοποιηθεί για να αναπαραστήσει αρτικόλεξα και κύρια ονόματα ή τοπωνύμια για τα οποία δεν υπάρχουν συγκεκριμένα νοήματα. Κατά τον νοηματισμό ονόματος μιας τοποθεσίας, το νόημα συχνά θα περιλαμβάνει μια συλλαβισμένη με τα δάχτυλα σύντμηση του ονόματος μαζί με χρήση ανάγνωσης των χειλιών. Με βάση αυτόν τον μηχανισμό δημιουργούνται και νέα νοήματα, όπως τα νοήματα για κύρια ονόματα, αλλά και λέξεις που δεν είχαν προηγούμενη νοηματική απεικόνιση και συχνά χαρακτηρίζονται από την

Σχήμα 2.2 Άποψη χειρομορφών της ΕΝΓ

επανάληψη του συλλαβισμένου δακτυλικού νοήματος για το πρώτο γράμμα της λέξης (π.χ. ΚΟΥΖΙΝΑ, ΒΙΒΛΟΣ κ.α.).

2.2 Συστήματα αναπαράστασης της Νοηματικής Γλώσσας

Στη ενότητα αυτή θα παρουσιάσουμε τεχνολογίες για την ψηφιακή αναπαράσταση και απόδοση της πληροφορίας στη ΝΓ. Παρακάτω αναφέρουμε συνοπτικά τις κυριότερες τεχνολογίες και μεθόδους που χρησιμοποιούνται αυτή τη στιγμή στην παγκόσμια κοινότητα.

Βίντεο της Νοηματικής Γλώσσας: Η νοηματική γλώσσα μπορεί να παραχθεί σε μια οθόνη υπολογιστών χρησιμοποιώντας κινούμενη εικόνα (ακολουθίες εικόνων). Οι ακολουθίες δημιουργούνται από τη βιντεοσκόπηση ενός χρήστη της νοηματικής που εκτελεί ένα νόημα και έπειτα την ψηφιοποίηση του βίντεο προς την παραγωγή ενός υλικού σε μορφή QuickTime ή AVI και άλλες κωδικοποιήσεις. Αυτά τα πρότυπα, όπως και άλλα πολλά (π.χ. Real Player, Mpeg, Wmv, Swf) μπορούν έπειτα να αναπαραχθούν είτε σε έναν υπολογιστή είτε σε απευθείας σύνδεση (online) είτε κατά απαίτηση (on video demand).

Σχήμα 2.3 Ελληνικό δακτυλικό αλφάβητο

Τα βίντεο νοηματικής γλώσσας θα μπορούσαν να χρησιμοποιηθούν για να αντιπροσωπεύσουν τη γλώσσα νοημάτων με τη σύλληψη μιας ακολουθίας βίντεο για κάθε λέξη.

Σκίτσα - Στατικά Δισδιάστατα Γραφικά: Μια άλλη μέθοδος αναπαράστασης της ΝΓ είναι τα στατικά αρχεία εικόνας του νοήματος. Η κίνηση για κάθε νόημα μπορεί να απεικονιστεί μέσω των κατευθυνόμενων βελών. Έτσι δημιουργούνται αναπαραστάσεις της νοηματικής, συνήθως σχεδιασμένες με το χέρι. Παρατίθεται ένα παράδειγμα (Σχήμα 2.4) από το βιβλίο "Εγχειρίδιο Νοηματικής Γλώσσας" (Μαγγανάρης, 2002).

Σχήμα 2.4 Καλημέρα στην ΕΝΓ

Κινούμενα Δισδιάστατα Γραφικά– 2D Μια εναλλακτική λύση στα παραπάνω στατικά γραφικά, είναι να δοθεί κίνηση σε κάθε μέρος του σώματος του νοηματιστή ξεχωριστά. Ένα λογισμικό όπως το "ANIMATE CC"[1] μπορεί να δημιουργήσει μια ποικιλία κινούμενου περιεχομένου, όπως κινούμενα σχέδια για δημοσίευση σε πλατφόρμες όπως οι HTML 5 Canvas, Flash Player & Air, WebGSL ή custom πλατφόρμες όπως η Snap SVG.

Το μέγεθος χρόνου που απαιτεί η δημιουργία εμψύχωσης (animation) για κάθε μεμονωμένο μέρος του σώματος θα μπορούσε να ελαχιστοποιηθεί με τη χρησιμοποίηση της τεχνικής που ονομάζεται "tweening". Το "tweening" είναι μια μέθοδος υπολογιστικής εμψύχωσης με την οποία ο χρήστης καθορίζει καρέ-κλειδιά (keyframes) και έπειτα ο υπολογιστής παράγει τα απαιτούμενα καρέ και τα ταξινομεί έτσι ώστε να κινούνται διαδοχικά αυτόματα. Ένα "keyframe" είναι η περιγραφή μιας κατάστασης για οτιδήποτε εμψυχώνεται. Περιέχει τις πληροφορίες για όλες τις ιδιότητες του αντικειμένου, όπως θέση, μέγεθος και χρώμα. Υπάρχουν διαφορετικές μέθοδοι "tweening" κίνησης, από απλοϊκές έως πολύ σύνθετες. Η απλούστερη μέθοδος είναι η γραμμική. Κατά τη χρησιμοποίηση της γραμμικής μεθόδου, η θέση του σημείου που κινείται στο αρχικό πλαίσιο αφαιρείται από τη θέση του σημείου που κινείται στο τελικό πλαίσιο. Από τη διαίρεση αυτής της διαφοράς με το ποσό πλαισίων που χρειάζονται να παραχθούν, προκύπτει το ποσό κίνησης που

[1] https://helpx.adobe.com/animate/user-guide.html

απαιτείται σε κάθε ενδιάμεσο πλαίσιο. Οι πιο περίπλοκες μέθοδοι χρησιμοποιούν σύνθετες μαθηματικές εξισώσεις για την παραγωγή keyframes, με εξισώσεις που παράγουν μια προσέγγιση της επιθυμητής κίνησης.

Τρισδιάστατα Γραφικά: Πριν εξερευνήσουμε τον τρισδιάστατο προγραμματισμό, πρέπει πρώτα να ρίξουμε μια ματιά στην εξέλιξη των τρισδιάστατων γραφικών. Τα τρισδιάστατα γραφικά προέκυψαν περίπου στη δεκαετία του '70 από την επιθυμία δημιουργίας εικονικών μοντέλων των αντικειμένων του πραγματικού κόσμου. Ειδικά σχεδιαστικά πακέτα (Computer Aided Design (CAD)) αναπτύχθηκαν για χρήση στη βιομηχανία σχεδίου, επιτρέποντας στους χρήστες να δημιουργήσουν και να χειριστούν τα μοντέλα σε τρισδιάστατο επίπεδο και έπειτα να τα εμφανίσουν στις οθόνες μέσω δισδιάστατων προβολών. Μέχρι τη δεκαετία του '90, τα τρισδιάστατα γραφικά παρέμεναν ένα πεδίο για εξειδικευμένους χρήστες, που δραστηριοποιούνταν κυρίως στην αυτοκινητοβιομηχανία. Η τρισδιάστατη τεχνολογία μπήκε για πρώτη φορά στα σπίτια μας όταν, το 1992, η IMB παρήγαγε το τρισδιάστατο παιχνίδι "Wolfenstein" για Η/Υ, που χαρακτήρισε την ψευδοτρισδιάστατη γραφική παράσταση. Η ανάγκη για ταχύτερη ανάπτυξη οδήγησε σε συγκεκριμένες προγραμματιστικές διαπροσωπείες (Application Program Interfaces (APIs)) για τον προγραμματισμό σε τρισδιάστατο επίπεδο. Οι APIs επέκτειναν τις υπάρχουσες γλώσσες προγραμματισμού με την παροχή λειτουργιών δημιουργίας 3D αντικειμένων. Πολλές από αυτές τις APIs είναι διαθέσιμες σήμερα, συμπεριλαμβανομένων των Direct3D της Microsoft, OpenGL της Silicon Graphics και Java3D της Sun Microsystems. Οι πιο προηγμένες λειτουργίες των APIs μπορούν να χρησιμοποιηθούν για να προσθέσουν ρεαλισμό στις σκηνές. Οι προβολές των αντικειμένων που δημιουργήθηκαν με τη χρησιμοποίηση αυτών των APIs δίνουν μια πειστική αίσθηση τρισδιάστατου κόσμου, αλλά ενδεχομένως να στερούνται ρεαλισμού. Μια άλλη μέθοδος παραγωγής τρισδιάστατων γραφικών είναι η γλώσσα VRML (Virtual Reality Modeling Language). Η VRML εμφανίστηκε αρχικά ως επέκταση της Hypertext Markup Language (HTML), με σκοπό να παρέχει την τρισδιάστατη λειτουργία. Εκτός από γλώσσα προγραμματισμού, η VRML είναι και γλώσσα περιγραφής σκηνής που περιγράφει έναν τρισδιάστατο εικονικό κόσμο. Η VRML παρέχει παρόμοιες αλλά λιγότερο περίπλοκες λειτουργίες από μια τρισδιάστατη API.

Τα τελευταία χρόνια έχουν αναπτυχθεί ενδιαφέρουσες τεχνολογίες που βασίζονται στη μελέτη των μηχανικών νόμων σχετικά με την κίνηση ή τη δομή των ζωντανών οργανισμών (kinematic). Για παράδειγμα αναφέρουμε τη γλώσσα STEP (Scripting Technology for Embodied Persona) η οποία αποτελεί το ενδιάμεσο επίπεδο επικοινωνίας μεταξύ του χρήστη και του ανθρωποειδούς (avatar). Επίσης, υπάρχει το σύστημα της γλώσσας SiGML

που τροφοδοτεί ένα ανθρωποειδές του πανεπιστημίου "East Anglia" (UEA)[1] (Efthimiou et al., 2010), προκειμένου να παρουσιαστεί το αποτέλεσμα του συστήματος μηχανικής μετάφρασης.

Κλείνοντας, θα πρέπει να επισημάνουμε ότι οι τελευταίες εξελίξεις στην τεχνολογία των 3D ανθρωποειδών διευκολύνουν εξαιρετικά τη δημιουργία και τη διαχείριση δεδομένων ΝΓ, με απώτερο σκοπό τη δυναμική σύνθεση νοημάτων και νοηματικών φράσεων, πέρα από τη χρήση του βίντεο, για την απεικόνιση του νοηματικού λόγου.

Πικτογραφικά συστήματα της ΝΓ: Το πικτογραφικό σύστημα της ΝΓ αποτελεί ένα σύστημα γραπτής μεταγραφής της ΝΓ. Πρόκειται για ένα σύστημα σημειογραφίας ή γραπτών συμβόλων που αναπαριστούν τις χειρομορφές και τις κινήσεις τους σε γραπτή μορφή (Notation Systems for Signed Languages).

Η έλλειψη συστημάτων μεταγραφής της ΝΓ που σημειώνεται από την αρχή της εμφάνισης νοηματικών γλωσσών ως συστημάτων επικοινωνίας σε κοινωνίες κωφών πληθυσμών, δημιούργησε σημαντικές ελλείψεις και κενά στην έρευνα της ΝΓ. Με την εξέλιξη της τεχνολογίας άρχισαν τα πρώτα δείγματα καταγραφής μέσω αναλογικών οπτικοακουστικών μέσων, αλλά και πάλι η χρήση βίντεο δεν προσφέρει δυνατότητες εύκολης επεξεργασίας. Έτσι, η διάδοση των νοηματικών γλωσσών με τη χρήση ψηφιακών βίντεο παραμένει ακόμα και σήμερα δύσκολη. Με την εξέλιξη των τηλεπικοινωνιών, η μεταφορά ψηφιακού βίντεο μέσω διαδικτύου είναι πλέον δεδομένη, αλλά τα ψηφιακά βίντεο δεν προσφέρουν τα αντίστοιχα πλεονεκτήματα που έχει ο γραπτός λόγος. Επομένως, κρίθηκε επιτακτική η ανάγκη δημιουργίας ενός συστήματος σημειογραφίας της ΝΓ που να είναι αποδεκτό από την παγκόσμια κοινότητα.

Να επισημάνουμε ότι η σημειογραφία (πικτογραφικό σύστημα) της ΝΓ είναι πολύ σημαντική, καθώς χάρη σε αυτή μπορεί κανείς, για αρχή, να πραγματοποιήσει ενέργειες σχετικές με τη ΝΓ που πριν την ανάπτυξη της σημειογραφίας δεν ήταν εφικτές, όπως:

- να διαβάσει μια Νοηματική Γλώσσα
- να γράψει μια Νοηματική Γλώσσα
- να μελετήσει στοιχεία μιας Νοηματικής Γλώσσας
- να καταγράψει στοιχεία πολιτισμού σε Νοηματική Γλώσσα

Παρ' όλα αυτά, τα συστήματα που αναπτύχθηκαν τις τελευταίες δεκαετίες δεν έγιναν ευρέως αποδεκτά, επειδή δεν καταφέρνουν να μεταφέρουν την τρισδιάστατη άρθρωση με τρόπο που μπορεί να θεωρηθεί "φυσικός" από τους νοηματιστές. Αντίθετα,

[1] http://vh.cmp.uea.ac.uk

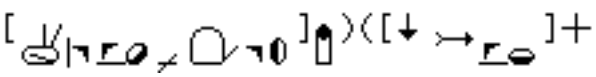
τα συστήματα μεταγραφής ΝΓ χρησιμοποιήθηκαν σχεδόν αποκλειστικά από ερευνητικές/επιστημονικές κοινότητες, ενώ κάθε επιστημονική κοινότητα δημιούργησε το δικό της σύστημα σημειογραφίας.

Στην ενότητα αυτή παρουσιάζουμε τα πιο ενδεικτικά συστήματα σημειογραφίας της ΝΓ, τα οποία είναι:

- SignWriting (Sutton-Spence and Woll, 1999)

- Stokoe System (Baker and Battison, 1980)

- HamNoSys (Prillwitz et al., 1989)



Στον δικτυακό τόπο "Signwritting"[1] υπάρχει αρκετό υλικό σωμάτων (corpora) κειμένων σημειογραφίας, σχετικά άρθρα, μαθήματα, λεξικά, βιβλιοθήκες, καταγραφές σημειογραφίας ανά τον κόσμο, ενώ επίσης μπορεί να βρει κανείς δωρεάν λογισμικό συγγραφής σημειογραφίας ΝΓ και ειδικότερα της Αμερικανικής Νοηματικής Γλώσσας (ASL).

Τέλος, θα ήταν χρήσιμο να αναφερθούμε στο Hamburg Notation System (HamNoSys) Prillwitz et al. (1989), ένα φωνητικό σύστημα σημειογραφίας της ΝΓ για το οποίο μπορεί κανείς να αντλήσει περισσότερες πληροφορίες από την ιστοσελίδα "DGS - Korpus"[2]. Αυτή τη στιγμή, το σύστημα HamNoSys βρίσκεται στην έκδοση 4.0. Υπάρχει επίσης το λογισμικό σημειογραφίας "syncWRITER"[3] που δημιουργήθηκε στο πανεπιστήμιο του Αμβούργου.

2.3 Γλωσσικοί πόροι της Ελληνικής Νοηματικής Γλώσσας

Όπως αναφέραμε ήδη, μια νοηματική γλώσσα είναι σύστημα νευμάτων το οποίο συνδυάζει εκφράσεις του προσώπου και κινήσεις των χεριών και του σώματος που αποτελούν έναν ολοκληρωμένο κώδικα φυσικής ανθρώπινης επικοινωνίας.

Η καταγραφή των νοηματικών γλωσσών σε παγκόσμιο επίπεδο εξακολουθεί να είναι εξαιρετικά ελλιπής και η μελέτη τους ιδιαίτερα περιορισμένη. Είναι προφανές ότι το πρόβλημα αυτό ισχύει και στην περίπτωση της ΕΝΓ, για την οποία η καταγραφή οποιασδήποτε πληροφορίας γινόταν μέχρι και την προηγούμενη δεκαετία μόνο με φωτογραφίες ή σκίτσα, από τα οποία έλειπε ένα βασικό συστατικό των νοημάτων, η κίνηση.

Στην πραγματικότητα, ό,τι έχουμε στη διάθεσή μας από πρώιμες προσπάθειες λεξικογράφησης της ΕΝΓ, οι οποίες αφορούσαν πάντα δίγλωσσα λεξικά (ελληνικής-ΕΝΓ) είναι αποσπασματικές καταγραφές εξαιρετικά περιορισμένου εύρους. Παράλληλα, η διδασκαλία της ΕΝΓ βασίστηκε σε αποσπασματικές προσπάθειες, συχνά ιδιωτικές, ενώ λείπει εντελώς από την εκπαίδευση ένα βιβλίο γραμματικής, ένα συστηματικό εγχειρίδιο περιγραφής και ερμηνείας των κανόνων που διέπουν την ΕΝΓ. Οι ελλείψεις αυτές γίνονται εντονότερες μετά την αναγνώριση της ΕΝΓ από το Ελληνικό Κοινοβούλιο, με το Νόμο 2817/2000 (ΦΕΚ 78/14-3-00).

Η καλύτερη απεικόνιση των στοιχείων μιας ΝΓ επιτυγχάνεται με τη μεταφορά όλων των χαρακτηριστικών της τρισδιάστατης άρθρωσης, απαίτηση που συνήθως ικανοποιείται μέσω της χρήσης βίντεο.

[1] http://www.signwriting.org/

[2] https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/hamnosys-97.html

[3] https://www.sign-lang.uni-hamburg.de/software/syncwriter/info.english.html

Στις μέρες μας, η τελειοποίηση της δυναμικής σύνθεσης νοημάτων και νοηματικών φράσεων με χρήση εικονικού βοηθού αποτελεί ζητούμενο. Παρ' όλα αυτά, η τρέχουσα τεχνολογία επιτρέπει ήδη μεγάλο βαθμό φυσικότητας από την πλευρά του εικονικού νοηματιστή.

Τρέχοντες λεξικογραφικοί πόροι που αφορούν την ΕΝΓ περιλαμβάνουν τόσο εφαρμογές στο διαδίκτυο όσο και εφαρμογές πολυμέσων (CD-ROM, DVD-ROM). Ενδεικτικά αναφέρουμε το βασικότερο λεξικογραφικό έργο για την ΕΝΓ "Το έργο ΝΟΗΜΑ για την Λεξικογράφηση της Ελληνικής Νοηματικής Γλώσσας" (Ευθυμίου et al., 2000) και το διαδικτυακό πολύγλωσσο λεξικό Dicta-Sign (Efthimiou et al., 2010) που αναπτύχθηκε στο πλαίσιο του ευρωπαϊκού έργου Dicta-Sign.

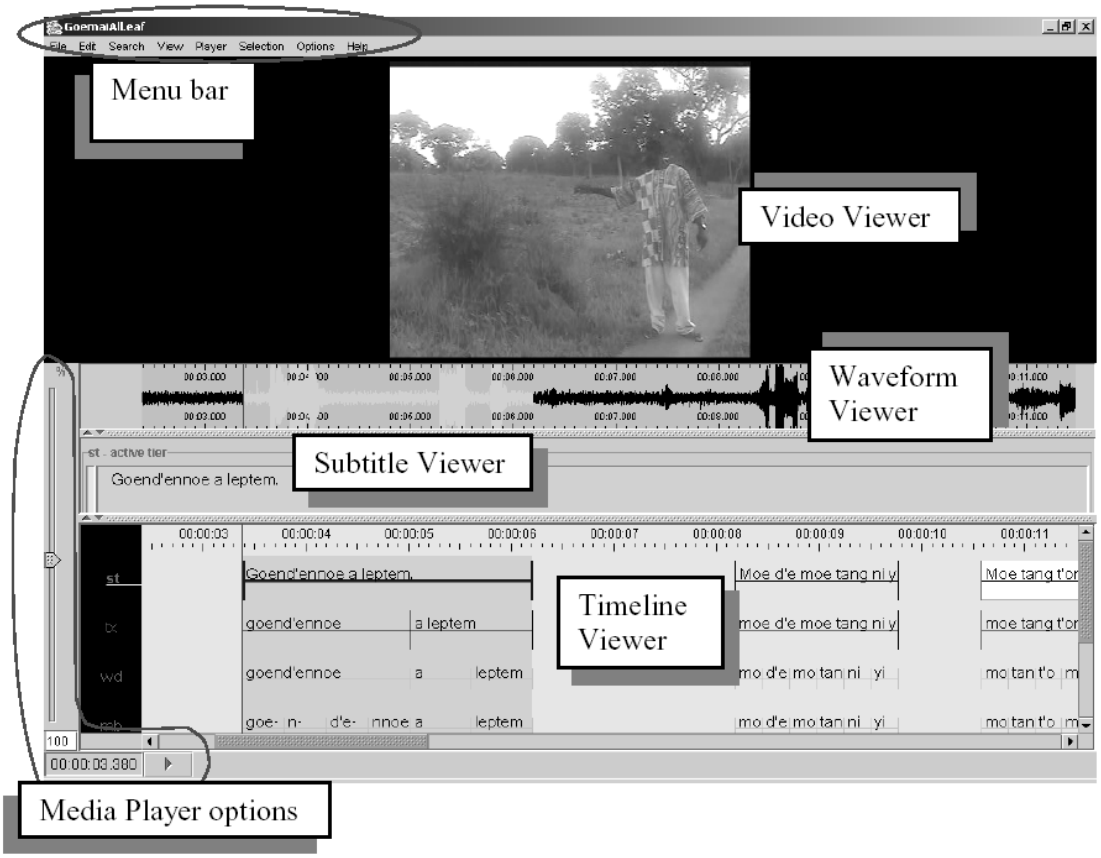
2.4 Εργαλεία επισημείωσης βίντεο-σωμάτων της ΕΝΓ

2.4.1 Εργαλείο ELAN

Ένα εργαλείο επισημείωσης (annotation tool) μπορεί να δημιουργήσει, να εκδώσει, να απεικονίσει και να αναζητήσει τις επισημειώσεις ή τους σχολιασμούς που έχουν προστεθεί σε οπτικά (βίντεο) ή ακουστικά δεδομένα. Υπάρχουν διάφορα λογισμικά επισημείωσης. Ένα από τα πλέον δημοφιλή είναι το λογισμικό ELAN (Hellwig and Van Uytvanck, 2005) (επισημειωτής EUDICO), που αναπτύχθηκε στο Ινστιτούτο Max Planck για την Ψυχογλωσσολογία. Το ELAN σχεδιάστηκε κυρίως για την επισημείωση δεδομένων βίντεο που δεν περιλάμβαναν νοηματική γλώσσα, για λόγους σχολιασμού, ανάλυσης και τεκμηρίωσης οπτικοακουστικών δεδομένων.

Οι δυνατότητες του λογισμικού ELAN περιλαμβάνουν τα εξής:

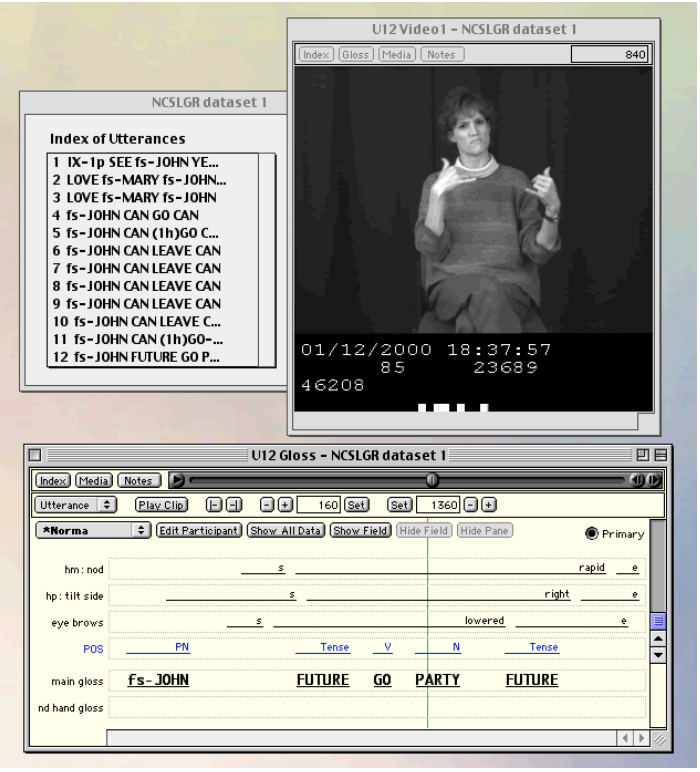
- Ταυτόχρονη επίδειξη σήματος ομιλίας και/ή βίντεο
- Χρονική σύνδεση των σχολιασμών με τη ροή ήχου και/ή εικόνας
- Απεριόριστο αριθμό σχολιασμών, όπως καθορίζεται από τους χρήστες
- Σύνδεση/υπόταξη των σχολιασμών σε άλλους σχολιασμούς
- Διαφορετικά σύνολα χαρακτήρων
- Εισαγωγή και εξαγωγή αρχείων εργασίας μεταξύ ELAN και Shoebox
- Επιλογές αναζήτησης με βάση τις επισημειώσεις



Σχήμα 2.5 Άποψη του λογισμικού ELAN

2.4.2 Ψηφιακή βάση δεδομένων βίντεο νοηματικής γλώσσας Sign Stream

Ένα άλλο γνωστό εργαλείο βάσης δεδομένων ψηφιακών βίντεο είναι το Sign Stream 2.6 που αρχικά αναπτύχθηκε με σκοπό την ανάλυση και καταχώρηση δεδομένων σχετικά με τη δομή (linguistic structure) της Αμερικανικής Νοηματικής Γλώσσας (ASL). Το εργαλείο Sign Stream μπορεί να χρησιμοποιηθεί και για οποιαδήποτε άλλη ΝΓ εκτός της Αμερικανικής.



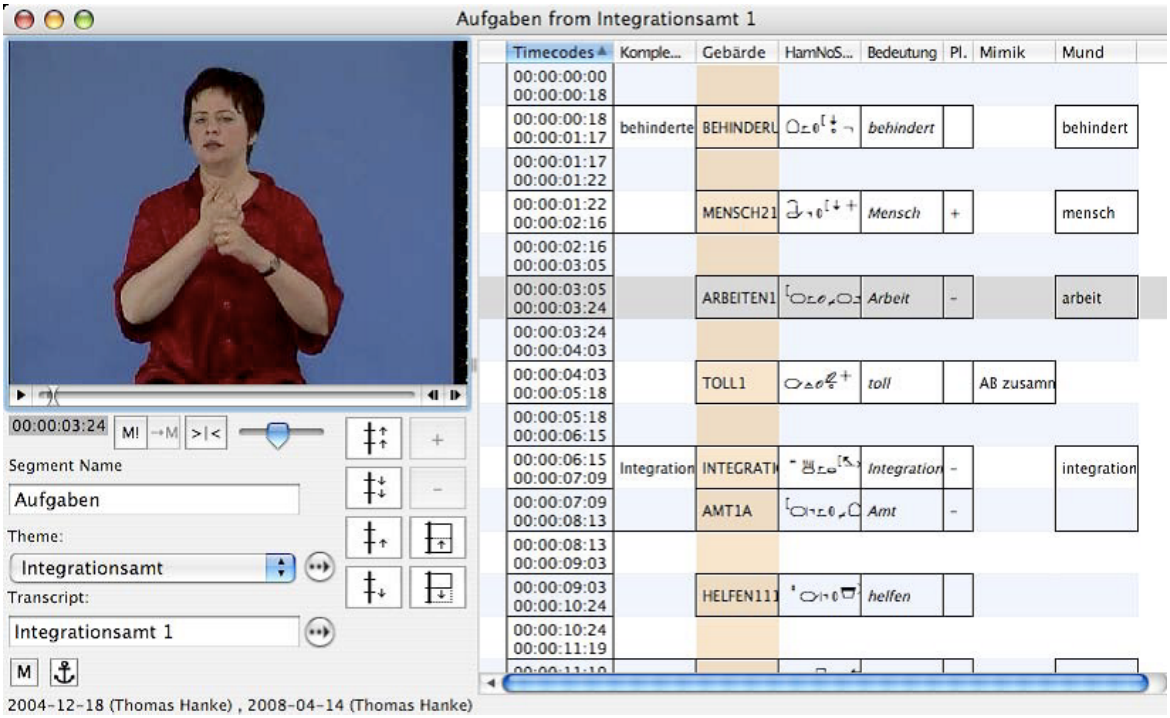
Σχήμα 2.6 Άποψη του λογισμικού SignStream

2.4.3 Εργαλείο iLEX: A tool for Sign Language Lexicography and Corpus Analysis

Το iLex (integrated lexicon) είναι ένα εργαλείο βάσης δεδομένων για τη διαχείριση γλωσσικών πόρων της ΝΓ και Λεξικού της Νοηματικής Γλώσσας. (Hanke, 2002).

Το iLex συνδυάζει δύο προσεγγίσεις, καθώς είναι μια βάση δεδομένων γραπτής μεταγραφής της ΝΓ σε όλη της την πολυπλοκότητα, συνδυασμένη με μια βάση δεδομένων λεξικού. Σήμερα, στο IDGS του Αμβούργου, το εργαλείο iLex δεν χρησιμοποιείται μόνο

στην ανάλυση του λόγου και τη λεξικογραφία, αλλά και σε άλλες περιοχές εφαρμογών που αντλούν δεδομένα από τη βάση του, π.χ. στα avatar projects ViSiCAST και eSIGN όπου αντλούνται δεδομένα από τη βάση iLex και αναπαράγονται από τους εικονικούς νοηματιστές (Hanke, 2002), (Hanke, 2004).



Σχήμα 2.7 iLex, a corpus and lexicography tool

2.5 Μηχανική Μετάφραση και Νοηματική Γλώσσα

Η ΜΜ από ή προς νοηματικές γλώσσες αποτελεί ενεργό αντικείμενο έρευνας, με στόχο να αναπτυχθεί ένα υπολογιστικό σύστημα που να αναιρεί τα εμπόδια της επικοινωνίας μεταξύ των γλωσσικών ομάδων προφορικών και νοηματικών γλωσσών (Efthimiou et al., 2007)). Στον διεθνή χώρο έχουν καταγραφεί σημαντικές ερευνητικές προσπάθειες, όπως αυτές που αφορούν τη μετάφραση από κείμενο σε ΝΓ για την ASL (Huenerfauth, 2007), αλλά και άλλες που αφορούν άλλα γλωσσικά ζεύγη, όπως ερευνητικές προσπάθειες με αντικείμενο τη ΝΓ της Νοτίου Αφρικής (Van Zijl and Combrink, 2006) και την ΕΝΓ (Fotinea et al., 2008), καθώς και πιο πρόσφατες, στο πλαίσιο του ευρωπαϊκού έργου Sign-Speak[1].

Σε κάθε περίπτωση, τα συστήματα που αναφέρονται στη βιβλιογραφία και αφορούν μετάφραση από τη γραπτή μορφή μιας προφορικής γλώσσας σε νοηματική γλώσσα, πε-

[1]http://www.signspeak.eu/

ριλαμβάνουν ένα υποσύστημα μεταφοράς («ταιριάσματος») από τη γλώσσα-πηγή (ΓΠ) στη γλώσσα-στόχο (ΓΣ), προκειμένου να επιτευχθεί το ταίριασμα από τη γραμμική δομή των προφορικών γλωσσών στην πολυεπίπεδη δομή των νοηματικών γλωσσών.

Η δυνατότητα διαχείρισης μεγάλου εύρους γλωσσικών φαινομένων, καθώς και μεγάλου όγκου λεξιλογίου, από τέτοιου τύπου συστήματα είναι συνάρτηση του συνεχούς εμπλουτισμού των συστημάτων αυτών με νέους κανόνες και λεξιλόγιο. Ωστόσο, παραμένει γεγονός ότι τα συστήματα αυτά έχουν καλύτερες επιδόσεις όταν περιλαμβάνουν περιορισμένο μέγεθος γλωσσικών δεδομένων.

Από την άλλη πλευρά, συστήματα που κάνουν χρήση στατιστικών αλγορίθμων είναι ακόμη αδύνατο να χρησιμοποιηθούν, αφού απαιτούν για την εκπαίδευση των μηχανών όγκους επισημειωμένων και παράλληλων ή συγκρίσιμων δεδομένων που προς το παρόν δεν διατίθενται για τις νοηματικές γλώσσες.

2.6 Επίλογος

Τα πεδία της γλωσσικής τεχνολογίας και της τεχνολογίας νοηματικής γλώσσας οδηγούν στην ανάπτυξη συστημάτων που στο μέλλον θα επιτρέψουν τη χρήση της ΝΓ σε πλήθος εφαρμογών στον χώρο της ψηφιακής επικοινωνίας. Η παρούσα διατριβή πραγματεύεται τη δημιουργία μεγάλων παράλληλων σωμάτων της ΕΝΓ, τη μοντελοποίησή τους και την εφαρμογή τους σε συστήματα στατιστικής μετάφρασης.

Κεφάλαιο 3

Μηχανική Μετάφραση και Νοηματικές Γλώσσες

3.1 Μηχανική Μετάφραση

3.1.1 Εισαγωγή

Η γλώσσα είναι ένα αποτελεσματικό μέσο επικοινωνίας. Εκφράζει με σαφήνεια τις ιδέες και τις σκέψεις του ανθρώπινου νου. Το γεγονός ότι υπάρχουν περισσότερες από 5.000 γλώσσες ανά τον κόσμο αντανακλά τη γλωσσική πολυμορφία. Είναι δύσκολο για ένα άτομο να γνωρίζει και να κατανοεί όλες τις γλώσσες του κόσμου. Ως εκ τούτου, η μέθοδος της αυτόματης μετάφρασης δημιουργήθηκε για να γεφυρωθεί η επικοινωνία από τη μια γλώσσα στην άλλη (Tripathi and Sarkhel, 2010).

Η Επεξεργασία Φυσικής Γλώσσας (ή Natural Language Processing - NLP) αποτελεί το σταυροδρόμι της επιστήμης των υπολογιστών, της τεχνητής νοημοσύνης και της γλωσσολογίας και ασχολείται με την ανάλυση και κατανόηση της φυσικής γλώσσας μέσω της χρήσης υπολογιστών.

Η ραγδαία εξέλιξη της Κοινωνίας της Πληροφορίας (ICT) έχει επηρεάσει δραματικά τη διαδικασία της αυτόματης μετάφρασης. Γίνονται συνεχείς προσπάθειες μελέτης των δυνατοτήτων της αυτόματης μετάφρασης από μια συγκεκριμένη γλώσσα (ΓΠ) σε μια άλλη γλώσσα (ΓΣ), σε ερευνητικό επίπεδο. Σήμερα διατίθενται πολλά εργαλεία, είτε επί πληρωμή είτε δωρεάν, τα οποία υποστηρίζουν την αυτόματη μετάφραση ενός κειμένου σε μία ή περισσότερες γλώσσες. Για παράδειγμα, οι διαδικτυακές υπηρεσίες Yandex

translate[1], Babylon online system[2] και Altavista Babelfish [3], οι οποίες παρέχουν απευθείας μετάφραση. Επίσης, οι ζωντανοί διαδικτυακοί μεταφραστές Microsoft Bing Translator[4] και Google Traslator[5] αποτελούν ζωντανά δικτυακά εργαλεία που χρησιμοποιούνται ευρέως για αυτόματη μετάφραση από βιβλιοθηκονόμους και άλλα μέλη της δικτυακής κοινότητας.

Στις μέρες μας, τα συστήματα ΜΜ πειραματίζονται κάνοντας χρήση μεγάλου όγκου κειμενικών δεδομένων στη βάση αξιοποίησης, τα οποία χρησιμοποιούνται όχι μόνο ως παράλληλα κείμενα αλλά και ως αντικείμενο σύγκρισης με άλλα κείμενα που μπορούν να συλλεχθούν με τεχνικές ανεύρεσης από το διαδίκτυο σε πολλές διαφορετικές γλώσσες, όπως στην περίπτωση του προγράμματος Accurat (FP7-ICT)[6]. Επίσης, μια αξιοσημείωτη βάση δεδομένων ανοιχτού περιεχομένου παράλληλων κειμένων είναι η συλλογή OPUS [7] (Tiedemann, 2012) που περιλαμβάνει αναφορές σε ανοιχτές βιβλιοθήκες σωμάτων κειμένων της Ευρωπαϊκής Ένωσης, όπως το Europarl[8] (Koehn, 2004) και άλλες.

Υπήρξαν σημαντικές πρωτοβουλίες από διάφορους ερευνητικούς οργανισμούς και κυβερνητικές υπηρεσίες για την ανάπτυξη εργαλείων αυτόματης μετάφρασης, με σκοπό την επίτευξη μιας ευρύτερης προσέγγισης και τη γεφύρωση του χάσματος της γλωσσικής πολυμορφίας. Ιδιαίτερα χρήσιμη για τους ερευνητές πληροφορικής της μαθηματικής γλωσσολογίας στάθηκε η ύπαρξη στο ευρύτερο διαδίκτυο ενός σημαντικού πλήθους σωμάτων κειμένων της λογοτεχνίας και σε άλλες γλώσσες εκτός από τα αγγλικά. Έτσι, τα εργαλεία αυτόματης μετάφρασης έχουν καταστεί ιδιαίτερα χρήσιμα για τους βιβλιοθηκονόμους, αλλά και γενικά για τους επαγγελματίες της πληροφορικής, ώστε να βελτιώνουν τις εφαρμογές πληροφορικής και να τις καθιστούν προσβάσιμες από περισσότερο κόσμο. Στο παρόν κεφάλαιο θα ασχοληθούμε με διάφορες προσεγγίσεις που έχουν υιοθετηθεί μέχρι σήμερα για την επίτευξη της αυτόματης μετάφρασης των κειμένων.

3.1.2 Μεθοδολογίες Μηχανικής Μετάφρασης

Σε γενικές γραμμές, η ΜΜ αποτελεί ένα υποσύνολο του ευρύτερου πεδίου της Γλωσσικής Τεχνολογίας και ασχολείται με την ανάπτυξη υπολογιστικών εργαλείων, λογισμικών και αλγορίθμων για τη μετάφραση κειμένων ή ομιλίας από τη μια φυσική γλώσσα

[1] https://translate.yandex.com
[2] https://translation.babylon-software.com/english/to-greek/
[3] https://www.babelfish.com/
[4] https://www.bing.com/translator
[5] https://translate.google.com/
[6] http://www.accurat-project.eu/
[7] http://opus.nlpl.eu/
[8] http://statmt.org/europarl/

στην άλλη. Κατά καιρούς έχουν υιοθετηθεί διάφορες μέθοδοι για την αυτοματοποίηση της διαδικασίας της μετάφρασης. Ωστόσο, ο κοινός σκοπός των μεθόδων αυτών είναι "η μεταφορά της έννοιας του κειμένου-πηγή στο κείμενο-στόχο". Ανεξάρτητα από τη μέθοδο που υιοθετείται, η ΜΜ αφορά κατά βάση τις εξής διαδικασίες:

- Αποκωδικοποίηση της έννοιας στη ΓΠ και
- Κωδικοποίηση της ίδιας έννοιας στη ΓΣ

Σε γενικές γραμμές, η διαδικασία της αυτόματης μετάφρασης περιλαμβάνει τα εξής δύο επίπεδα (Tripathi and Sarkhel, 2010):

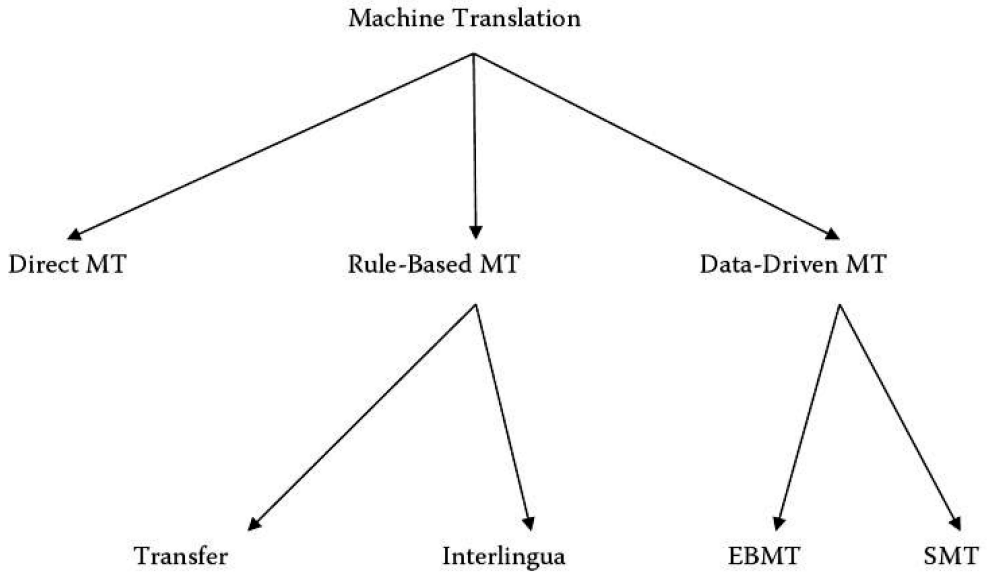
- **Metaphrase**. Σημαίνει "λέξη-προς-λέξη" μετάφραση. Πρόκειται δηλαδή για μια "κατά λέξη" μετάφραση του κειμένου. Ωστόσο, το "κατά λέξη" μεταφρασμένο κείμενο δεν μπορεί να μεταφέρει οπωσδήποτε την έννοια του αρχικού κειμένου. Αυτό σημαίνει ότι μερικές φορές η σημασιολογία μπορεί να διαφέρει από το πρωτότυπο κείμενο.
- **Paraphrase**. Αφορά στη "δυναμική αντιστοιχία" των δύο κειμένων, δηλαδή το μεταφρασμένο κείμενο περιέχει την ουσία του αρχικού κειμένου, οπότε ενδέχεται να μην περιέχει απαραιτήτως τη "λέξη-προς-λέξη" μετάφραση.

Σε κάθε περίπτωση, η ανάπτυξη επιτυχημένων στατιστικών αλγορίθμων ΜΜ απαιτεί μεγάλους όγκους δεδομένων για τον έλεγχο και εμπλουτισμό τους, αλλά πρωτίστως ικανό μέγεθος επισημειωμένων δεδομένων για την αρχική εκπαίδευση των συγκεκριμένων συστημάτων.

Διαφορετικές μεθοδολογίες ΜΜ (Σχήμα 3.1) αναλύονται στις επόμενες ενότητες.

Μηχανική Μετάφραση βάσει λεξικού (Dictionary-based machine translation)

Η ΜΜ βάσει λεξικού αποτελεί "λέξη-προς-λέξη" μετάφραση, με χρήση βάσης λεξικού λημμάτων. Αυτή η μέθοδος αποτελεί την πρώτη γενιά ΜΜ (τέλη του '40 προς τα μέσα του '60) και είναι ιδανική για μεταφράσεις μακροσκελών καταλόγων προϊόντων και απλών εγχειριδίων προϊόντων ή υπηρεσιών. Τα συστήματα ΜΜ αυτής της κατηγορίας μπορούν να χρησιμοποιηθούν συνήθως ως εργαλείο υποβοήθησης της μετάφρασης από έναν επαγγελματία μεταφραστή, ο οποίος επιμελείται το συντακτικό και τη γραμματική του τελικού κειμένου. Η χρήση τους ως βασικό εργαλείο για τη μετάφραση κειμένων δεν ενδείκνυται, καθώς το μεταφράσμα που προκύπτει είναι αποτέλεσμα της απλής μεταφοράς λέξεων από τη μια γλώσσα στην άλλη, χωρίς την κατάλληλη γραμματική και



Σχήμα 3.1 Μεθοδολογίες Μηχανικής Μετάφρασης

συντακτική επεξεργασία που απαιτείται για τη μεταφορά του νοήματος στη ΓΣ. Στη συνέχεια αναπτύχθηκαν διάφορες παραλλαγές συστημάτων ΜΜ βάσει λεξικού, κάνοντας χρήση δίγλωσσων λεξικών μαζί με γραμματικούς κανόνες (Gerber and Yang, 1997), με αποτέλεσμα την ενσωμάτωση των εν λόγω συστημάτων στην επόμενη κατηγορία της ΜΜ βάσει κανόνων (Rule based machine translation - RBMT).

Μηχανική Μετάφραση βάσει κανόνων (Rule-based machine translation – RBMT)

Η ΜΜ βάσει κανόνων (RBMT) είναι ο γενικός όρος που περιγράφει τα συστήματα ΜΜ που βασίζονται στις γλωσσικές πληροφορίες μεταξύ της ΓΠ και της ΓΣ. Οι γλωσσικές πληροφορίες καλύπτουν τις βασικές σημασιολογικές, μορφολογικές και συντακτικές κανονικότητες της κάθε γλώσσας αντίστοιχα. Η μέθοδος που υιοθετούν τα εν λόγω συστήματα βασίζεται σε σύνολα γλωσσικών κανόνων που ορίζονται ως αντιστοιχίες μεταξύ των δομών της ΓΠ και της ΓΣ. Χάρη στη μεθοδολογία τους, τα συστήματα RBMT μπορούν να διαχειρίζονται μεγάλη ποικιλία γλωσσικών φαινομένων, είναι επεκτάσιμα και συντηρούνται εύκολα (Kaji, 1988). Ωστόσο, οι γραμματικές εξαιρέσεις αποτελούν σκόπελο στην ομαλή λειτουργία τους. Επίσης, η έρευνα για την ανάπτυξη ενός συστήματος RBMT απαιτεί πολύ χρόνο. Ο κύριος σκοπός ενός συστήματος RBMT είναι να μετατρέψει τη γλωσσική δομή (language structures) της ΓΠ στη δομή της ΓΣ.

Μηχανική Μετάφραση άμεσης προσέγγισης (Direct approach MT)

Οι λέξεις μεταφράζονται από τη ΓΠ χωρίς να διέλθουν μέσω μιας πρόσθετης/ενδιάμεσης αναπαράστασης. Το σύστημα RBMT "Anusaarka" είναι ένα σύστημα ΜΜ που βασίζεται στην άμεση προσέγγιση. Αναπτύχθηκε στο Ινδικό Ινστιτούτο Πληροφορικής, IIIT Hyderabad[1] και καλύπτει όλες τις βασικές Ινδικές γλώσσες (Josan and Lehal, 2008). Το σύστημα ΜΜ άμεσης προσέγγισης βασίζεται σε δίγλωσσα λεξικά που περιέχουν και γραμματικούς κανόνες.

Μηχανική Μετάφραση που βασίζεται στη μεταφορά (Transfer-based MT)

Η προσέγγιση αυτής της μεθοδολογίας ανήκει στη δεύτερη γενιά ΜΜ (από τα μέσα του '60 έως το 1980). Η διαδικασία της μεθοδολογίας αυτής περιλαμβάνει την ανάλυση του κειμένου-πηγή ως προς τη μορφολογία, το συντακτικό και τη σημασιολογία, δημιουργώντας μια όσο γίνεται αφηρημένη αναπαράσταση. Οι κανόνες "μεταφέρουν" ή "ταιριάζουν" αυτή την "αναπαράσταση" της ΓΠ σε μια νέα αναπαράσταση της ΓΣ. Αυτή η διαδικασία περιλαμβάνει τα εξής τρία βασικά στάδια (Tripathi and Sarkhel, 2010):

Ανάλυση: Η ανάλυση της ΓΠ γίνεται με βάση γλωσσικές πληροφορίες όπως η μορφολογία, τα μέρη του λόγου, η σύνταξη, η σημασιολογία κλπ. Επιπλέον, εφαρμόζονται ευρετικοί αλγόριθμοι για να αναλύσουν τη ΓΠ και να αντλήσουν:

- τη συντακτική δομή (syntactic structure) της ΓΣ ή
- τη σημασιολογική δομή (semantic structure) της ΓΣ, εφόσον πρόκειται για γλωσσικά ζεύγη διαφορετικών οικογενειών.

Μεταφορά: Η συντακτική/σημασιολογική δομή της ΓΠ "μεταφέρεται" στη συντακτική/σημασιολογική δομή της ΓΣ.

Σύνθεση: Σε αυτό το στάδιο η μηχανή αντικαθιστά τα συστατικά της ΓΠ με τα αντίστοιχα της ΓΣ. Ωστόσο, η προσέγγιση αυτή εξαρτάται από το ζεύγος γλωσσών που εμπλέκονται στη διαδικασία της μετάφρασης. Για τον λόγο αυτό, το ερευνητικό έργο "Eurotra project4" (Tripathi and Sarkhel, 2010) προτείνει τη χρήση δύο ανεξάρτητων μονόγλωσσων λεξικών. Επίσης, υπάρχουν διαφορετικές αναπαραστάσεις για διαφορετικές γλώσσες. Το σύστημα PaTrans (Σύστημα Μηχανικής Μετάφρασης για διπλώματα ευρεσιτεχνίας) βασίζεται σε Transfer-based MT και είναι ένα από τα αποτελέσματα του ερευνητικού έργου

[1] https://www.iiit.ac.in/

“Eurotra project4”. Το σύστημα Mantra είναι επίσης ένα πρότυπο μοντέλο ΜΜ για ινδικές γλώσσες που βασίζεται στη προσέγγιση της μεταφοράς κανόνων. Η δημιουργία του συστήματος χρηματοδοτήθηκε από την κυβέρνηση της Ινδίας και ο αναλυτής χρησιμοποιείται για την επεξεργασία της γλώσσας που είναι γνωστή ως Vyakarta.

Διαγλωσσική Μηχανική Μετάφραση (InterLingua Machine Traslation) Η διαγλωσσική ΜΜ θεωρείται ότι ανήκει στην τρίτη γενιά της ΜΜ. Αποτελεί αναπόσπαστο μέρος ενός κλάδου που ονομάζεται διαγλωσσολογία (Interlinguistics). Η διαγλωσσική ΜΜ στοχεύει στη δημιουργία γλωσσικής ομοιογένειας σε παγκόσμιο επίπεδο. Η λέξη “Interlingua” (διαγλωσσική) προέρχεται από τον συνδυασμό των δύο λατινικών λέξεων “Inter” και “Lingua”, που σημαίνουν “ενδιάμεσος” και “γλώσσα” αντίστοιχα.

Στη διαγλωσσική ΜΜ (InterLingua), η ΓΠ μετατρέπεται σε μία βοηθητική/ενδιάμεση γλώσσα (αναπαράσταση), η οποία είναι ανεξάρτητη από οποιαδήποτε από τις γλώσσες που εμπλέκονται στη μετάφραση.

Στη συνέχεια, η μετάφραση του κειμένου στη ΓΣ προκύπτει μέσω αυτής της ενδιάμεσης αναπαράστασης. Σε αυτό το είδος ΜΜ, χρειάζονται μόνο δύο μονάδες, η ανάλυση και η σύνθεση. Επίσης, λόγω της ανεξαρτησίας του από το ζεύγος γλωσσών, το σύστημα αυτό παρουσιάζει μεγάλο ενδιαφέρον στην πολύγλωσση μηχανική μετάφραση. Σύμφωνα με αυτή την προσέγγιση, πολλές γλώσσες μπορούν να υποβληθούν σε μια ενιαία μορφή ανάλυσης (interlingua), έπειτα από την οποία μπορεί να συσταθεί το μετάφρασμα σε οποιαδήποτε ΓΣ, μέσω της διαδικασίας της σύνθεσης. Η μέθοδος αυτή βρίσκει ιδανικό πεδίο εφαρμογής σε πολυγλωσσικά συστήματα ΜΜ και σε πολύ συγκεκριμένους τομείς χρήσης της γλώσσας. Το σύστημα UNITRAN6 είναι μια εφαρμογή διαγλωσσικής ΜΜ. Χρησιμοποιεί παραμετροποιήσεις τόσο σε συντακτικό όσο και λεξιλογικό επίπεδο. Το Ινδικό Ινστιτούτο Τεχνολογίας, Powai[1] εργάζεται πάνω στην ανάπτυξη συστημάτων ΜΜ για τις Ινδικές γλώσσες που βασίζονται στη μεθοδολογία της διαγλωσσικής ΜΜ.

Μηχανική Μετάφραση που βασίζεται στη γνώση (Knowledge-based machine translation - KBMT)

Αυτό το είδος συστήματος ΜΜ είναι βασισμένο στη γενική “αντίληψη” του λεξικού που αντιπροσωπεύει μια περιοχή. Το σύστημα “KANT7” Nyberg and Mitamura (1992) είναι ένα παράδειγμα KBMT για πολύγλωσση μετάφραση, που αναπτύχθηκε μέσα από μια μεγάλη κλίμακα βάσης γνώσεων και ένα ελεγχόμενο σύστημα γλώσσας.

[1] http://www.cfilt.iitb.ac.in/machine-translation/

Μηχανική Μετάφραση που βασίζεται σε σώματα κειμένων (Corpus-based MT)

Από το 1989, η προσέγγιση της ΜΜ που βασίζεται σε σώματα κειμένων (Corpus-based MT) έχει αναδειχθεί ως ένα από τα πιο ευρέως διερευνώμενα πεδία στον τομέα της ΜΜ. Η μέθοδος αυτή έχει κυριαρχήσει έναντι άλλων προσεγγίσεων, χάρη στο υψηλό επίπεδο ακρίβειας της μετάφρασης που επιτυγχάνει. Μερικές από τις προσεγγίσεις ΜΜ που βασίζονται σε σώματα κειμένων παρουσιάζονται παρακάτω:

Στατιστική Μηχανική Μετάφραση (Statistical machine translation - SMT)

Η Στατιστική ΜΜ (SMT) είναι ένα παράδειγμα μηχανικής μετάφρασης, όπου τα μεταφράσματα δημιουργούνται με βάση στατιστικά μοντέλα των οποίων οι παράμετροι προκύπτουν κυρίως από τη στατιστική ανάλυση δίγλωσσων σωμάτων κειμένων. Ο Warren Weaver, το 1949 (Shannon and Weaver, 1963), είχε εισαγάγει την ιδέα της SMT. Οι πρώτες μηχανές μετάφρασης που χρησιμοποιούσαν τη μέθοδο στατιστικής ανάλυσης παρουσιάστηκαν από τους ερευνητές του ερευνητικού κέντρου της IBM, στις αρχές της δεκαετίας του 1990 (Brown et al., 1993). Στο πλαίσιο αυτό, οι στατιστικές μέθοδοι που εφαρμόζονται για την παραγωγή μετάφρασης χρησιμοποιούν δίγλωσσα σώματα κειμένων. Συναντούμε, για παράδειγμα, συστήματα SMT με βάση τα n-γράμματα (n-grams) (Schwenk et al., 2007), με βάση τον αριθμό εμφανίσεων (Chiang, 2010) κλπ. Ο Macherey (Macherey et al., 2008) είχε πειραματιστεί με διάφορες στατιστικές μεθόδους για την κατανόηση του προφορικού λόγου για συστήματα SMT.

Στατιστικά μοντέλα που βασίζονται στη λέξη: Τα συστήματα στατιστικής μετάφρασης που βασίζονται σε λέξεις έχουν τα παρακάτω χαρακτηριστικά:

- Θεμελιώδης μονάδα είναι η λέξη (Fundamental unit – Word)
- Αναδιάταξη (Reordering): Οι αλγόριθμοι που σχετίζονται με την ευθυγράμμιση των λέξεων, η οποία απαιτείται για να επιτευχθεί μεγαλύτερη ακρίβεια στη μεταφρασμένη φράση.
- Σύνθετες λέξεις, ιδιωματισμοί και ομώνυμα δημιουργούν αυξημένη πολυπλοκότητα για απλή μετάφραση με βάση τη λέξη.

Στατιστικά μοντέλα που βασίζονται στη φράση: Τα συστήματα στατιστικής μετάφρασης που βασίζονται σε φράσεις έχουν τα παρακάτω χαρακτηριστικά:

- Θεμελιώδης μονάδα είναι η φράση ή σειρά λέξεων (Fundamental unit – a phrase or sequence of words)

- Αναπτύσσονται ακολουθίες λέξεων από τη ΓΠ και τη ΓΣ. Η αποκωδικοποίηση γίνεται με βάση το διάνυσμα των χαρακτηριστικών που ταιριάζουν με τις τιμές της γλώσσας ζεύγους (του μεταφραστικού ζεύγους της γλώσσας).

Σε ένα σύστημα ΜΜ που βασίζεται σε φράσεις (Example-based machine translation (EBMT)) έχουμε ένα σύνολο από προτάσεις στη ΓΠ και τις αντίστοιχες μεταφράσεις κάθε πρότασης στη ΓΣ, οι οποίες χαρτογραφούνται σημείο προς σημείο. Αυτά τα παραδείγματα χρησιμοποιούνται για τη μετάφραση παρόμοιων φράσεων. Η βασική αρχή είναι ότι, στην περίπτωση που μια φράση έχει μεταφραστεί και εμφανίζεται πάλι, τότε η ίδια μετάφραση είναι πιθανό να είναι και πάλι σωστή (Cavalli-Sforza et al., 2004).

Τα πλεονεκτήματα ενός συστήματος EBMT σε σχέση με ένα σύστημα SMT, όπως διατυπώνονται από τον Frederking στο βιβλίο του[1] είναι τα εξής:

- Το σύστημα EBMT μπορεί να λειτουργήσει με μικρά σετ δεδομένων.
- Η εκπαίδευση της ΜΜ και η αποκωδικοποίηση γίνεται πιο γρήγορα.
- Απαιτούνται λιγότερες τυποποιήσεις (Less principled).

Εντούτοις, μερικές μελέτες επιβεβαιώνουν τα συστήματα SMT ως ένα από ένα από τα παραδείγματα του EBMT (Dale et al., 2000).

Στατιστικά μοντέλα που βασίζονται στη σύνταξη: Η στατιστική μετάφραση βάσει κανόνων αποτελείται από μια ακολουθία λέξεων και μεταβλητών της ΓΠ, ένα συντακτικό δέντρο στη ΓΣ (λέξεις ή μεταβλητές σε μορφή φύλλων δέντρου) και ένα διάνυσμα τιμών με χαρακτηριστικά που περιγράφουν την πιθανοφάνεια (likehood) (DeNeefe et al., 2010, Yamada and Knight, 2001) του μεταφραστικού ζεύγους της γλώσσας.

Μηχανική Μετάφραση που βασίζεται σε παραδείγματα (Example-based machine translation - EBMT)

Η ΜΜ που βασίζεται σε παραδείγματα (EBMT), επίσης γνωστή και ως ΜΜ μνήμης, βασίζεται στην ανάκληση/εύρεση ανάλογων παραδειγμάτων από δίγλωσσα παράλληλα κείμενα με ευθυγραμμισμένα ζεύγη προτάσεων.

Σε αυτή την περίπτωση χρησιμοποιείται η "αρχή της αναλογίας", μια ιδέα που παρουσιάστηκε για πρώτη φορά από τον Makoto Nagao το 1981 (Hutchins, 2007). Επίσης, χρησιμοποιείται και η ιδέα του Case Based Reasoning (CBR). Τα συστήματα CBR χρησιμοποιούνται εκτενώς σε εφαρμογές εξυπηρέτησης πελατών (help desk), όπως για παράδειγμα το Compaq SMART system (Nguyen et al., 1993).

[1] http://www.cs.cmu.edu/afs/cs/project/cmt-55/lti/Courses/731/www/ebmt2007.pdf

Μηχανική Μετάφραση που βασίζεται στο περιεχόμενο (Context-based machine translation - CBMT)

Η ΜΜ που βασίζεται στο περιεχόμενο (CBMT) αναπτύσσεται ως σύστημα ΜΜ που βασίζεται σε σώματα κειμένων, αλλά δεν απαιτεί ούτε κανόνες ούτε παράλληλα σώματα κειμένων. Αντ' αυτών, η CBMT για να λειτουργήσει απαιτεί ένα εκτεταμένο σώμα μονόγλωσσων κειμένων της ΓΣ, ένα πλήρες δίγλωσσο λεξικό και προαιρετικά (για περαιτέρω βελτίωση της ποιότητας της μετάφρασης) ένα μικρότερο σώμα μονόγλωσσων κειμένων της ΓΠ.

Τα πλεονεκτήματα ενός συστήματος CBMT είναι τα εξής:

- Ακριβή σώματα κειμένων ΜΜ που εκπαιδεύονται από τα μονόγλωσσα κείμενα. Αυτό σημαίνει ότι η CBMT είναι εφαρμόσιμη σχεδόν σε οποιοδήποτε ζεύγος γλωσσών.
- Διατηρεί το περιεχόμενο μέσα στο μεταφραστικό πλαίσιο.
- Δυνατότητα να χειριστεί μεγαλύτερες συμβολοσειρές σε σύγκριση με άλλα συστήματα.
- Μπορεί να χειριστεί τα φαινόμενα της λεκτικής ασάφειας.
- Έχει τη δυνατότητα να προτείνει εναλλακτικές φράσεις (συνώνυμα φράσεων), σε περίπτωση που δεν βρεθεί κατάλληλη αντιστοιχία στη ΓΣ.
- Μπορεί να διαχωρίζει τμήματα της μετάφρασης ανάλογα με τον βαθμό αξιοπιστίας (υψηλή ή χαμηλή).

Το τελευταίο συντελεί στην εξοικονόμηση χρόνου και χρήματος, αφού μετά την επεξεργασία η προσοχή του χρήστη επικεντρώνεται μόνο στα τμήματα με χαμηλό βαθμό αξιοπιστίας.

Τα συστήματα CONTRAST (Isahara and Uchida, 1995) και REFTEX (Kjærsgaard, 1987) αποτελούν παραδείγματα CBMT.

Υβριδικά μοντέλα Μηχανικής Μετάφρασης (Hybrid machine translation - HMT)

Τα υβριδικά μοντέλα ΜΜ αξιοποιούν τα πλεονεκτήματα της στατιστικής προσέγγισης σε συνδυασμό με τη χρήση των γλωσσικών κανόνων (Boretz and Adam, 2009). Τα συστήματα αυτά αποτελούν και την πιο ελπιδοφόρα εκδοχή των συστημάτων ΜΜ προς την κατεύθυνση της ποιοτικής μηχανικής μετάφρασης ευρείας γκάμας κειμένων.

Μηχανική Μετάφραση που βασίζεται στο Βαθύ Νευρωνικό Δίκτυο (Deep Neural Network machine translation)

Τα συστήματα ΜΜ που βασίζονται στο Βαθύ Νευρωνικό Δίκτυο χρησιμοποιούν δίκτυο διασυνδεμένων νευρώνων, μέσω αλγορίθμων τεχνητής νοημοσύνης (Collobert and Weston, 2008, Kombrink et al., 2011, Mikolov et al., 2013). Το βαθύ νευρωνικό δίκτυο (deep neural network-DNN) είναι ένα έξυπνο νευρωνικό δίκτυο (artificial neural network-ANN) με πολλαπλά κρυφά επίπεδα (multiple hidden layers) ανάμεσα στα επίπεδα εισόδου και εξόδου (Deng et al., 2014, Schmidhuber, 2015). Γενικώς, υπάρχει μια τάση στην ερευνητική κοινότητα προς τη χρήση έξυπνων συστημάτων νευρωνικών δικτύων και τεχνητής νοημοσύνης.

3.2 Αποτύπωση της κατάστασης της ΜΜ ως προς τις Νοηματικές Γλώσσες

Σύμφωνα με τους Porta et al. (2014) σχετικά με τα θεμελιώδη προβλήματα των ΝΓ, τα περισσότερα νέα έργα της ΝΓ υιοθετούν τις γλωσσικές θεωρίες που αναπτύχθηκαν για τις προφορικές γλώσσες, αντί να δοκιμάσουν νέες θεωρίες. Από την άποψη της υπολογιστικής επεξεργασίας των γλωσσών, οι ΝΓ εξακολουθούν να μην διαθέτουν επαρκείς γλωσσικούς πόρους. Επίσης ελάχιστες τεχνολογίες είναι διαθέσιμες και χρησιμοποιούνται για τις γλώσσες αυτές.

Επιπλέον, ένα άλλο σημαντικό πρόβλημα των ΝΓ είναι η έλλειψη ενός συστήματος γραφής. Συνήθως, εφαρμόζεται η αναλογική (παλιότερα) ή ψηφιακή (σήμερα) βιντεοσκόπηση της ΝΓ ως καθιερωμένος τρόπος για να αναπαρασταθεί η ΝΓ και να δημιουργηθούν βιντεοσώματα. Για τον λόγο αυτό, υπάρχει μεγάλη έλλειψη σωμάτων της ΝΓ ειδικά σε γραπτή μορφή. Οι περιορισμοί στη σύνταξη, επεξεργασία και επαναχρησιμοποίηση των εκφράσεων (utterances) της ΝΓ, καθώς και οι συνέπειές τους για την εκπαίδευση και την επικοινωνία κωφών έχουν αναφερθεί συστηματικά στη βιβλιογραφία των ΝΓ, από το δεύτερο μισό του 20ου αιώνα (Efthimiou et al., 2016). Ωστόσο, διάφορες επιστημονικές κοινότητες έχουν αναπτύξει και χρησιμοποιούν διάφορα συστήματα αναπαράστασης της ΝΓ, κυρίως για ερευνητικούς σκοπούς, παρά ως επίσημα συστήματα γραφής της ΝΓ.

Τα σημαντικότερα από αυτά είναι το σύστημα του Stokoe Jr (2005), του SignWriting (Sutton, 1995), του HamNoSys Prillwitz et al. (1989) και του Neidle (Neidle et al., 2001). Το σύστημα SignWriting σχεδιάστηκε πρωτίστως ως σύστημα γραφής και έχει τις ρίζες του στο DanceWriting (Sutton, 1973), μια σημειογραφία για την ανάγνωση και τη συγγραφή κινήσεων χορού. Το σύστημα HamNoSys σχεδιάστηκε ως ένα σύστημα φωνολογικής

μεταγραφής για τη ΝΓ, χρησιμοποιώντας την ίδια φιλοσοφία με το Διεθνές Φωνητικό Αλφάβητο (IPA) για τις προφορικές γλώσσες. Ένα πολλά υποσχόμενο σύστημα είναι το SiGML (Elliott et al., 2004), που αντιπροσωπεύει τις 3D ιδιότητες των ΝΓ. Τέλος, το σύστημα γραφής "si5s" system (Augustus et al., 2013) αποτελεί μια αξιόλογη προσπάθεια και πρόταση για την αμερικανική νοηματική γλώσσα (ASL).

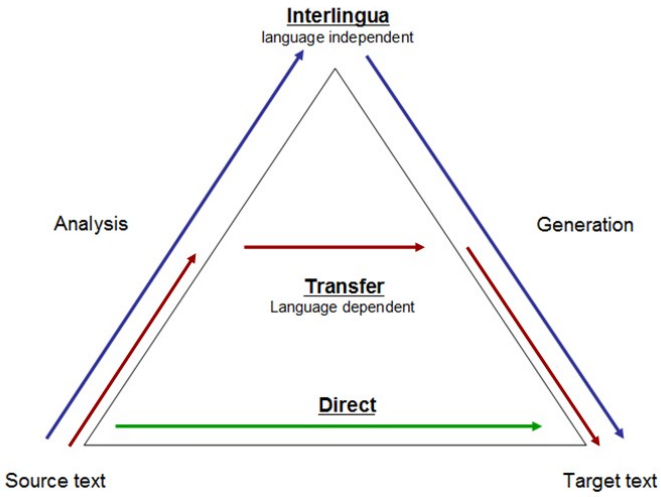
Όσον αφορά την ΕΝΓ, προς το παρόν δεν υπάρχουν γλωσσικά μοντέλα λόγω έλλειψης μεγάλων σωμάτων της ΕΝΓ. Προκειμένου να αντιμετωπιστεί το σχετικό κενό, στην παρούσα διατριβή προτείνεται ένα καινοτόμο σύστημα μετάφρασης με κανόνες (RBMT), το οποίο είναι σε θέση να παράγει γρήγορα και σε υψηλή ποιότητα, μεγάλα παράλληλα σώματα ΕΝΓ σε μορφή Σύντομης Μεταγραφής της Ελληνικής Νοηματικής Γλώσσας (ΣΜΕΝΓ).

3.2.1 Επισκόπηση βιβλιογραφίας σχετικά με τα συστήματα μηχανικής μετάφρασης Νοηματικής Γλώσσας

Υπόβαθρο

Η μηχανική μετάφραση των προφορικών γλωσσών έχει τις ρίζες της στη δεκαετία του '40, με σημαντική αύξηση του ενδιαφέροντος για το πεδίο αυτό στα τέλη της δεκαετίας του '70 και του '80 (Trujillo, 1999). Παρόλα αυτά, δεν συναντούμε παρόμοιο επίπεδο ενδιαφέροντος για τη μηχανική μετάφραση της ΝΓ. Μέχρι τη δεκαετία του '90, οπότε και εμφανίστηκε μια εργασία για τη γλωσσική ανάλυση των ΝΓ από τον Morrissey (2008), δεν έχουμε κάποια εκτεταμένη έρευνα. Παρά την καθυστέρηση αυτή, η ανάπτυξη των συστημάτων μηχανικής μετάφρασης των ΝΓ, σε γενικές γραμμές, ακολούθησε τις θεωρητικές προσεγγίσεις των συστημάτων μηχανικής μετάφρασης των προφορικών γλωσσών. Συγκεκριμένα δοκιμάστηκαν συστήματα δεύτερης γενιάς, βασισμένα σε κανόνες (rule-based), τα οποία εμφανίστηκαν κατά τη δεκαετία του '70, με την ανάπτυξη συστημάτων όπως το Meteo (Chandioux, 1976, 1989) και το Systran (Toma, 1977). Αυτά αποτελούν παραδείγματα των πρώτων εμπορικά υιοθετημένων συστημάτων ΜΜ για επιτυχή μετάφραση των προφορικών γλωσσών.

Τα συστήματα που βασίζονται σε κανόνες μπορούν να υποδιαιρεθούν σε συστήματα μεταφοράς κανόνων ή σε διαγλωσσικά συστήματα. Η πυραμίδα Vauquois (Σχήμα 3.2) (Hutchins and Somers, 1992) χρησιμοποιείται ευρέως σε κύκλους ΜΜ για να απεικονίσει την προσπάθεια που σχετίζεται με τις διαδικασίες μετάφρασης. Στη βάση της πυραμίδας συναντούμε τη ΜΜ άμεσης προσέγγισης (Direct) που αφορά κυρίως τη μετάφραση λέξη προς λέξη και βρίσκει εφαρμογή περισσότερο σε τυποποιημένες μεταφράσεις τεχνικών εγχειριδίων και οδηγιών χρήσης συσκευών. Στη συνέχεια συναντούμε τη ΜΜ που



Σχήμα 3.2 Η πυραμίδα Vauquois

βασίζεται στη μεταφορά (Transfer), κατά την οποία τα συστήματα πρέπει να γνωρίζουν τη ΓΠ και τη ΓΣ. Τέλος, στην κορυφή της πυραμίδας συναντούμε τη διαγλωσσική ΜΜ (Interlingua), όπου τα συστήματα τείνουν να αναλύουν βαθύτερα τη φράση της ΓΠ, δημιουργώντας περισσότερες δομές σημασιολογικής φύσης. Όλες οι μέθοδοι έχουν πλεονεκτήματα και μειονεκτήματα.

Ο πολλαπλασιασμός των εμπειρικών τεχνολογιών στην έρευνα των γλωσσικών τεχνολογιών της πληροφορικής, σε συνδυασμό με την αποτυχία των μεθόδων που βασίζονται σε κανόνες για τη δημιουργία αξιόπιστων, επεκτάσιμων και ευρείας κάλυψης συστημάτων ΜΜ, άνοιξε τον δρόμο για τις εμπειρικές προσεγγίσεις ΜΜ (Morrissey, 2008). Τα συστήματα ΜΜ που βασίζονται σε δεδομένα άρχισαν να κερδίζουν έδαφος κατά τη δεκαετία του ΄90 και τώρα πλέον κυριαρχούν στον τομέα της έρευνας. Αυτά τα συστήματα, που συχνά αποκαλούνται ως "συστήματα ΜΜ που βασίζονται σε σώματα κειμένων (corpora)", μπορούν να υποδιαιρεθούν σε συστήματα Στατιστικής ΜΜ (ΣΜΜ) που βασίζονται σε λέξεις, σε φράσεις, στη σύνταξη ή σε παραδείγματα (Tripathi and Sarkhel, 2010). Σε σύγκριση με τα συστήματα ΜΜ που βασίζονται σε κανόνες, τα συστήματα ΜΜ που βασίζονται σε δεδομένα έχουν θεμελιώδεις διαφορές τόσο ως προς τις διεργασίες όσο και στις ίδιες τις παραμέτρους. Γενικά, οι γλωσσικές πληροφορίες και οι κανόνες αποφεύγονται έναντι πιθανοτικών μοντέλων που συλλέγονται από ένα μεγάλο παράλληλο σώμα κειμένων.

Οι στατιστικές μέθοδοι προέρχονται σε μεγάλο βαθμό από την περιοχή της αναγνώρισης ομιλίας (Brown et al., 1988, 1991). Ουσιαστικά, η μετάφραση μιας πρότασης θεωρείται ως η πιθανότητα της πιο πιθανής συμβολοσειράς στόχου e, δεδομένης της συμβολο-

σειράς πηγής f: $Pr(e|f)$. Το log-γραμμικό μοντέλο (Koehn, 2004) υπολογίζει την $Pr(e|f)$ απευθείας από το παράλληλο σώμα κειμένων, χρησιμοποιώντας τον ακόλουθο τύπο:

$$argmax_e Pr(e|f) = e^{\Sigma_i \lambda_i h_i(e,f)} \tag{3.1}$$

Για τα συστήματα ΜΜ που βασίζονται σε δεδομένα πρέπει να υπάρχουν διαθέσιμα μεγάλα σώματα δίγλωσσων δεδομένων, από τα οποία να εξάγονται πιθανότητες ή παραδείγματα. Στην περίπτωση της μετάφρασης των ομιλούμενων γλωσσών, μεγάλες ποσότητες δεδομένων καθίστανται ολοένα και πιο διαθέσιμες. Από την άλλη πλευρά, η συλλογή δεδομένων παραμένει σοβαρό πρόβλημα για τα συστήματα ΜΜ στη ΝΓ, ειδικά στην περίπτωση της ΕΝΓ.

Σύστηματα Μηχανικής Μετάφρασης με κανόνες (Rule-based SL MT Systems)

Όλα τα συστήματα ΜΜ για τις ΝΓ που δημοσιεύτηκαν μέχρι το 2003 ήταν μόνο έργα σε εξέλιξη ή απλές δοκιμαστικές εκδόσεις (Huenerfauth, 2003). Ωστόσο, ορισμένα από αυτά τα συστήματα διακρίθηκαν, όπως το σύστημα ZARDOZ (Veale and Conway, 1994), ο μεταφραστής ViSiCAST (Bangham et al., 2000), το έργο ASL Workbench (Speers, 2002), η μετάφραση ΝΓ μέσω DRT και HPSG του Sáfár and Marshall (2002) και το έργο TEAM (Zhao et al., 2000). Όλα αυτά τα συστήματα βασίζονταν σε κανόνες και χρησιμοποίησαν προσεγγίσεις είτε διαγλωσσικές (Interlingua) είτε βασιζόμενες σε μεταφορές (Transfer-based). Η μοναδική προσέγγιση που ασχολείται με τα ευρήματα ταξινομητή (classifier predicates) ήταν αυτή των Huenerfauth et al. (2006), που πρότεινε μια μέθοδο πολλαπλών διαδρομών που συνδυάζει τη διαγλωσσική προσέγγιση (Interlingua), την προσέγγιση μεταφοράς γλώσσας (Transfer-based) και την άμεση προσέγγιση (Direct approach).

Για την ισπανική γλώσσα προς την Ισπανική Νοηματική Γλώσσα (ΙΝΓ), οι Baldassarri and Royo-Santas (2009) περιέγραψαν ένα δοκιμαστικό σύστημα ΜΜ που βασίζεται σε κανόνες. Τα ισπανικά αναλύονται σε μέρη του λόγου, χρησιμοποιώντας έναν αναλυτή βασισμένο σε ανάλυση εξάρτησης (dependency analysis) (Atserias et al., 2006). Η ανάλυση εξάρτησης μέσω γραμματικών κανόνων μετατρέπεται σε σειρά glosses (σύστημα σύντομης γραπτής αναπαράστασης). Το σύστημα δοκιμάστηκε με 92 προτάσεις που περιείχαν συνολικά 561 λέξεις. Για την αξιολόγηση δημιουργήθηκαν κατάλληλες καταχωρίσεις λεξικού με πολύ ικανοποιητικά αποτελέσματα: το 96% των λέξεων μεταφράστηκε σωστά και το 93,7% ήταν σε σωστή σειρά. Ένα άλλο ενδιαφέρον Ισπανικό σύστημα ΜΜ είναι το Apertium (Forcada et al., 2011), το σύστημα που βασίζεται στην εφαρμογή κανόνων της ισπανικής γλώσσας στην ΙΝΓ και στην πλατφόρμα ανοιχτού λογισμικού. Δεν υπάρχουν δημοσιευμένα αποτελέσματα για αυτό το σύστημα, αλλά είναι διαθέσιμο στο διαδίκτυο.

Ο Grieve-Smith (1999) περιέγραψε ένα σύστημα βασισμένο σε μεταφορές για τη μετάφραση της αγγλικής γλώσσας στην Αμερικάνικη Νοηματική Γλώσσα (ASL), στο πεδίο της πρόβλεψης του καιρού. Για το έργο αυτό, ο Grieve-Smith υιοθέτησε ένα σύστημα σύντομων γραπτών μεταγραφών της ΝΓ με βάση το ρωμαϊκό αλφάβητο, που επιτρέπει την εύκολη ενσωμάτωση με τα συστήματα Unix που βασίζονται σε ASCII. Παρόλα αυτά δεν τεκμηριώνονται πειράματα, ενώ ο Grieve-Smith δηλώνει ότι σκοπός του έργου ήταν να αποδείξει ότι μια τέτοια προσέγγιση ήταν εφικτή.

Όσον αφορά την ΕΝΓ, οι Kouremenos et al. (2010) παρουσίασαν ένα πρωτότυπο σύστημα μετατροπής του ελληνικού κειμένου στην ΕΝΓ. Σε αυτό το σύγγραμμα παρέχεται η λεπτομερής εφαρμογή της συνιστώσας επεξεργασίας γλώσσας, εστιάζοντας στα εγγενή προβλήματα της άντλησης γνώσης της γραμματικής της ΝΓ και της εφαρμογής της μέσα σε ένα πλαίσιο ανάλυσης. Πρόσφατα η Efthimiou et al. (2016) παρουσίασε την υλοποίηση ενός σταδίου μετά-επεξεργασίας (post-processing) σε ένα σύστημα ΜΜ βασισμένο σε γραμματικούς κανόνες από το γραπτό ελληνικό κείμενο στην ΕΝΓ, όπου η μετάφραση εξάγεται και αποδίδεται από ένα εικονικό avatar. Η μετά-επεξεργασία εφαρμόζεται στην έξοδο της μονάδας μεταφοράς του συστήματος ΜΜ, ενεργοποιώντας ένα γραφικό περιβάλλον επεξεργασίας για την ΕΝΓ, το οποίο χρησιμοποιεί ένα λεξικό νοημάτων που κωδικοποιούνται για μεταγραφές HamNoSys και μη χειροκίνητα νοήματα, μαζί με γνώσεις γραμματικής. Αυτό το περιβάλλον μετά-επεξεργασίας μπορεί, επίσης, να χρησιμοποιηθεί ως αυτόνομο περιβάλλον επεξεργασίας για έμπειρους χρήστες, επιτρέποντας ένα ευρύ φάσμα ενεργειών επεξεργασίας, καθώς και ως ένα εργαλείο επεξεργασίας ΝΓ για το μη εξοικειωμένο κοινό.

Συστήματα Μηχανικής Μετάφρασης με δεδομένα (Data-Driven Based SL MT Systems)

Τον τελευταίο καιρό, η ΜΜ που βασίζεται σε παραδείγματα (EBMT), η στατιστική ΜΜ (SMT) και άλλοι τύποι συστημάτων μετάφρασης με βάση τα δεδομένα, αντικατέστησαν τις προηγούμενες προσεγγίσεις ΜΜ με κανόνες (RBMT). Ωστόσο, οι προσεγγίσεις που βασίζονται σε δεδομένα υπολογίζουν τις παραμέτρους τους από ένα ευθυγραμμισμένο δίγλωσσο σώμα κειμένων και η ακρίβειά τους εξαρτάται σε μεγάλο βαθμό από την ποιότητα και το μέγεθος αυτού του σώματος. Δυστυχώς, τα σώματα για τις ΝΓ είναι ακόμα πολύ μακριά από το να προσεγγίζουν την αιχμή της τεχνολογίας των παραπάνω συστημάτων για τις προφορικές γλώσσες. Επιπλέον, το πρόβλημα της τροπικότητας (modality) και η έλλειψη ενός τυποποιημένου συστήματος γραφής καθιστούν την απόκτηση δεδομένων για τις ΝΓ μια χρονοβόρα και δαπανηρή εργασία. Τα δεδομένα ΝΓ συνήθως αποκτώνται με βιντεοσκοπήσεις ενός ηθοποιού της ΝΓ. Στη συνέχεια γίνεται σχολιασμός και χαρακτηρισμός των βίντεο με πολυκαναλικούς σχολιασμούς που

περιγράφουν με στιλπνότητα και σύντομες γραπτές μεταγραφές (glosses) και τα μη χειροκίνητα νοήματα. Η χρήση εργαλείων σχολιασμού όπως το ANVIL (Kipp, 2001), το SignStream (Neidle et al., 2001), το iLex (Hanke, 2002), το ELAN (Brugman et al., 2002), ή τα εργαλεία σχολιασμού βίντεο που παρουσιάζονται στη μελέτη (Braffort et al., 2004), απλοποιούν τις διαδικασίες κατασκευής και σχολιασμού νέων σωμάτων για τις ΝΓ. Συνολικά, εάν τα σώματα είναι απαραίτητα για την έρευνα της ΝΓ, τα παράλληλα σώματα είναι απαραίτητα για τη ΜΜ. Το πλήθος των παράλληλων δεδομένων που είναι σήμερα διαθέσιμα για ΓΠ είναι συνήθως μερικές εκατοντάδες προτάσεις ή, ακόμα χειρότερα, και μεμονωμένες λέξεις σε περιορισμένους τομείς και σε πολύ λίγους θεματικούς τομείς. Ωστόσο, ορισμένα παράλληλα σώματα χρησιμοποιούνται για έρευνα στη μηχανική μετάφραση. Αυτά συνοψίζονται στον Πίνακα 3.1 (Porta et al., 2014).

Παρά την έλλειψη παράλληλων σωμάτων, η επιτυχία των προσεγγίσεων ΜΜ που βασίζονται σε δεδομένα μεταξύ των προφορικών γλωσσών οδήγησε στην εφαρμογή των ίδιων τεχνικών και στις ΝΓ. Ωστόσο, σύμφωνα με τον Morrissey (Morrissey, 2011), οι περισσότερες έρευνες στη ΜΜ των ΝΓ προέρχονται από σποραδικά και βραχυπρόθεσμα έργα, σε αντίθεση με τις μακροπρόθεσμες επενδύσεις στην έρευνα της ΜΜ των προφορικών γλωσσών. Αξίζει να αναφερθούν ορισμένα έργα ακόμη. Το σύστημα ΜΜ της ταϊλανδικής γλώσσας προς την Ταϊλανδική Νοηματική Γλώσσα (Morrissey, 2011) παρουσιάζει ένα σύστημα άμεσης μετάφρασης με κανόνες αναδιοργάνωσης. Το σύστημα ΜΜ για την ταϊλανδική γλώσσα φθάνει ένα F-score περίπου 97% για ένα σύνολο 297 δοκιμαστικών προτάσεων. Στη μελέτη του Bauer et al. (1999) παρουσιάζεται η πρώτη στατιστική προσέγγιση της ΜΜ για την Γερμανική Νοηματική Γλώσσα (DGS). Στη μελέτη αυτή αναφέρεται ότι για 52 νοήματα επιτυγχάνεται ακρίβεια αναγνώρισης 94% και για 100 νοήματα 91,6%. Ο Morrissey (Morrissey, 2008) παρουσίασε λεπτομερή πειράματα του MaTrEx, μια υβριδική προσέγγιση που συνδυάζει EBMT και SMT (Stroppa and Way, 2006).

Τα αποτελέσματα του MaTrEx στο corpus ATIS έφθασαν το 0,39 BLEU για μετάφραση από την αγγλική γλώσσα στην Ιρλανδική Νοηματική Γλώσσα και περίπου 50% για μετάφραση από τη γερμανική γλώσσα στη Γερμανική Νοηματική Γλώσσα (DGS). Πρόσφατα, οι Morrissey και Way Way (Morrissey and Way, 2013) εκμεταλλεύτηκαν την αμφίδρομη λειτουργία του συστήματος MaTrEx, επιδεικνύοντας τον τρόπο με τον οποίο πρόσθετες ενότητες (modules), όπως η αναγνώριση και η SL animation, μπορούν να δημιουργήσουν ένα πλήρες μοντέλο ΜΜ ΓΠ για επικοινωνία μέσω ομιλίας και ΓΠ (Stein et. al. (Stein et al., 2006)). Εργάστηκαν στη Γερμανική Νοηματική Γλώσσα (DGS). Ο βασικός μηχανισμός ΜΜ που χρησιμοποιείται για τη μετάφραση ΓΠ είναι το σύστημα SMT βασισμένο σε φράσεις, που αναπτύχθηκε στο RWTH (Matusov et al., 2006). Το μοντέ-

Corpus	Languages pair	Sents	Tokens		Types		Reference
			Words	Signs	Words	Signs	
–	Italian - Italian SL	585	15000	6000	1442	300	Nicola et al. (2010)
RWTH-Phoenix	German - German SL	2468	16500	10500	1302	1895	Bungeroth et al. (2006)
–	Chinese - Thaiwanese SL	1983	–	11501	–	2159	Chiu et al. (2007)
–	Catalan - Catalan SL	199	2416	4305	446	648	Massó and Badia (2010)
ATIS	English - Irish SL	595	4436	4333	600	544	Bungeroth et al. (2008)
ATIS	German - German SL	595	4903	4291	627	498	Bungeroth et al. (2008)
ATIS	English - South African SL	595	4436	2525	600	422	Bungeroth et al. (2008)
ID	Spanish - Spanish SL	2000	20000	15000	800	400	San Segundo Hernández et al. (2010)
DL	Spanish - Spanish SL	2000	11000	8000	1000	500	San Segundo Hernández et al. (2010)

Πίνακας 3.1 Στατιστικά παράλληλων σωμάτων ΝΓ για διάφορα ζεύγη γλωσσών

λο γλώσσας χρησιμοποιεί τριγράμματα και εξομαλύνεται χρησιμοποιώντας την μέθοδο εξομάλυνσης Kneser-Ney (Kneser and Ney, 1995). Το μοντέλο μετάφρασης βασίζεται σε φράσεις. Ωστόσο, οι βαθμολογίες των αυτόματων και χειρωνακτικών αξιολογήσεων μπορούν να θεωρηθούν συγκρίσιμες και η προσέγγισή τους καταφέρνει να λάβει υπόψη τα γλωσσικά φαινόμενα. Ένα άλλο σύστημα από τον Wu et al. (2007) περιγράφει ένα υβριδικό στατιστικό μοντέλο κανόνων και δεδομένων για μεταφράσεις της κινεζικής γλώσσας στην Ταϊβανική Νοηματική Γλώσσα (TSL). Οι συγγραφείς παραθέτουν ένα δίγλωσσο σώμα 2.036 προτάσεων στην κινεζική, το οποίο είναι ένα από τα μεγαλύτερα σύνολα δεδομένων που έχουν σημειωθεί μέχρι σήμερα. Πρόκειται για ένα σώμα δεδομένων τύπου δέντρου (Treebank) για την κινεζική γλώσσα που περιέχει 36.925 σημειώσεις με χειρονακτικούς σχολιασμούς. Επίσης, αξίζει να σημειωθεί ότι και στα δύο σώματα χρησιμοποιήθηκε για την εξαγωγή μια πιθανολογική γραμματική χωρίς περιεχόμενο (probabilistic context-free grammar - PCFG). Τέλος, για την αξιολόγηση της μετάφρασης χρησιμοποιήθηκαν τέσσερις αλγόριθμοι: τρεις αυτόματες αξιολογήσεις, συγκεκριμένα το AER (Och and Ney, 2000), το TopN (Karypis, 2001) και το BLEU (Papineni et al., 2002) και μια χειρωνακτική αξιολόγηση με βάση τη μέση βαθμολογία γνώμης (MOS).

Στην Ελλάδα, οι πιο ολοκληρωμένες εργασίες στον τομέα της ΜΜ στην ΕΝΓ έχουν παρουσιαστεί στις μελέτες (Efthimiou et al., 2016, Fotinea et al., 2005, Kouremenos et al., 2010). Σε αυτά τα σχήματα, η ΜΜ βασίζεται σε ένα σύστημα κανόνων, χρησιμοποιώντας μια εφαρμογή μεταφοράς κανόνων. Το σύστημα εξάγει νοήματα που κωδικοποιούνται σε μεταγραφές HamNoSys, οι οποίες συνοδεύονται από μη χειρωνακτικά νοήματα και τελικά απεικονίζονται από ένα εικονικό avatar. Επιπλέον, το σύστημα εξάγει και σε μορφή SiGML, παρέχοντας την επιλογή για οπτική προεπισκόπηση του νοήματος. Το σύστημα, στην τρέχουσα κατάστασή του, μπορεί να μεταφράσει αυτόματα μόνο προτάσεις (όχι παραγράφους), ενώ χρησιμοποιεί προσεγγίσεις βασισμένες σε κανόνες και εξάγει σε εικονικό avatar μέσω της χρήσης SiGML (Elliott et al., 2004).

Κεφάλαιο 4

Ένα σύστημα Μηχανικής Μετάφρασης της Ελληνικής προς την ΕΝΓ

Η αρχιτεκτονική του προτεινόμενου συστήματος μηχανικής μετάφρασης της ελληνικής γλώσσας προς την ΕΝΓ (4.2) βασίζεται σε ένα καινοτόμο σύστημα Σύντομης Μεταγραφής Γλώσσας (gloss) της Ελληνικής Νοηματικής Γλώσσας (ΕΝΓ), ή αλλιώς ΣΜΕΝΓ. Στην τελευταία ενότητα 4.3 παρουσιάζουμε ένα νέο σώμα κειμένων που δημιουργήσαμε για τις ανάγκες της εργασίας μας πάνω στη ΜΜ της ΕΝΓ.

4.1 Σύντομη Μεταγραφή της Ελληνικής Νοηματικής Γλώσσας

4.1.1 Εισαγωγή

Στο παρόν κεφάλαιο περιγράφουμε το σύστημα γραπτής Μεταγραφής Γλώσσας της Ελληνικής Νοηματικής Γλώσσας σε σημασιολογικό επίπεδο, για τις ανάγκες της παρούσας διατριβής. Είναι ένα σύστημα το οποίο χρησιμοποιούμε για να αποτυπώσουμε γραπτώς μια ενδιάμεση μορφή σύντομης μεταγραφής γλώσσας (gloss) από τη μια γλώσσα στην άλλη. Η λέξη “gloss” προκύπτει από το λατινικό “glossa” που σημαίνει “γλώσσα” στην Ελληνική και δηλώνει μια οριακά σύντομη μεταγραφή της έννοιας ή σημασίας μια λέξης στο κείμενο.

Για τις ανάγκες της παρούσας διατριβής προτείνουμε ένα καινοτόμο πρότυπο σύστημα Σύντομης Μεταγραφής Γλώσσας (gloss) της Ελληνικής Νοηματικής Γλώσσας (ΕΝΓ),

ή αλλιώς ΣΜΕΝΓ, δίνοντας έμφαση στα γλωσσολογικά χαρακτηριστικά και στη σημασία του νοήματος, παρά στη καταγραφή λεπτομερών κινηματικών χαρακτηριστικών του νοήματος. Η παντελής έλλειψη ενός υπάρχοντος σώματος κειμένων της ΕΝΓ και η αρκετά χρονοβόρα διαδικασία καταγραφής ενός σώματος κειμένων της ΕΝΓ σε επίπεδο κινηματικών χαρακτηριστικών μας οδήγησαν στη δημιουργία ενός συστήματος ΣΜΕΝΓ. Πρόκειται για ένα σύστημα που διατηρεί τα γλωσσολογικά στοιχεία της γλώσσας και τη σημασία του νοήματος που εκφράζει, χωρίς να καταπιάνεται με τις κινηματικές λεπτομέρειες. Εξάλλου, στην πράξη, σε μια γλώσσα (είτε είναι η ελληνική είτε η αγγλική) δίνουμε μεγαλύτερη έμφαση στη σημασία της λέξης, τα γλωσσικά της χαρακτηριστικά, τη γραμματική της, παρά στην προφορά της, αφού για αυτόν τον σκοπό υπάρχουν ειδικά φωνολογικά συστήματα όπως είναι όπως το γνωστό IPA (Decker et al., 1999).

Όπως θα δείτε παρακάτω, για τη δημιουργία του συστήματος ΣΜΕΝΓ μελετήθηκαν διάφορα συστήματα, ένα εκ των οποίων είναι το σύστημα μεταγραφής της ΝΓ του Πανεπιστημίου Berkeley (Berkeley Transcription System - BTS) (Slobin et al., 2001) που περιλαμβάνει κινηματικές πληροφορίες, όπως χειρομορφές, τοποθέτηση στον χώρο, είδη κινήσεων, καθώς και χειροκίνητα (manual) ή μη χειροκίνητα (non manual) στοιχεία. Στην παγκόσμια κοινότητα υπάρχουν και χρησιμοποιούνται διάφορα συστήματα μεταγραφής της ΝΓ, με πολλές παραλλαγές, από απλά μέχρι πολύ αναλυτικά, όπως είναι το IPA που περιέχει περισσότερες λεπτομέρειες καταγραφής της κινησιολογίας του νοήματος. Το κάθε σύστημα που χρησιμοποιείται έχει τα δικά του πλεονεκτήματα και μειονεκτήματα, αναλόγως τον σκοπό για τον οποίο δημιουργήθηκε και τις ανάγκες που καλείται να καλύψει. Για τις ανάγκες της παρούσας διατριβής, λόγω της παντελούς έλλειψης σωμάτων κειμένων της ΕΝΓ, προτιμήθηκε ένα υβριδικό σύστημα μεταγραφής ΝΓ που περιλαμβάνει τα χαρακτηριστικά ενός τυπικού συστήματος σημείωσης γλώσσας (gloss) και μερικά χαρακτηριστικά από το σύστημα μεταγραφής νοηματικής BTS, χωρίς τις αναλυτικές κινηματικές πληροφορίες τύπου IPA ή HamNoSys (Prillwitz et al., 1989).

Το τελικό σύστημα ΣΜΕΝΓ, λόγω της απλότητας του, βοήθησε στην άμεση δημιουργία αξιόπιστων παράλληλων σωμάτων κειμένων της ελληνικής γλώσσας και της ΕΝΓ. Πέρα από τα δύο συστήματα μεταγραφής της ΝΓ που αναφέραμε παραπάνω, χρησιμοποιούνται και άλλα συστήματα γραφής και μεταγραφής της ΝΓ στην παγκόσμια κοινότητα, μερικά εκ των οποίων παρουσιάσαμε αναλυτικά στην ενότητα 2.2. Τα πιο γνωστά και ευρέως χρησιμοποιούμενα συστήματα αυτή τη στιγμή είναι τα εξής:

- Stokoe Notation (Stokoe Jr, 2005),
- SignWriting (Sutton, 1973),
- Hamburg Notation System (Prillwitz et al., 1989),

- Neidle (Neidle et al., 2001)
- SiGML (Elliott et al., 2004), ένα σύστημα αναπαράστασης των τρισδιάστατων κινηματικών ιδιοτήτων της ΝΓ και
- "si5s" ένα πρότυπο σύστημα γραφής (Augustus et al., 2013)

4.1.2 Περιγραφή

Χρησιμοποιώντας τη ΣΜΕΝΓ, γίνεται προσπάθεια να αποδοθεί η έννοια του νοήματος της ΕΝΓ, νόημα προς νόημα, περιλαμβάνοντας όλες τις πρόσθετες σημαντικές γραμματικές πληροφορίες που το συνοδεύουν, όπως είναι οι εκφράσεις προσώπου, οι κινήσεις σώματος, οι κινήσεις των χειλιών, η κατεύθυνση βλέμματος και η αλλαγή ρόλου (role-shift). Με το σύστημα αυτό δεν προσπαθούμε να μεταφράσουμε τη γλώσσα, αλλά να καταγράψουμε περιεκτικά την έννοια του κάθε νοήματος, αποτυπώνοντας τη σημασία του κάθε νοήματος σε γραπτή μορφή που γίνεται κατανοητή κυρίως από γνώστες της ΕΝΓ. Σκοπός μας είναι η μελέτη της ΕΝΓ σε επίπεδο μορφοσυντακτικής δομής και η άμεση δημιουργία παράλληλων σωμάτων κειμένων για τις ανάγκες της παρούσας διατριβής.

Πρέπει να τονίσουμε ότι ένα σύστημα ΣΜΕΝΓ της ΕΝΓ δεν μεταγράφει την ΕΝΓ μαζί με τα πλήρη χειροκίνητα ή μη χειροκίνητα στοιχεία της (χειρομορφές, κινήσεις, εκφράσεις προσώπου κλπ.), γεγονός που επιτάχυνε την όλη διαδικασία της παρούσας εργασίας και μείωσε αρκετά το κόστος παραγωγής των παράλληλων σωμάτων κειμένων. Αν λόγου χάρη χρησιμοποιούσαμε ένα σύστημα μεταγραφής ΝΓ όπως το σύστημα Hamnosys (Hanke, 2004), το οποίο μόνο για τις χειρομορφές έχει 200+ κωδικοποιήσεις[1] και προσθέταμε τις κινήσεις, τις τοποθετήσεις στον χώρο, τα σημεία επαφής μεταξύ χειρομορφών και σημείων στο σώμα, τότε θα μιλούσαμε για έναν τεράστιο αριθμό πληροφοριών και κωδικοποιήσεων. Σε αυτή την περίπτωση, η γραφή ή μεταγραφή της ΕΝΓ για τη δημιουργία σωμάτων κειμένων θα καθιστούσε το εγχείρημά μας εξαιρετικά χρονοβόρο και την όλη διαδικασία και αρκετά δαπανηρή.

4.1.3 Χαρακτηριστικά

Το σύστημα μεταγραφής σχολίου της ΕΝΓ είναι πλήρως συμβατό με τους Η/Υ, μπορεί να μεταγράψει την ΕΝΓ γρήγορα και απλά, δημιουργώντας ψηφιακά παράλληλα σώματα κειμένων της νέας ελληνικής γλώσσας και της ΕΝΓ, για τη μελέτη των γλωσσικών φαινομένων, τη μελέτη και ανάπτυξη γλωσσικών μοντέλων σε Η/Υ, την ανάπτυξη γλωσσικών εφαρμογών της πληροφορικής και την επεξεργασία της πληροφορίας της ΕΝΓ. Το

[1] http://aslfont.github.io/Symbol-Font-For-ASL/asl/handshapes.html

σύστημα μεταγραφής σχολίου της ΕΝΓ που χρησιμοποιούμε έχει τα παρακάτω κύρια χαρακτηριστικά:

- δυνατότητα γραμμικής αναπαράστασης σε μια συνεχόμενη γραμμή εκτύπωσης χρησιμοποιώντας μόνο χαρακτήρες ASCII (linear representation on a continuous typed line, using only ASCII characters).

- δυνατότητα αναπαράστασης μόνο στο επίπεδο "σημασίας" των συστατικών του νοήματος, αποφεύγοντας τη λεπτομερή καταγραφή της κινησιολογίας του νοήματος, δηλαδή την καταγραφή της μορφή της παλάμης ενός χεριού ή και των δύο, καθώς και την(τις) θέση(εις) του(ς) στον χώρο και στον χρόνο καθόλη τη διάρκεια του νοήματος (representation at the level of meaning components).

- πλήρης αναπαράσταση των πολυδιάστατων στοιχείων για τα ρήματα της ΕΝΓ, που είναι ένα ιδιαίτερο χαρακτηριστικό γνώρισμα της ΕΝΓ, αλλά και άλλων ΝΓ στο κόσμο (full representation of elements of polycomponential verbs).

- αναπαράσταση των χειροκίνητων (manual) αλλά και μη χειροκίνητων (non-manual) στοιχείων της ΕΝΓ, όπως εκφράσεις προσώπου και σώματος (representation of manual and non-manual elements).

- αναπαράσταση της κατεύθυνσης βλέμματος, μετατόπιση ρόλου και οπτικής προσοχής (representation of gaze direction, role shift, visual attention).

- αναπαράσταση ειδικών νευμάτων, καθώς και διαφόρων ιδιαίτερων επικοινωνιακών πράξεων (representation of gestures and other communicative acts).

Μια μεταγραφή σχολίου γράφεται πάντα με κεφαλαία ελληνικά γράμματα και αποτυπώνει ένα νεύμα (χειροκίνητο/manual) ή συνδυασμό νεύματος ή νευμάτων, μαζί με μη χειροκίνητα στοιχεία (non-manual) της ΕΝΓ που εκφράζουν μια σημασία ή έννοια. Επίσης, είναι σημαντικό να τονίσουμε ότι το "σχόλιο" (gloss) που αποτυπώνουμε με κεφαλαία αντιστοιχεί στο λήμμα της ελληνικής γλώσσας που συνήθως εκφράζει το συγκεκριμένο νόημα ή τη συγκεκριμένη σημασία στην ΕΝΓ. Τα λήμματα που χρησιμοποιούμε ως σχόλιο είναι άκλιτα και γυμνά από γραμματικές πληροφορίες, γιατί στην ΕΝΓ τα νοήματα δεν περιέχουν κλιτικές πληροφορίες. Οι πληροφορίες αυτές δίνονται με ξεχωριστά ειδικά νεύματα που προσδίδουν τις κλιτικές πληροφορίες στο κύριο νεύμα. Τα κλιτικά αυτά νεύματα έχουν κωδικοποιηθεί, όπως και διάφορα ειδικά χαρακτηριστικά τα οποία μπορείτε να δείτε συγκεντρωτικά παρακάτω.

Αυτόνομο νόημα σχολίου

Για να αποδώσουμε αυτόνομα νοήματα σε μορφή σχολίου χρησιμοποιούμε τις παρακάτω κωδικοποιήσεις:

ΣΧΟΛΙΟ	Για την απόδοση της έννοιας ενός νοήματος της ΕΝΓ το οποίο περιλαμβάνει από ένα νεύμα έως μια σειρά νευμάτων μαζί με μη χειροκίνητα νοήματα, σε μια φυσιολογική και ουδέτερη κατάσταση.
ΣΧΟΛΙΟ+, ΣΧΟΛΙΟ++	Για την επανάληψη ενός νεύματος χρησιμοποιούμε το σύμβολο "+" δίπλα από ένα σχόλιο στα ελληνικά, με κεφαλαία γράμματα. Με χρήση δύο "+" έχουμε διπλή επανάληψη, δηλαδή το ίδιο νόημα αποδίδεται συνολικά 3 φορές όταν έχουμε δύο "+" και δύο φορές όταν έχουμε μόνο ένα "+".
ΣΧΟΛΙΟ-ΣΧΟΛΙΟ, ΣΧΟΛΙΟΣΧΟΛΙΟ	Δύο νοήματα μαζί τα οποία δημιουργούν ένα νέο νόημα. π.χ. ΔΕΝ-ΧΡΕΙΑΖΟΜΑΙ.
ΣΧΟΛΙΟ_ΣΧΟΛΙΟ	Δύο ελληνικές λέξεις που περιγράφουν ένα νόημα.
ΣΧΟΛΙΟ(ΔΑ)	Λέξεις που δακτυλίζονται με δακτυλικό αλφάβητο γρήγορα στην ΕΝΓ, σαν να είναι νοηματική.
Σ-Χ-Ο-Λ-Ι-Ο(ΔΑ)	Λέξη άγνωστη που δακτυλίζεται "αργά" στην ΕΝΓ, σαν να είναι δανεική λέξη.

Αντωνυμίες

Για να αποδώσουμε σε μορφή σχολίου τις αντωνυμίες πρώτου, δεύτερου και τρίτου προσώπου χρησιμοποιούμε τη κωδικοποίηση σχολίου ΑΝΤ_1, ΑΝΤ_2 και ΑΝΤ_3, δηλαδή το ΑΝΤ_1 σημαίνει "ΕΓΩ", ενώ για τις κτητικές αντωνυμίες χρησιμοποιούμε το ΚΤΗΤ_1, ΚΤΗΤ_2 και ΚΤΗΤ_3. Η προσθήκη "+" στο τέλος προσθέτει το αντίστοιχο πληθυντικό μέρος της συγκεκριμένης αντωνυμίας. π.χ. το ΑΝΤ_1+ σημαίνει "ΕΜΕΙΣ" και το ΚΤΗΤ_1+ είναι το "ΜΑΣ".

ANT_1,2,3 ANT_1+,2+,3+	Προσωπικές αντωνυμίες όπου + το πληθυντικό μέρος.
ΚΤΗΤ_1,2,3, ΚΤΗΤ_1+,2+,3+	Κτητικές αντωνυμίες όπου + το πληθυντικό μέρος.

Πολυσυστατικά ρήματα

Τα πολυσυστατικά ρήματα της ΕΝΓ είναι κυρίως τα ρήματα που χρησιμοποιούν μετακίνηση ή την κίνηση στον χώρο για να αποδώσουν την τελική σημασία. Στην περίπτωση των ρημάτων αυτών χρησιμοποιούμε ως μορφή σχολίου την αρχική στατική μορφή του βασικού συστατικού του νοήματος, δηλαδή της χειρομορφής χωρίς την κίνηση, και στη συνέχεια περιγράφουμε την ενέργεια/κίνηση που επιτελείται με τη συγκεκριμένη χειρομορφή-συστατικό, με την οποία επιτυγχάνουμε την τελική αποδομένη σημασία. Δηλαδή έχουμε τη μορφή "ΣΧΟΛΙΟ:ΕΝΕΡΓΕΙΑ". Ας δούμε για παράδειγμα την πρόταση "Παρκάρω το αμάξι". Σε αυτή την περίπτωση έχουμε τη μεταγραφή σχολίου "ΑΜΑΞΙ:ΠΑΡΚΑΡΩ". Το σχόλιο "ΑΜΑΞΙ" είναι η στατική χειρομορφή "αμάξι" την οποία αποδίδουμε στην ΕΝΓ σαν ουσιαστικό και στη συνέχεια προσθέτουμε την κίνηση/ενέργεια "ΠΑΡΚΑΡΩ", με την οποία η ίδια χειρομορφή επιτελεί μια συγκεκριμένη κίνηση που στη ΝΓ σημαίνει "παρκάρω το αμάξι". Ας δούμε άλλο ένα παράδειγμα, όπως την πρόταση "Ο ήλιος ανατέλλει", όπου έχουμε τη μορφή σχολίου "ΗΛΙΟΣ:ΑΝΑΤΕΛΛΕΙ". Ο "ΗΛΙΟΣ" εκφράζει τη στατική χειρομορφή "ήλιος" με την οποία γίνεται η ενέργεια/κίνηση "ΑΝΑΤΟΛΗ" για να έχουμε τελικά "την ανατολή του ήλιου" στην ΕΝΓ.

Σε κάποιες περιπτώσεις πολυσυστατικών ρημάτων της ΕΝΓ ίσως χρειαστεί να αναφέρουμε πρόσθετες πληροφορίες για την τοποθεσία, το είδος κίνησης, αλλά και μη χειροκίνητα στοιχεία (non-manual). Τότε συναντούμε τις παρακάτω κωδικοποιήσεις σχολίου, κατά περίπτωση:

/ΤΠΘ(Χ)	Όπου "ΤΠΘ" η κωδικοποίηση τοποθεσίας και "Χ" η περιγραφή χώρου.
/ΚΝΣ(Χ)	Όπου "ΚΝΣ" η κωδικοποίηση της κίνησης του νοήματος και "Χ" η περιγραφή της κίνησης. π.χ. κίνηση ζιγκ-ζαγκ, κίνηση κύματος κλπ.

Μη χειροκίνητα συστατικά (Non-manual components)

Το μη χειροκίνητο (non-manual) συστατικό νοήματος της ΕΝΓ αφορά τα μέρη του σώματος που συνοδεύουν ένα νόημα, εκτός από τη χειρομορφή (manual). Τέτοια μέρη του σώματος είναι οι εκφράσεις προσώπου, οι κινήσεις των χειλιών, οι ώμοι, οι κινήσεις του κεφαλιού, οι κινήσεις του σώματος, τα οποία μπορούν να προσθέσουν σημαντικές

γραμματικές πληροφορίες σε ένα χειροκίνητο νόημα (manual sign) που συντελείται με τα χέρια.

Χρησιμοποιώντας την κωδικοποίηση ΜΧ(Χ) περιγράφουμε το σύνολο μη χειροκίνητων συστατικών του νοήματος, τα οποία αποδίδουν κάποιες ιδιότητες ή χαρακτηριστικά στο νόημα μέσω της περιγραφής του "Χ". Για παράδειγμα, ένα νόημα που γίνεται πιο αργά από το συνηθισμένο τέμπο (ΜΧ(ΑΡΓΑ)), ή γρήγορα (ΜΧ(ΓΡΗΓΟΡΑ)), με έκφραση απορίας (ΜΧ(ΑΠΟΡΙΑ)), ή με έμφαση (ΜΧ(ΕΜΦΑΣΗ)). Με την κωδικοποίηση ΜΧ(Χ) δεν καταπιανόμαστε με τεχνικές λεπτομέρειες των μη χειροκίνητων συστατικών, μιας και ένας γνώστης της ΕΝΓ είναι εξοικειωμένος με αυτές, όπως συμβαίνει αντίστοιχα και στην γραπτή ελληνική γλώσσα με τη χρήση των σημείων στίξεως του θαυμαστικού ή ερωτηματικού.

Ο σκοπός του συστήματος ΣΜΕΝΓ, όπως έχουμε αναφέρει και παραπάνω, είναι να δώσουμε έμφαση στην απόδοση της σημασίας και της έννοιας ενός νοήματος, συμπεριλαμβάνοντας και τα μη χειροκίνητα συστατικά. Θεωρούμε ότι ένας γνώστης της ΕΝΓ γνωρίζει πώς μπορεί να τα αποδώσει αυτά.

/ΜΧ(Χ)	Όπου ΜΧ το Μη Χειροκίνητο (Non-Manual) συστατικό και Χ η περιγραφή του.
/ΧΛ(Χ)	Όπου ΧΛ η κωδικοποίηση που αφορά τα χείλια και Χ η περιγραφή κινήσεων των χειλιών.
/ΜΤ(Χ)	Όπου ΜΤ η κωδικοποίηση που αφορά τα μάτια και Χ η περιγραφή κινήσεων των ματιών.
/ΜΓΛ(Χ)	Όπου ΜΓΛ η κωδικοποίηση που αφορά τα μάγουλα και Χ η περιγραφή των κινήσεών τους. Παράδειγμα ΜΓΛ(ΦΟΥΣΚΩΜΕΝΑ), σημαίνει ότι έχουμε φουσκωμένα μάγουλα.

Ιδιόμορφα σχόλια

Υπάρχουν κάποια νοήματα στην ΕΝΓ που δύσκολα μπορούν να μεταφραστούν και να μεταγραφούν με τη μορφή λήμματος σχολίου με κεφαλαία γράμματα στην ελληνική. Ο λόγος είναι ότι είναι αμφίσημα και έχουν διαφορετική σημασία κατά περίπτωση. Σε αυτές τις περιπτώσεις, τα νεύματα αποτυπώνονται ξεχωριστά, με ειδικά σχόλια, που είναι κατανοητά μόνο από γνώστες της ΕΝΓ. Ένα χαρακτηριστικό παράδειγμα είναι η μεταγραφή σχολίου "ΠΑ!" της ΕΝΓ που εκφράζει γενικώς κάτι που είναι τετελεσμένο, και συνεργάζεται πολύ με ρήματα πράξεων και χρόνου, στοχεύοντας στο να δείξει ότι μια πράξη ή ενέργεια έχει τελειώσει οριστικά και αμετάκλητα, δίνοντας μια ιδιαίτερη έμφαση σε αυτό. Το gloss "ΠΑ!" γράφτηκε έτσι γιατί παράλληλα με το νεύμα συνδυάζεται και

η κίνηση των χειλιών που εκφέρουν την λέξη "ΠΑ" και μάλιστα σε αρκετές περιπτώσεις οι νοηματιστές το εκφέρουν έτσι.

/ΠΑ!	Κάτι που είναι τετελεσμένο, έχει τελειώσει οριστικά και αμετάκλητα, δίνοντας μια έμφαση σε αυτό.

Εναλλαγή ρόλων (Role shift)

Ένα ιδιαίτερα σημαντικό στοιχείο που συναντάμε συχνά στην ΕΝΓ, αλλά και σχεδόν σε όλες της ΝΓ ανά τον κόσμο είναι η εναλλαγή ρόλων (Role shift), δηλαδή όταν η ενέργεια ενός ρήματος πραγματοποιείται από διαφορετικά υποκείμενα με εναλλαγές ρόλων. Αυτό γίνεται με συγκεκριμένες κινήσεις στην ΝΓ, σε συνδυασμό και με τη χρήση βλέμματος. Για παράδειγμα, ο Πέτρος έδωσε νερό στη γάτα και η γάτα πίνει νερό. Εδώ χρησιμοποιείται η εναλλαγή ρόλων για να υποδείξει τον Πέτρο ή την γάτα κάθε φορά. Για τον σκοπό αυτόν στο σύστημά μας χρησιμοποιούμε τις δεικτικές αντωνυμίες και μέσα σε παρένθεση το υποκείμενο της πράξης /ΕΡ(υποκείμενο), προκειμένου να υποδείξουμε την εναλλαγή ρόλων, ενώ ταυτόχρονα κρατάμε στη μνήμη το "σημείο" του χώρου που αντιπροσωπεύει τον σκύλο, ώστε εάν αργότερα ξαναγίνει αναφορά στον σκύλο να "δείξουμε" το ίδιο σημείο.

/ΕΡ(Χ)	Όπου ΕΡ η κωδικοποίηση της αντωνυμίας όπως έχουμε περιγράψει στο παρόν κεφάλαιο και Χ το υποκείμενο του θέματος.

Για παράδειγμα έχουμε:
ΣΚΥΛΟΣ ΕΡ(ΣΚΥΛΟΣ) ΓΑΥΓΙΖΕΙ/ΜΧ(ΕΝΤΟΝΑ).
που σημαίνει ότι ο νοηματιστής αναφέρει ως κύριο θέμα τον σκύλο και στη συνέχεια παίρνει το ρόλο του σκύλου και κάνει νοήματα σαν να γαυγίζει ο ίδιος, αλλά στην ουσία γαυγίζει ο σκύλος. Αυτό ονομάζεται εναλλαγή ρόλων και είναι συχνό φαινόμενο που συναντάται όχι μόνο στην ΕΝΓ αλλά σε όλες τις ΝΓ του κόσμου.

Κατεύθυνση βλέμματος (Gaze)

Είναι συχνά σημαντικό να γνωρίζουμε που κοιτά ένας νοηματιστής όταν νοηματίζει. Η κατεύθυνση του βλέμματος προσδιορίζεται από την κωδικοποίηση /ΒΛΜ και όπου Χ είναι ο στόχος/αντικείμενο που κοιτάζει το βλέμμα του νοηματιστή.

/ΒΛΜ(Χ)	Όπου /ΒΛΜ η κωδικοποίηση της κατεύθυνσης βλέμματος και Χ ο στόχος του βλέμματος

Για παράδειγμα :
/ΒΛΜ(ΒΙΒΛΙΟ) κοιτάει το βιβλίο.
/ΒΛΜ(ΚΩΣΤΑΣ) κοιτάει τον Κώστα.

4.1.4 Γενική συζήτηση και στόχοι

Βάσει της ανάλυσης του Κεφαλαίου 3, είναι προφανές ότι τα υπάρχοντα συστήματα ΜΜ της ΝΓ ακολούθησαν τις σύγχρονες τάσεις της ΜΜ. Στο πλαίσιο του συστήματος RBMT, δόθηκε έμφαση στη μοντελοποίηση συγκεκριμένων φαινομένων, χρησιμοποιώντας θεωρίες και φορμαλισμούς από την υπολογιστική γλωσσολογία. Αντίθετα, το μοντέλο με βάση τα δεδομένα εκμεταλλεύεται τις στατιστικές κανονικότητες που υπάρχουν στα διαθέσιμα παράλληλα δεδομένα, τα οποία είναι σπάνια για τις ΝΓ. Ωστόσο, τα παράλληλα σώματα είναι απαραίτητα για τη ΜΜ της ΝΓ και αναμένεται να παραμείνουν σπάνια στο προσεχές μέλλον, εκτός εάν οι ερευνητικές προσπάθειες επικεντρωθούν στην απόκτησή τους.

Η τρέχουσα εργασία επιχειρεί να αντιμετωπίσει δύο πολύ σοβαρά προβλήματα της ΕΝΓ: (α) την έλλειψη γραμματικής και (β) την έλλειψη μεγάλων παραλλήλων σωμάτων μεταξύ της ελληνικής και της ΕΝΓ. Προς αυτές τις κατευθύνσεις προτείνεται μια μεθοδολογία επεξεργασίας για τη δημιουργία μεγάλων παράλληλων δεδομένων καλής ποιότητας για την ΕΝΓ, με τη βοήθεια ενός επαγγελματία μεταφραστή και ενός απλού συστήματος ΜΜ με κανόνες. Ο μεταφραστής χρησιμοποιεί ένα απλό σύστημα ΜΜ με κανόνες, που βασίζεται στη γλώσσα προγραμματισμού Python, εργαλεία ανοιχτού κώδικα που ενσωματώνουν ένα υποσύστημα μεταφοράς και ένα εύρωστο υποσύστημα (module) μεταφοράς γραμματικής τύπου δέντρου. Τέλος, με τη δημιουργία παράλληλων σωμάτων μπορούμε στη συνέχεια να εκπαιδεύσουμε ένα στατιστικό σύστημα ΜΜ για τη δική μας περίπτωση, χρησιμοποιώντας την εργαλειοθήκη της πλατφόρμας Moses (Koehn et al., 2007).

Η πλατφόρμα Moses είναι μια εργαλειοθήκη εφαρμογών της στατιστικής προσέγγισης στη ΜΜ. Αποτελεί την κυρίαρχη τάση στον τομέα της ΜΜ αυτή τη στιγμή και χρησιμοποιείται από συστήματα μετάφρασης που αναπτύσσονται από εταιρείες-κολοσσούς όπως η Google και η Microsoft. Στη στατιστική μηχανική μετάφραση (ΣΜΜ) τα μεταφραστικά συστήματα εκπαιδεύονται: (α) από μεγάλες ποσότητες παράλληλων δεδομένων (σώματα/corpora), από τα οποία τα συστήματα μαθαίνουν να μεταφράζουν μικρά τμήματα/προτάσεις και (β) από ακόμη μεγαλύτερες ποσότητες μονογλωσσικών δεδομένων, από τα οποία τα συστήματα μαθαίνουν κατά πόσον η ΓΣ θα πρέπει να μοιάζει με τη ΓΠ. Τα παράλληλα δεδομένα είναι μια συλλογή προτάσεων σε δύο διαφορετικές γλώσσες, η οποία είναι ευθυγραμμισμένη ανά πρόταση, καθώς κάθε πρόταση σε μια γλώσσα αντιστοιχεί σε

μία μεταφρασμένη πρόταση στην άλλη γλώσσα. Η κατηγορία αυτή είναι επίσης γνωστή ως δίγλωσσο σώμα κειμένων (bitext).

4.2 Αρχιτεκτονική του προτεινόμενου συστήματος

Το προτεινόμενο σύστημα ΜΜ έχει λάβει υπόψη τις αρχές της Βασικής Γραμματικής Μοντελοποίησης (Basic Unification Grammar) (Carpenter, 2005, 1992, Kay, 1984, Shieber, 2003). Αποτελείται από δύο βασικά στάδια: στο πρώτο στάδιο χρησιμοποιούμε την εφαρμογή ΜΜ με κανόνες (Rule Based Machine Translation / RBMT), την οποία αναπτύξαμε με σκοπό την παραγωγή των παράλληλων σωμάτων υπό την επίβλεψη επαγγελματία μεταφραστή και στο δεύτερο στάδιο έχουμε την εφαρμογή ενός έτοιμου ανοιχτού λογισμικού συστήματος στατιστικής μετάφρασης που βασίζεται σε ευθυγραμμισμένα παράλληλα κείμενα. Για τη συνολική ανάπτυξη της παρούσας εργασίας έχουν συνδυαστεί διάφορα εργαλεία και τεχνολογίες: (α) ο αναλυτής σε μέρη του λόγου AUEB's POS Parser που χρησιμοποιήθηκε για την πραγματοποίηση μορφολογικού σχολιασμού (annotation) στο σώμα της ΓΣ, (β) η βιβλιοθήκη γλωσσικών εφαρμογών NLTK έκδοσης 3.0 (Natural Language Toolkit) της γλώσσας προγραμματισμού Python, η οποία είναι ένα εργαλείο ανοιχτού κώδικα που διατίθεται δωρεάν, αναπτύσσεται από την κοινότητα χρηστών και αποτελεί μια κορυφαία πλατφόρμα για την ανάπτυξη προγραμμάτων σε γλώσσα προγραμματισμού Python, με σκοπό την επεξεργασία και ανάλυση γλωσσικών δεδομένων, (γ) η γλώσσα προγραμματισμού Java, (δ) η αντικειμενοστραφής γλώσσα προγραμματισμού Perl scripts και (ε) η εργαλειοθήκη ΣΜΜ της πλατφόρμας Moses.

Για το πρώτο στάδιο της παρούσας εργασίας, η διαδικασία της ΜΜ με κανόνες (RBMT) που προτείνεται για την υποβοήθηση στη μετάφραση και τη δημιουργία των παράλληλων σωμάτων κειμένων της ελληνικής γλώσσας και της ΕΝΓ είναι αναδρομική. Χρησιμοποιούμε κάθε φορά τους κανόνες μεταφοράς που αναπτύσσουμε σταδιακά (βήμα-βήμα) κατά τη διαδικασία της δημιουργίας των παράλληλων σωμάτων κειμένων. Οι κανόνες παράγουν ένα κείμενο στη ΓΣ (την ΕΝΓ), το οποίο ελέγχουμε και διορθώνουμε με τη βοήθεια επαγγελματία μεταφραστή. Κατά τη διαδικασία ελέγχου και διόρθωσης, αν παρατηρήσουμε νέα γλωσσικά φαινόμενα, δημιουργούμε κάθε φορά νέους κανόνες μεταφοράς του συστήματος ΜΜ κανόνων (transfer rules of RBMT system), οι οποίοι λαμβάνονται υπόψη για τα επόμενα τμήματα του σώματος κειμένων, ώστε να ελαχιστοποιείται ο χρόνος διόρθωσης και ελέγχου. Για παράδειγμα, ας υποθέσουμε ότι στα πρώτα τμήματα του σώματος κειμένων παρατηρείται ότι οι χρονικές προτάσεις τύπου ονομασίας, π.χ. "Τα Χριστούγεννα", πηγαίνουν πάντα στην αρχή της πρότασης και το άρθρο "Τα" φεύγει. Σε αυτή την περίπτωση, φτιάχνουμε τους ανάλογους κανόνες τεμαχισμού (chunks rules) και

κανόνες λέξεων (words rules) και στη συνέχεια, όταν θα εξετάσουμε τα επόμενα τμήματα του σώματος κειμένων, δεν θα χρειαστεί να ξαναδιορθώσουμε το ίδιο φαινόμενο, παρά να ασχοληθούμε με κάποιο νέο. Με αυτόν τον τρόπο επιταχύνουμε τη δημιουργία των παράλληλων σωμάτων κειμένων και ελέγχουμε πάντα κατά πόσον οι τρέχοντες κανόνες εφαρμόζονται σωστά κάθε φορά στα τρέχοντα σώματα κειμένων.

Καθόλη τη διαδικασία της ΜΜ του πρώτου σταδίου, η ΜΜ εποπτεύεται από έναν επαγγελματία μεταφραστή, έτσι ώστε να διορθώνονται τα κείμενα εξόδου και δημιουργούνται νέοι κανόνες μεταφοράς ή λεξικοί κανόνες, εφόσον αυτό απαιτείται, οι οποίοι προστίθενται στο σύστημα ΜΜ με κανόνες (RBMT), ενώ παράλληλα δημιουργείται λεξικό δεδομένων, έτσι ώστε να καλύπτονται τυχόν νέες περιπτώσεις.

Με τη βοήθεια του συστήματος RBMT δημιουργούμε, λοιπόν, παράλληλα κείμενα γραπτού λόγου της ελληνικής και γραπτής μεταγραφής της ΕΝΓ. Στη συνέχεια, έπειτα από ειδική προετοιμασία και προεργασία, εκπαιδεύουμε την ανοιχτή πλατφόρμα στατιστικής μετάφρασης Moses με το μεγαλύτερο μέρος των κειμένων αυτών (90%). Τέλος, το μέρος των κειμένων που εξαιρέθηκε από τη διαδικασία εκπαίδευσης του συστήματος (10%), χρησιμοποιήθηκε αποκλειστικά για την αξιολόγηση του συστήματος με χρήση του αλγόριθμου αξιολόγησης μετάφρασης BLUE Scoring Translation Accuracy της Papineni et al. (2002).

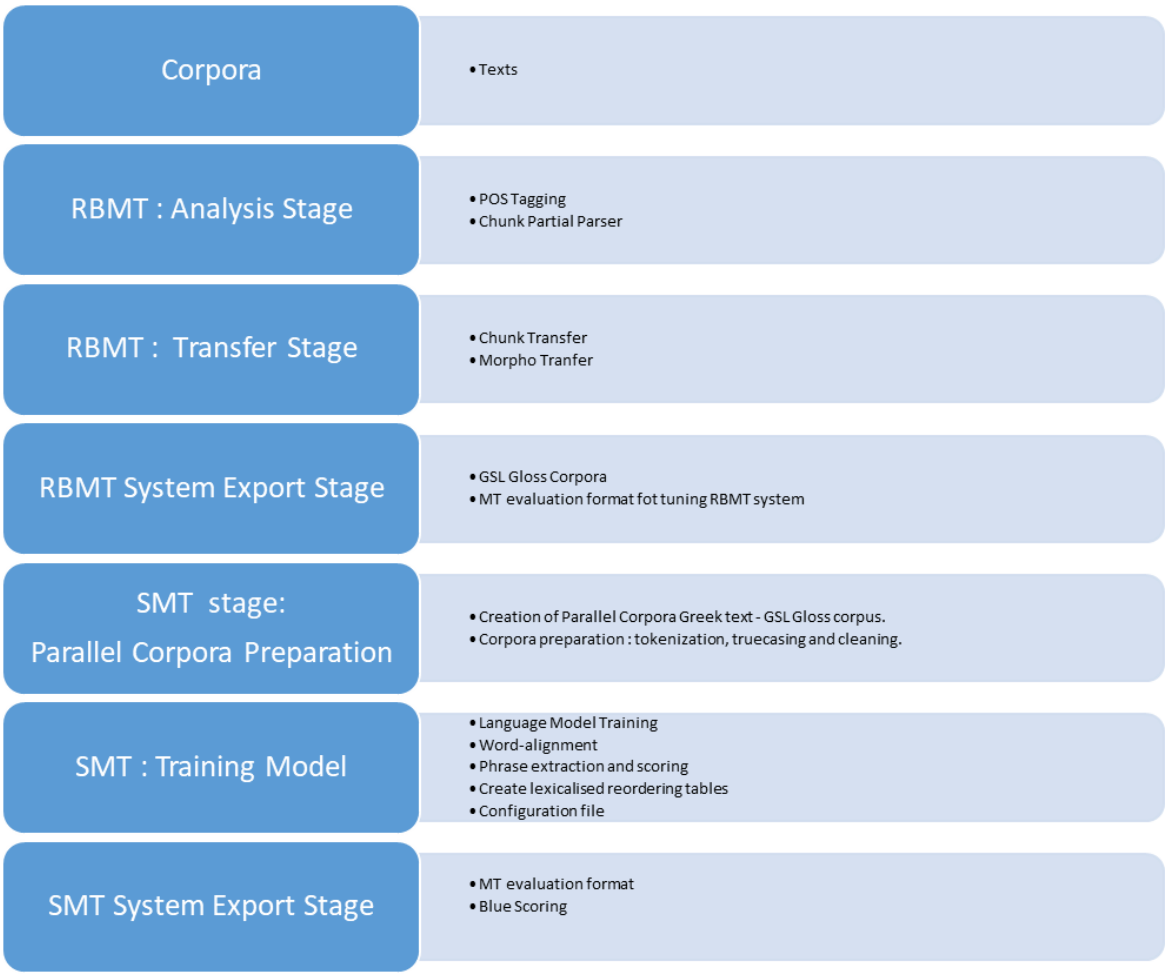
4.2.1 Αρχιτεκτονική συστήματος

Προκειμένου να μεταφράσουμε τα ελληνικά σώματα κειμένων στην ΕΝΓ, πραγματοποιείται μεταφορά στο συντακτικό επίπεδο, μέσω της χρήσης μερικής ανάλυσης δέντρων. Το Σχήμα 4.1 απεικονίζει την αρχιτεκτονική του συνολικού συστήματος, συμπεριλαμβανομένων των συνδέσεών του με τις υπόλοιπες εξωτερικές μονάδες.

- Σώματα κειμένων: Αρχικά, πραγματοποιώντας εξόρυξη κειμένων από αρκετές ιστοσελίδες που σχετίζονται με την πρόβλεψη του καιρού, δημιουργήσαμε ένα μεγάλο σώμα απλών κειμένων σε μορφή ASCII , αποτελούμενο από 1.015 προτάσεις και 20.287 λέξεις (tokens). Περισσότερες πληροφορίες μπορείτε να βρείτε στην ενότητα 7.1.

- POS Tagging: Μορφολογικός αναλυτής που παρέχεται δωρεάν μέσω της εφαρμογής του Πανεπιστημίου AUEB's Greek POS Parser και τη μελέτη της Koleli (Koleli, 2011). Για περισσότερες πληροφορίες μπορείτε να ανατρέξετε στην ενότητα 5.2.1.

- Chunk Partial Parser: Χρησιμοποιώντας τον αναλυτή φράσεων (chunk parser) και τη γραμματική κανονικών εκφράσεων (regular grammar) από την εργαλειοθήκη Python's



Σχήμα 4.1 Αρχιτεκτονική Συστήματος

NLTK Toolkit, εφαρμόζεται ένας μερικός τεμαχισμός φράσεων (partial chunking) και οι προτάσεις διαιρούνται σε υπο-προτάσεις, δημιουργώντας ένα μορφολογικό δέντρο (constituency tree). Για πιο αναλυτική περιγραφή μπορείτε να ανατρέξετε στην ενότητα 5.2.2.

- Chunk Transfer: Υποεφαρμογή μεταφοράς μορφολογικού δέντρου του συστήματός μας (chunk transfer module), με χρήση του δίγλωσσου λεξικού και της βάσης δεδομένων των κανόνων μεταφοράς (που περιέχει γνώσεις με γραμματικά φαινόμενα της ΕΝΓ), μέσω της οποίας πραγματοποιούμε μεταφορά της μορφολογικής δομής δέντρου της ελληνικής στην αντίστοιχη μορφολογική δομή δέντρου της ΕΝΓ. Για περισσότερες πληροφορίες μπορείτε να ανατρέξετε στην ενότητα 5.2.1.

- Morpho Transfer module: Υποεφαρμογή μορφολογικής μεταφοράς με χρήση: (α) ενός δίγλωσσου λεξικού ΕΝΓ-ελληνικής, (β) λημματοποιητή της εφαρμογής μας σε γλώσσα Python (python lemmatizer) και (γ) μορφολογικών και γραμματικών κανόνων που ενεργούν βάσει των γλωσσολογικών πληροφοριών του ελληνικού μορφολογικού αναλυτή της Koleli (POS Parser) (Koleli, 2011). Η υποεφαρμογή παράγει μια σειρά ΣΜΕΝΓ με μορφολογικές ετικέτες και ετικέτες με μη χειρωνακτικά νοήματα. Δείτε περισσότερα στην ενότητα 5.2.2.

- GSL gloss corpora: Το προτεινόμενο σύστημα λαμβάνει τις εξόδους από τα στάδια ανάλυσης και μεταφοράς και παράγει διαφορετικά είδη σωμάτων κειμένων της ΕΝΓ. Δείτε περισσότερα στην ενότητα 5.2.4.

- Corpora preparation: Κατά την προετοιμασία των σωμάτων κειμένων γίνεται σε πρώτη φάση ο συγχρονισμός των σωμάτων κειμένων της ΓΠ, δηλαδή της ελληνικής, με το σώμα κειμένων της ΓΣ, δηλαδή της ΕΝΓ, και συγκεκριμένα της σύντομης γραπτής μεταγραφής της ΕΝΓ (ΣΜΕΝΓ) που παράγει ως έξοδο το σύστημα ΜΜ κανόνων. Στη συνέχεια, κάνουμε χρήση των εργαλείων της πλατφόρμας Moses για βελτιστοποίηση των σωμάτων κειμένων (Tokenization, Truecasing, cleaning). Τα εργαλεία αυτά εφαρμόζονται κυρίως στο σώμα κειμένων της ΓΠ, αφού η έξοδος καθαρίζεται και βελτιστοποιείται από τα εργαλεία του συστήματός μας. Σε κάθε βήμα κάνουμε έλεγχο με χρήση ειδικών εργαλείων σύγκρισης κειμένου (differencing tool), όπως το WinMerge, το οποίο στην περίπτωσή μας παρουσιάζει κάθε φορά τις διαφορές που εντοπίζονται στο σώμα κειμένων μας με οπτικό τρόπο, ώστε να είναι εύκολο να τις απομονώσουμε και να τις διαχειριστούμε. Στον Πίνακα 4.1 παρακάτω θα δείτε ένα παράδειγμα παράλληλων συγχρονισμένων προτάσεων.

- MT evaluation (Αξιολόγηση ΜΜ): Το προτεινόμενο σύστημα εξάγει τα σώματα κειμένων με την κατάλληλη μορφή, έτσι ώστε να υπολογιστεί η μετρική βαθμολογία του αλγόριθμου BLEU. Για περισσότερες πληροφορίες μπορείτε να ανατρέξετε στις ενότητες 8.7 και 9.1.7.

4.3 Ένα νέο σώμα κειμένων για τη Μηχανική Μετάφραση

Όπως έχουμε ήδη αναφέρει, για τις ανάγκες της παρούσας εργασίας πραγματοποιήθηκε εξόρυξη κειμένων από αρκετές ιστοσελίδες[1] που σχετίζονται με τον καιρό, δημιουργώντας ένα μεγάλο corpus κειμένων ελληνικής γλώσσας, αποτελούμενο από 1.015 φράσεις και 20.287 λέξεις (tokens). Στη συνέχεια το σώμα χωρίστηκε σε 10 υποσώματα, με περίπου 100 προτάσεις το καθένα. Για τις ανάγκες της εργασίας μας συνεργαστήκαμε με έναν έμπειρο επαγγελματία μεταφραστή που είχε την ευθύνη της επιμέλειας των μεταφράσεων των ελληνικών γραπτών σωμάτων σε γραπτές σύντομες μεταγραφές της ΕΝΓ, χαρακτηρισμένες με ετικέτες NmCs (Non-Manual Component Sign)(μη χειρωνακτικά νοήματα). Ο εν λόγω μεταφραστής βοήθησε στην αποτίμηση των γλωσσικών φαινομένων της ΕΝΓ, στη δημιουργία των κανόνων μεταφοράς του συστήματος ΜΜ με κανόνες και τέλος στη δημιουργία των σωμάτων κειμένων αναφοράς για την αξιολόγηση της μετάφρασης μέσω BLUE της (Papineni et al., 2002).

[1] http://www.deltiokairou.gr/, http://www.weather.gr/, http://meteo.gr/

Parallel Corpora (Greek text with GSL gloss text)	
Βροχές και καταιγίδες θα εκδηλωθούν κατά τόπους στη Δυτική Ελλάδα τα Χριστούγεννα.	ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΧΡΙΣΤΟΥΓΕΝΝΑ ΕΧΕΙ ΒΡΟΧΗ ΚΑΙ ΚΑΤΑΙΓΙΔΑ-/ΜΧ(ΕΝΤΑΣΗ)/ΜΓΛ(ΦΟΥΣΚΩΜΕΝΑ) ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΤΟΠΟΣ/ΤΠΘ(Χ1)-ΤΟΠΟΣ/ΤΠΘ(Χ2) ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΛΛΑΔΑ ΔΥΤΙΚΟΣ.
Χιόνια θα πέσουν στα ορεινά της Δυτικής και Βόρειας Ελλάδας σε υψόμετρα μεγαλύτερα των 1400 μέτρων.	ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΧΙΟΝΙ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΟΡΕΙΝΑ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΛΛΑΔΑ ΒΟΡΕΙΟΣ ΔΥΤΙΚΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΥΨΟΜΕΤΡΟ ΜΕΓΑΛΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) 1400 ΜΕΤΡΟ.
Στην υπόλοιπη χώρα αναμένονται τοπικές νεφώσεις, ενώ πρόσκαιρες βροχές δεν αποκλείεται να εκδηλωθούν στα κεντρικά ηπειρωτικά.	ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΧΩΡΑ ΥΠΟΛΟΙΠΟΣ ΑΝΑΜΕΝΩ ΣΥΝΝΕΦΑ/ΜΤ(ΜΙΣΑΝΟΙΧΤΑ ΤΟΠΙΚΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΒΡΟΧΗ ΝΩΡΙΣ ΔΕΝ ΑΠΟΚΛΕΙΕΤΑΙ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΚΔΗΛΩΝΩ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΗΠΕΙΡΟΣ ΚΕΝΤΡΙΚΑ.
Λόγω των υψηλών ποσοστών υγρασίας και των αυξημένων συγκεντρώσεων σκόνης η ορατότητα θα είναι και πάλι τοπικά περιορισμένη.	ΛΟΓΩ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΠΟΣΟΣΤΟ ΨΗΛΟΣ ΥΓΡΑΣΙΑ ΚΑΙ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΣΚΟΝΗ ΜΑΖΕΥΕΙ ΠΟΛΥ/ΧΛ(ΠΟΛΥ) ΟΡΑΤΟΤΗΤΑ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΚΑΙ ΠΑΛΙ ΤΟΠΙΚΑ ΠΕΡΙΟΡΙΣΜΕΝΟΣ.
Οι νότιοι άνεμοι που αρχικά θα πνέουν στα πελάγη με εντάσεις έως 8 μπ.	ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΑΝΕΜΟΣ ΝΟΤΙΟΣ ΠΟΥ ΑΡΧΙΚΑ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΠΝΕΩ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΠΕΛΑΓΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΝΤΑΣΗ ΕΩΣ 8 ΜΠΟΦΩΡ .
Σταδιακά μετά το μεσημέρι θα στραφούν σε δυτικούς νοτιοδυτικούς και θα παρουσιάσουν εξασθένηση στα επίπεδα των 6 μπ.	ΣΤΑΔΙΑΚΑ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΜΕΣΗΜΕΡΙ ΣΤΡΕΦΩ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΔΥΤΙΚΟΣ ΝΟΤΙΟΔΥΤΙΚΟΣ ΚΑΙ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΜΕΙΩΣΗ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΠΙΠΕΔΑ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) 6 ΜΠΟΦΩΡ.

Πίνακας 4.1 Συγχρονισμένες παράλληλες προτάσεις

Κεφάλαιο 5

Στάδια Ανάλυσης, Μεταφοράς και Εξαγωγής

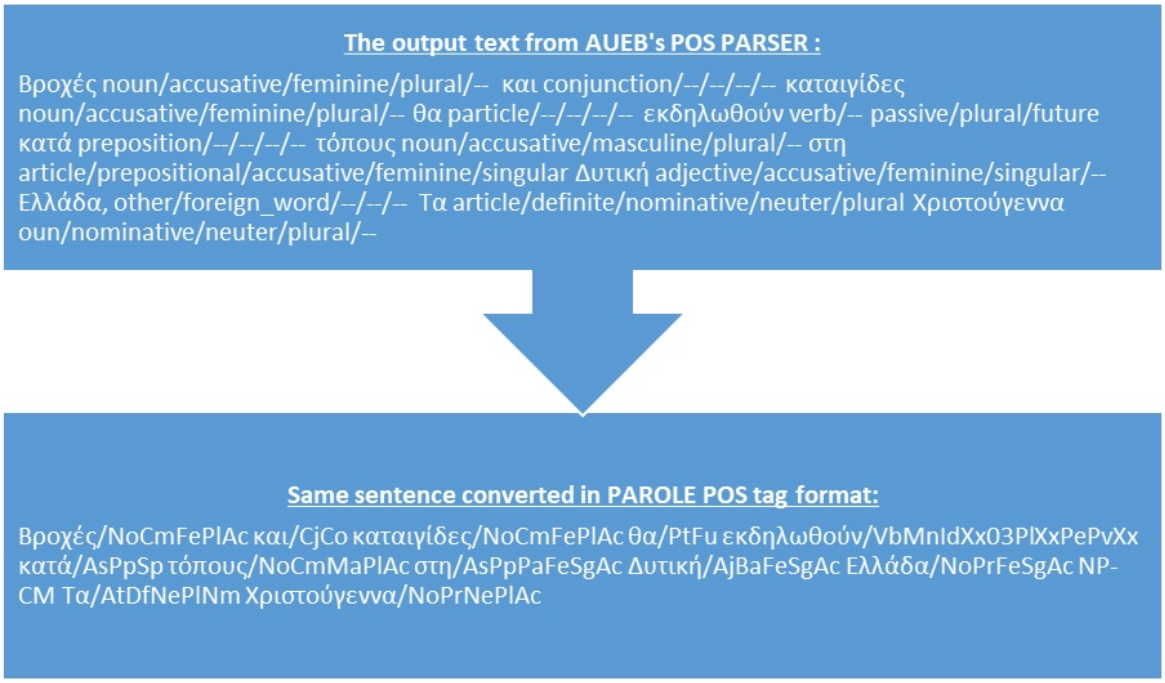
5.1 Στάδιο Ανάλυσης

5.1.1 Ελληνικός αναλυτής σε μέρη του λόγου (Greek POS Parser)

Η γραμματική ανάλυση σε μέρη του λόγου πραγματοποιείται από τον αντίστοιχο ελληνικό αναλυτή σε μέρη του λόγου (Greek Part Of Speech) (Koleli, 2011) του AUEB, που στηρίζεται σε ισχυρή αναζήτηση δεδομένων με βάση την ταξινόμηση μέγιστης εντροπίας. Ο αναλυτής χρησιμοποιεί ετικέτες για τη σήμανση του κειμένου σε μέρη του λόγου (POS tags). Στο Σχήμα 5.1 μπορείτε να δείτε ένα παράδειγμα ελληνικού κειμένου επισημασμένο με ετικέτες POS. Η πρόταση είναι: "Οι βροχές και οι καταιγίδες θα εκδηλωθούν κατά τόπους στην Δυτική Ελλάδα τα Χριστούγεννα". Στη συνέχεια, οι ετικέτες POS μετατρέπονται σε μορφή Parole (Labropoulou et al., 1996), έπειτα από εκτέλεση δικών μας script σε Perl.

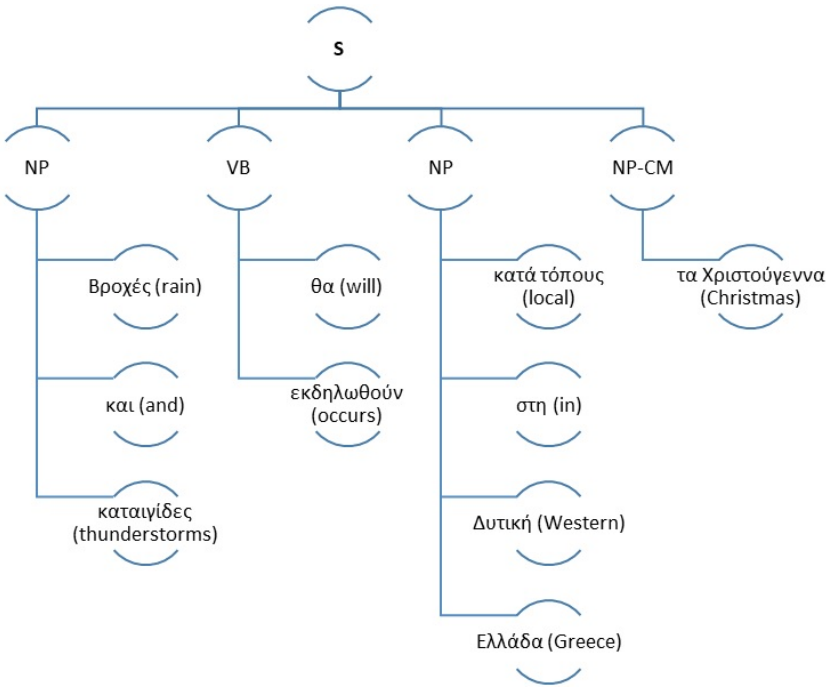
5.1.2 Αναλυτής φράσεων (Chunk parser)

Ο αναλυτής τμημάτων της Python αναλύει και σπάει κάθε πρόταση σε δευτερεύουσες προτάσεις, τις οποίες χαρακτηρίζει μορφολογικά, προσδίδοντάς τους έτσι τη μορφή συντακτικού δέντρου (constituency tree). Ο αναλυτής φράσεων χρησιμοποιεί τη γραμματική κανονικών εκφράσεων (regular expressions grammar). Αυτή η διαδικασία παράγει δομημένα σώματα με τη μορφή δέντρων δεδομένων της Python. Το συντακτικό/μορφολογικό δέντρο (tree constituency) παρέχει μια σύνοψη της δομής της ΕΝΓ. Στο Σχήμα 5.2



Σχήμα 5.1 POS Parsed Sentence

παρουσιάζουμε ένα παράδειγμα ανάλυσης φράσης σε συντακτικό δέντρο, για την προαναφερθείσα πρόταση.



Σχήμα 5.2 Sentence in graphical constituency tree

5.2 Στάδιο μεταφοράς (Transfer Stage)

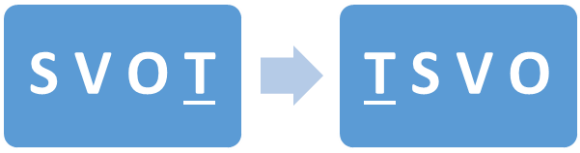
Το στάδιο μεταφοράς μπορεί να χωριστεί σε δύο επιμέρους στάδια: (α) το στάδιο ανάλυσης φράσεων και δημιουργίας του συντακτικού δέντρου (Chunk Transfer) (σε επίπεδο υποπροτάσεων/δέντρου) και (β) το στάδιο μορφολογικής μεταφοράς (Morpho Transfer) (σε επίπεδο λέξης). Κατά τη διάρκεια της μεταφοράς φράσεων, το δέντρο μερικής ανάλυσης μετατρέπεται σε ενδιάμεσο δέντρο ανάλυσης, κατά τα πρότυπα της ΕΝΓ. Στο στάδιο μορφολογικής μεταφοράς (Morpho Transfer), χρησιμοποιούμε ένα δίγλωσσο λεξικό ΕΝΓ-ελληνικής, την εφαρμογή λημματοποιητή και τους γραμματικούς κανόνες από τη βάση δεδομένων του συστήματος ΜΜ κανόνων, που βασίζεται στις γλωσσικές πληροφορίες κάθε λέξης.

Στη συνέχεια, το ενδιάμεσο δέντρο ανάλυσης της ΕΝΓ μετατρέπεται σε μια σειρά γραπτών μεταγραφών της ΕΝΓ, μερικές από τις οποίες είναι χαρακτηρισμένες με μηχειρωνακτικά νοήματα (NmN), όπως για παράδειγμα η λέξη "καταιγίδες" είναι χαρακτηρισμένη με τα στοιχεία NmN "ΜΧ (ΕΝΤΑΣΗ)" (ΣΦΑΙΡΙΚΗ ΕΚΦΡΑΣΗ ΣΗΜΑΤΩΝ) και "ΜΓΛ (ΦΟΥΣΚΩΜΕΝΑ ΜΑΓΟΥΛΑ-SWOLLEN CHEEKS)". Στο σημείο αυτό πρέπει να αναφερθούν τα εξής: (α) το δίγλωσσο λεξικό διευρύνεται με την ενσωμάτωση μορφόλεξικών και λεξικό-σημασιολογικών σχέσεων και (β) μη χειρωνακτικά χαρακτηριστικά προστίθενται σε συγκεκριμένες μεταγραφές της ΕΝΓ, χρησιμοποιώντας μια βάση δεδομένων γνώσεων που δημιουργήθηκε ειδικά για αυτή την εφαρμογή. Η βάση δεδομένων περιλαμβάνει αντιστοίχιση μη χειρωνακτικών χαρακτηριστικών σε λήμματα της ΕΝΓ.

5.2.1 Μεταφορά υποφράσεων (Chunk transfer)

Αρχικά, ένας μερικός αναλυτής φράσεων δημιουργεί ένα μερικό ενδιάμεσο συντακτικό δέντρο φράσεων, έτσι ώστε ο μετασχηματισμός να επιτυγχάνεται σε επίπεδο υποφράσεων του συντακτικού δέντρου, αλλάζοντας τη σειρά των υποφράσεων. Όλοι αυτοί οι μετασχηματισμοί δεν είναι υποχρεωτικοί. Μόνο αφού ικανοποιηθεί μια σειρά από λογικά σενάρια, υποδεικνύεται εάν πρέπει να μεταφερθεί ή όχι κάθε φορά ένα φαινόμενο. Οι κανόνες μεταφοράς του συστήματός μας εφαρμόζονται στην Python και χειρίζονται κυρίως τον τύπο μορφοτύπου δέντρων που παράγει ο αναλυτής της Pyhton (ενότητα 5.1.2). Αυτός ο τύπος δεδομένων (δέντρου) καθιστά την εφαρμογή του κώδικα μεταφοράς απλή, εύκολη, γρήγορη και ευέλικτη και τα αποτελέσματα μπορούν ανά πάσα στιγμή να εμφανίζονται σε γραφική προβολή, χρησιμοποιώντας το ενσωματωμένο εργαλείο προβολής της Python.

Μια συνήθης περίπτωση που συναντάμε είναι η υποφράση που σχετίζεται με τον χρόνο, η οποία μεταφέρεται από το τέλος στην αρχή της φράσης. Το Σχήμα 5.3 απεικονίζει τη νέα σύνταξη, όπου S: Subject, V: Verb, O: Object και T: Time.



Σχήμα 5.3 "SVOT" to "TSVO" transfer

Ο κανόνας του αναλυτή φράσεων για τη μεταφορά από SVOT σε TSVO εφαρμόζεται από τον ακόλουθο κώδικα (5.1):

```
def MoveChunkTime(tree):
    newTree = tree.copy() #Tree('S',[])
    for i in range(len(newTree)):
        if newTree[i].label()=='NP-CM' and newTree[i][1][0]=='Χριστουγ
   εννα':
            newTree.insert(0,tree[i])
            newTree.pop(i+1)
    return newTree
```

Listing 5.1 Python Code "MoveChunkTime(tree)"

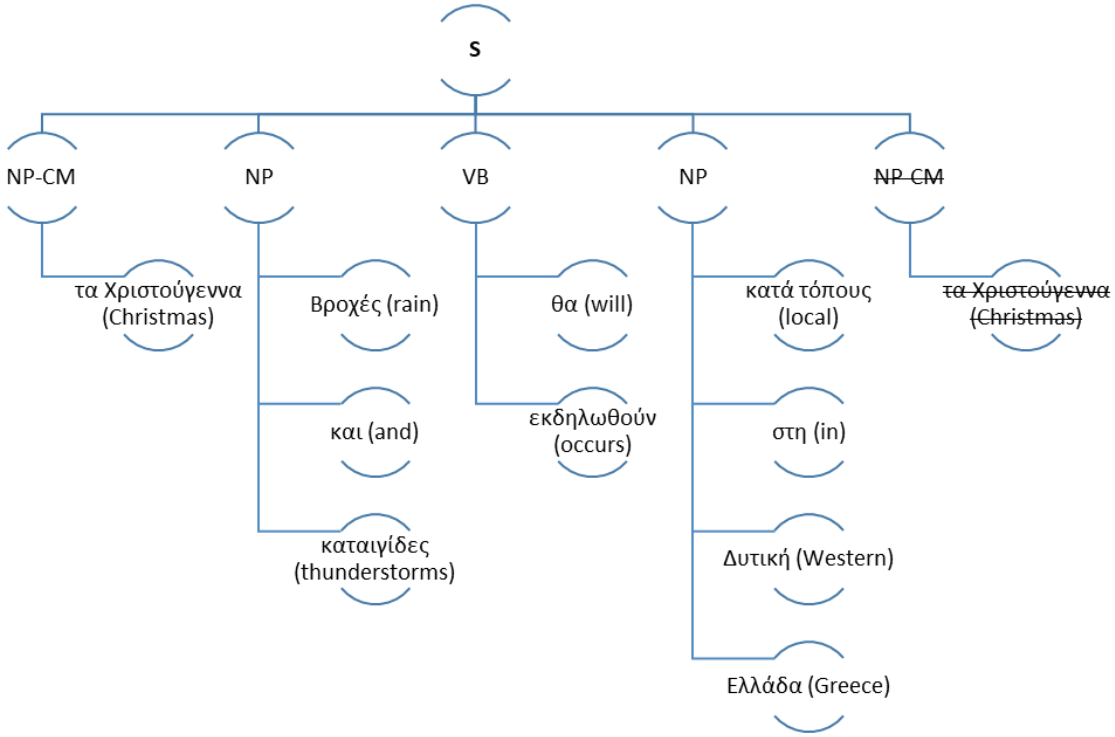
Με βάση τον παραπάνω κανόνα, η υποφράση «τα Χριστούγεννα» (Christmas) της πρότασης του παραδείγματος, μεταφέρεται στην αρχή της πρότασης (Σχήμα 5.4 και 5.5):

5.2.2 Μορφολογική μεταφορά (Morpho Transfer)

Η προτεινόμενη υποεφαρμογή μορφολογικής μεταφοράς (Morpho Transfer) χρησιμοποιεί (α) ένα δίγλωσσο λεξικό ΕΝΓ-ελληνικής, (β) τον λημματοποιητή Pyhton και (γ) μορφολογικούς και γραμματικούς κανόνες που βασίζονται στις γλωσσικές πληροφορίες κάθε λέξης. Στη συνέχεια, ακολουθεί το στάδιο παραγωγής της εξόδου του συστήματος μας, η οποία αποτελείται από μια ακολουθία γραπτών μεταγραφών με ενδείξεις μορφολογικών και μη χειρωνακτικών ετικετών, έτσι ώστε να σχηματιστεί η τελική πρόταση (Hoiting and Slobin, 2002, Slobin et al., 2001).

Η επεξεργασία της λημματοποίησης γίνεται με σκοπό τη μετατροπή όλων των λέξεων σε λήμματα και τη μετατροπή απλού κειμένου σε μορφή σωμάτων XCES corpora format[1] (Ide and Brew, 2000). Για να ολοκληρωθεί αυτή η διαδικασία, έχει ήδη υλοποιηθεί μια εφαρμογή-γέφυρα λεξικής μεταφοράς σε κώδικα της Python, χρησιμοποιώντας

[1] An XML based standard to encode text corpora.



Σχήμα 5.4 Tree structure of the applied time transfer rule

(S	
(NP-CM Τα/AtDfNePlNm Χριστούγεννα/NoPrNePlAc)	(Christmas)
(NP Βροχές/NoCmFePlAc και/CjCo καταιγίδες/NoCmFePlAc)	(rains and thunderstomrs)
(VB θα/PtFu εκδηλωθούν/VbMnIdXx03PlXxPePvXx)	(will occurs)
(NP	
κατά/AsPpSp	(in)
τόπους/NoCmMaPlAc	(local)
στη/AsPpPaFeSgAc	(in)
Δυτική/AjBaFeSgAc	(Western)
Ελλάδα/NoPrFeSgAc)	(Greece)
~~(NP-CM Τα/AtDfNePlNm Χριστούγεννα/NoPrNePlAc)~~	~~(Christmas)~~
)	

Σχήμα 5.5 Εφαρμογή του κανόνα χρονικής μεταφοράς

τα ανοιχτά δεδομένα του Ελληνικού Βικιλεξικού (Greek Wiktionary[1]) και μετατρέποντας τα σώματα σε μορφή σωμάτων κειμένων XCES (Κώδικας 5.2). Τα λήμματα που δεν περιλαμβάνονται στο Ελληνικό Βικιλεξικό έχουν προστεθεί χειρωνακτικά στο λεξικό του συστήματος. Το εργαλείο μας αρχικά μετατρέπει όλες τις λέξεις σε λήμματα και στη συνέχεια τις μεταγράφει με κεφαλαία γράμματα, έτσι ώστε να μπορούν να χρησιμοποιηθούν ως γραπτές μεταγραφές της ΕΝΓ πάντα συνοδευόμενες με τις αρχικές γραμματικές πληροφορίες της λέξης.

Γραπτές μεταγραφές της ΕΝΓ (GSL glosses): Το προτεινόμενο γραπτό σύστημα γραπτών μεταγραφών της ΕΝΓ χρησιμοποιεί μια παραλλαγή του κώδικα κωδικοποίησης του συστήματος Berkeley (Hoiting and Slobin, 2002, Slobin et al., 2001) ως σύστημα μεταγραφής, το οποίο αφαιρεί την κινηματική αναπαράσταση των νοημάτων. Η έξοδος του δικού μας συστήματος μεταγραφής θεωρείται ως ένα ενδιάμεσο αποτέλεσμα από το οποίο τα νοήματα, αφού μεταφραστούν σε κινηματικές μορφές της ΕΝΓ, θα μπορούσαν να αναπαρασταθούν από την τρισδιάστατη τεχνολογία των avatar.

Μορφολογικοί κανόνες (Morphological rules): Η εργαλειοθήκη γλωσσικών εφαρμογών της Python (NLTK) έχει χρησιμοποιηθεί για να υπολογίσει άτυπα στατιστικά στοιχεία των μορφολογικών δεδομένων. Κύριος στόχος μας ήταν να εξάγουμε χρήσιμα συμπεράσματα και να αναπτύξουμε κανόνες μορφολογικής και γραμματικής μεταφοράς. Ο Πίνακας 5.1 παρέχει ένα παράδειγμα στατιστικού δεδομένου με μια λίστα με τις δέκα πιο συχνά εμφανιζόμενες μορφολογικές ετικέτες του προτύπου Parole[2].

Επιπλέον, η στατιστική ανάλυση τύπου Concordance[13] (Sinclair, 1991) έχει ενσωματωθεί για να διερευνήσει τη συχνότητα των συγκεκριμένων λέξεων μέσα σε προτάσεις.

[1] https://el.wiktionary.org/
[2] http://nlp.ilsp.gr/nlp/tagset_examples/
[13] http://www.nltk.org/book/ch01.html (section): 1.3 Searching Text.

POS	Freq	Example	Comment
No	6,191	Βροχές/tag=NoCmFePlAc (rain)	Ουσιαστικό/Noun (No), γένους θηλυκού/feminine (Fe) στον πληθυντικό/plural (Pl)
Aj	3,353	Δυτική/AjBaFeSgAc (Western)	Επίθετο/Adjective(Aj), γένους θηλυκού/feminine (Fe) στον ενικό/in singular (Sg)
As	3,191	στα/ AsPpPaNePlAc (at)	Adposition (= Preposition)[8]
At	2,756	τα/AtDfNePlNm (the)	Άρθρο/Article (At), γένος ουδέτερο/gender neutral (Ne) στον πληθυντικό/plural (Pl)
Vb	1,825	εκδηλωθού-ν/VbMnIdXx03PlXxPePvXx (occurs)	Ρήμα/Verb (Vb), παθητικής φωνής/passive voice (Pv), πληθυντικός/plural (Pl)
DI	1,603	Αριθμός (Digit)	8 beaufort
PT	1,343	Θα/PtFu (will)	Μόριο μέλλοντα/future particle (Fu) (Particles)[9]
Cj	1,299	ενώ/CjCo(while)	Σύζευξη/ Conjunction[10]
Pt	1,251	Να/PtSj (π.χ. να είναι)(to)	Μόριο υποτακτικής κυρίως πριν από ρήματα/ subjunctive particle mainly used before verbs (Sj Subjunctive)[11]
Ad	1,118	κυρίως, σχετικά, έτσι, μόνο, όπου, πιθανόν, αύριο /AdXxBa (mainly, relatively, so, only, where, perhaps, tomorrow)	Επίρρημα/Adverb[12]

Πίνακας 5.1 Εφαρμογή του χρονικού κανόνα μεταφοράς

....νες καταιγίδες . Η θερμοκρασία θα	παρουσιάσει	μικρή περαιτέρω άνοδο ξεπερνώντας
..ταιγίδες DAYMONTHYEAR Μεταβολή θα	παρουσιάσει	ο καιρός από το μεσημέρι της Δευτ..
AYMONTHYEAR Σημαντική μεταβολή θα	παρουσιάσει	ο καιρός από τις πρωινές ώρες της
MONTHYEAR Σημαντική επιδείνωση θα	παρουσιάσει	από αύριο το μεσημέρι ο καιρός με
..εσημέρι της Πέμπτης Επιδείνωση θα	παρουσιάσει	ο καιρός στο μεγαλύτερο μέρος της
..μοκρασίας . Αναλυτικά Μεταβολή θα	παρουσιάσει	ο καιρός από νωρίς το μεσημέρι τη
GIT μποφόρ , ενώ η θερμοκρασία θα	παρουσιάσει	μικρή πτώση κυρίως στη δυτική και
..ράδυ στα βορειοδυτικά Μεταβολή θα	παρουσιάσει	ο καιρός από σήμερα το βράδυ και
..ιαστήματα , ενώ σταδιακή πτώση θα	παρουσιάσει	και η θερμοκρασία από το απόγευμα
θερμοκρασία έως και το Σάββατο θα	παρουσιάσει	σταδιακή πτώση από DIGIT έως και
Δευτέρας . Μεταβολή αναμένεται να	παρουσιάσει	ο καιρός στη χώρα μας από τη Δευτ..
τη χώρα μας . Αρχικά ο καιρός θα	παρουσιάσει	επιδείνωση την Παρασκευή DIGIT /
..τοιχεία , περαιτέρω επιδείνωση θα	παρουσιάσει	ο καιρός την Καθαρά Δευτέρα DIGIT

Σχήμα 5.6 Εφαρμογή του Concordance python για τον ρηματικό τύπο “παρουσιάσει”

Το Σχήμα 5.6 παρέχει ένα παράδειγμα για τον ρηματικό τύπο “παρουσιάσει”:

Source Text	**Morphological Rules' Transfers**	**Final Sentence in GSL glosses**
Τα	Remove “τα”	
Χριστού-	Use gloss “ΧΡΙΣΤΟΥΓΕΝΝΑ”	ΧΡΙΣΤΟΥΓΕΝΝΑ
γεννα		
θα	Remove “θα»	
	Add Predict gloss “ΜΕΤΑ” with MmCs	ΜΕΤΑ/ΧΛ(ΜΕΤΑ)
	“/ΧΛ(ΜΕΤΑ)”	
	Use gloss “ΓΙΝΕΙ” for “ΕΚΔΗΛΩΘΟΥΝ”	ΓΙΝΕΙ
εκδηλω-	Use gloss “ΒΡΟΧΗ”	ΒΡΟΧΗ
θούν		
βροχές	Use gloss “ΚΑΙ”	ΚΑΙ
και καται-	Use gloss “ΚΑΤΑΙΓΙΔΑ” with MmCs	ΚΑΤΑΙΓΙΔΑ-
γίδες	“/ΜΧ(ΕΝΤΑΣΗ)/ΜΓΛ(ΦΟΥΣΚΩΜΕΝΑ)”	/ΜΧ(ΕΝΤΑΣΗ)
		/ΜΓΛ(ΦΟΥΣΚΩΜΕΝΑ)
	Use gloss ANT_3 with MmCs	
	“/ΜΤ(ΑΝΟΙΧΤΑ)”	
	Use twice gloss ΤΟΠΟΣ with MmCs	ΤΟΠΟΣ/ΤΠΘ(Χ1)-
		ΤΟΠΟΣ/ΤΠΘ(Χ2)
κατά	“/ΤΠΘ(Χ1)” and “/ΤΠΘ(Χ2)”	
τόπους	Use gloss ANT_3 with MmCs	ANT_3/ΜΤ(ΑΝΟΙΧΤΑ)
	“/ΜΤ(ΑΝΟΙΧΤΑ)”	
	And Add glosses	ΕΛΛΑΔΑ
	“ΔΥΤΙΚΟΣ”-“ΕΛΛΑΔΑ”	
στη	and SWAP them:	ΔΥΤΙΚΟΣ.
Δυτική	“ΕΛΛΑΔΑ”-“ΔΥΤΙΚΟΣ”	
Ελλάδα		

```
<?xml version='1.0' encoding='UTF-8'?>
<cesDoc version="0.4" xmlns="http://www.xces.org/schema/2003">
 <text>
  <body>
    <p id="p">
      <s id="s0">
        <t id="t0" lemma="τα" tag="AtDfAcNePl" word="τα"/>
        <t id="t1" lemma="Χριστουγεννα" tag="NoAcNePlXx" word="Χριστουγεννα"
   />
        <t id="t2" lemma="θα" tag="Pt" word="τα"/>
```

```
10          <t id="t3" lemma="εκδηλωνω" tag="VbXxPvPlFu" word="εκδηλωθουν"/>
11          <t id="t4" lemma="Βροχη" tag="NoAcFePlXx" word="βροχες"/>
12          <t id="t5" lemma="και" tag="Cj" word="και"/>
13          <t id="t6" lemma="καταιγιδα" tag="NoAcFePlXx" word="καταιγιδες"/>
14          <t id="t7" lemma="κατα" tag="Pp" word="κατα"/><t id="t8" lemma="
    τοπος" tag="NoAcMaPlXx" word="τοπους"/>
15          <t id="t9" lemma=" " tag="AtPpAcFeSg" word="στε"/>
16          <t id="t10" lemma="δυτικος" tag="AjAcFeSgXx" word="Δυτικη"/>
17          <t id="t11" lemma="Ελλαδα" tag="NoAcFeSgXx" word="Ελλαδα"/>
18          <t id="t12" lemma="." tag="PTERM_P" word="."/>
19        </s>
20      </p>
21    </body>
22  </text>
23 </cesDoc>
```

Listing 5.2 XCES format for the example sentence

5.2.3 Κανόνες μεταφοράς σε επίπεδο λέξης (Word Transformation Rule Based)

Στο τελευταίο στάδιο το σύστημα ΜΜ κανόνων συνεχίζει τη διεργασία της μεταφοράς, στο επίπεδο λέξεων πλέον. Για τη διεργασία αυτή χρησιμοποιούμε πρότυπα κανόνων που βασίζονται στη λογική των κανονικών εκφράσεων και στη λογική του finite-State-Automata, χρησιμοποιώντας κώδικα σε επίπεδο script-xml για την εκτέλεση των κανόνων. Η πρόταση εισόδου περνάει μέσα σε μια "μηχανή" ελέγχου συνθηκών, χρησιμοποιώντας πρότυπα συνθηκών, και στη συνέχεια προβαίνει σε ενέργειες, και συγκεκριμένα στις ενέργειες "διαγραφής", "τροποποίησης", και "αντιμετάθεσης".

5.2.4 Η έξοδος του συστήματος ΜΜ κανόνων (RBMT System's Export Stage)

Κατά το στάδιο εξόδου του συστήματος, λαμβάνονται στις εισόδους του τα αποτελέσματα των σταδίων ανάλυσης και μεταφοράς και παράγονται διαφορετικά είδη σωμάτων κειμένων της ΕΝΓ, ακόμα και σε ειδική μορφή κειμένου τύπου XML, για την αξιολόγηση της ΜΜ.

Η γραπτή μεταγραφή της ΕΝΓ γίνεται σε μορφή απλού κειμένου, όπου τα glosses είναι χαρακτηρισμένα (α) με ή χωρίς τις ετικέτες POS (μέρη του λόγου) από το στάδιο της ανάλυσης και (β) με ή χωρίς ετικέτες NMC (μη χειρωνακτικών νοημάτων) από το στά-

Full POS without NmCs	**Simple POS without NmCs**	**Glosses with NmCs without POS**
ΧΡΙΣΤΟΥΓΕΝΝΑ- /NoAcNePlXx ΜΕΤΑ/Pt ΓΙΝΕΙ/Vb ΒΡΟΧΗ/NoAcFePlXx ΚΑΙ/Cj ΚΑΤΑΙΓΙΔΑ- /NoAcFePlXx ANT_3/PreDict ΤΟΠΟΣ/ΤΠΘ(X1)- ΤΟΠΟΣ/ΤΠΘ(X2)/No ANT_3/PreDict ΕΛΛΑΔΑ/NoAcFeSgXx ΔΥΤΙΚΟΣ/AjAcFeSgXx ./PTERM_P	ΧΡΙΣΤΟΥΓΕΝΝΑ/No ΜΕΤΑ/Pt ΓΙΝΕΙ/Vb ΒΡΟΧΗ/No ΚΑΙ/Cj ΚΑΤΑΙΓΙΔΑ/No ANT_3/PreDict ΤΟΠΟΣ/ΤΠΘ(X1)- ΤΟΠΟΣ/ΤΠΘ(X2)/No ANT_3/PreDict ΕΛΛΑΔΑ/No ΔΥΤΙΚΟΣ/Aj ./PTERM_P	ΧΡΙΣΤΟΥΓΕΝΝΑ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΓΙΝΕΙ ΒΡΟΧΗ ΚΑΙ ΚΑΤΑΙΓΙ- ΔΑ/ΜΧ(ΕΝΤΑΣΗ) /ΜΓΛ(ΦΟΥΣΚΩΜΕΝΑ) ANT_3/MT(ΑΝΟΙΧΤΑ) ΤΟΠΟΣ/ΤΠΘ(X1)- ΤΟΠΟΣ/ΤΠΘ(X2) ANT_3/MT(ΑΝΟΙΧΤΑ) ΕΛΛΑΔΑ ΔΥΤΙΚΟΣ.

Πίνακας 5.2 Different kinds of export corpora

διο μεταφοράς, παρέχοντας έτσι τη δυνατότητα να παράγουμε διαφορετικούς τύπους σωμάτων γραπτών μεταγραφών της ΕΝΓ (Πίνακας 5.2).

5.2.5 Χρονικό κόστος του συστήματος RBMT

Για να αξιολογηθεί η ποιότητα του συστήματος, έχει δημιουργηθεί μια διαδικασία δοκιμής και ελέγχου. Πιο συγκεκριμένα από τα 10 υποσώματα, τα 9 (90%) χρησιμοποιούνται για την εκπαίδευση του συστήματος και το 1 (10%) εξαιρείται εντελώς από την όλη διαδικασία της εκπαίδευσης και χρησιμοποιείται αποκλειστικά για τη δοκιμή και αξιολόγηση του συστήματος. Ένας έμπειρος επαγγελματίας μεταφραστής μεταφράζει τα 9 υποσώματα, με τη βοήθεια του προτεινόμενου συστήματος ΜΜ κανόνων με κανόνες μεταφοράς. Το σύστημα τροφοδοτείται συνεχώς με κανόνες μεταφοράς που προκύπτουν από τα γλωσσικά φαινόμενα που διαπιστώνονται από τα 9 υποσώματα. Στις επόμενες παραγράφους παρουσιάζονται τα αποτελέσματα του χρονικού κόστους που απαιτείται από τον μεταφραστή για να μεταφράσει τα σώματα κειμένων και να αναπτύξει τους κανόνες μεταφοράς σε συνάρτηση με τα υποσώματα.

Για το πρώτο υποσώμα και στην περίπτωση απουσίας συστήματος ΜΜ, ο επαγγελματίας μεταφραστής μπορεί να μεταφράζει κατά μέσο όρο 8,2 λέξεις ανά λεπτό. Αυτό σημαίνει ότι για κάθε λέξη έχουμε ρυθμό 0,099 λεπτά. Πάλι για το πρώτο υποσώμα, όταν χρησιμοποιείται το προτεινόμενο σύστημα ΜΜ κανόνων, ο επαγγελματίας μεταφραστής

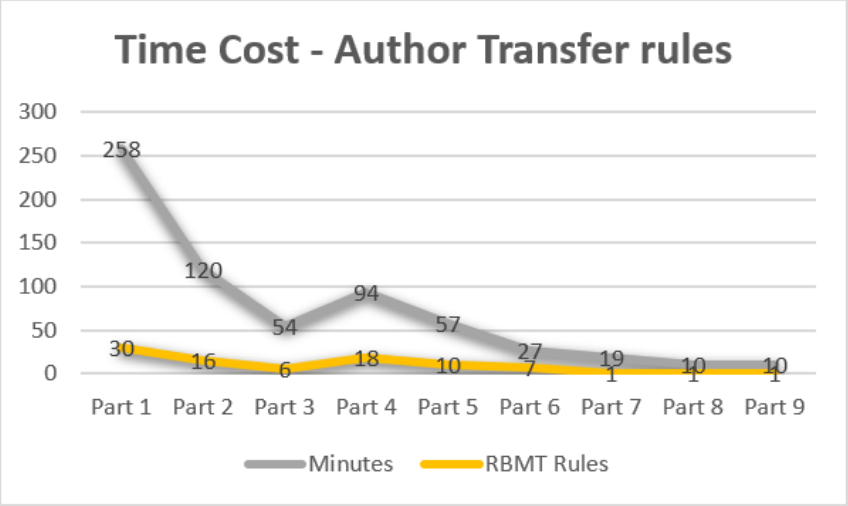
# of Sentence	Total number of words	Words per minute	Minutes per word
1-10	160	8.42	0.12
11-20	229	12.05	0.08
21-30	296	15.58	0.06
31-40	236	15.73	0.06
41-50	278	21.38	0.05
51-60	285	20.36	0.05
61-70	281	14.05	0.07
71-80	252	18.00	0.06
81-90	135	7.94	0.13
91-100	131	6.89	0.15

Πίνακας 5.3 Χρόνος μετάφρασης από σύστημα ΜΜ κανόνων

επιτυγχάνει ένα ρυθμό 14,5 λέξεων ανά λεπτό κατά μέσο όρο (ή 0,06 λεπτά ανά λέξη). Στον Πίνακα 5.3 παρουσιάζεται το μέσο χρονικό κόστος ανά δεκάδα προτάσεων για τις πρώτες 10 δεκάδες, όταν χρησιμοποιείται το προτεινόμενο σύστημα ΜΜ κανόνων.

Όπως μπορείτε να παρατηρήσετε, υπάρχει μια σημαντική βελτίωση του κόστους χρόνου μετάφρασης κατά περίπου 42,86%, όταν το προτεινόμενο σύστημα ΜΜ κανόνων ολοκληρώσει την εκπαίδευση. Συνολικά, 23.192 λέξεις έχουν μεταφραστεί (μαζί με τις ετικέτες NmCs) σε περίπου 1.600 λεπτά (26.6 ώρες). Όταν το προτεινόμενο σύστημα RBMT δεν χρησιμοποιείται, ο μεταφραστής χρειάζεται 2.800 λεπτά (46.6 ώρες).

Στο Σχήμα 5.7 παρουσιάζεται το κόστος χρόνου που απαιτείται για τη δημιουργία των κανόνων μεταφοράς για τα 9 υποσώματα εκπαίδευσης. Όπως μπορείτε να παρατηρήσετε, για το πρώτο υποσώμα χρειάστηκαν 258 λεπτά για να γίνει η εξαγωγή των γλωσσικών φαινομένων και στη συνέχεια να υλοποιηθούν οι κανόνες μεταφοράς για το σύστημα μας. Αυτό το κόστος χρόνου μειώνεται στο δεύτερο υποσώμα, όπου βλέπουμε ότι χρειάστηκαν 120 λεπτά για τον σκοπό αυτό. Από το έκτο υποσώμα και μετά, παρατηρούμε ότι το κόστος χρόνου εξαγωγής γλωσσικών φαινομένων και δημιουργίας κανόνων μεταφοράς μειώνεται δραστικά και αυτό εξηγείται με το γεγονός ότι τα περισσότερα γλωσσικά φαινόμενα έχουν ήδη εξαντληθεί από τα προηγούμενα υποσώματα. Ωστόσο, αφιερώνεται αρκετός χρόνος για μικρές διορθώσεις, τροποποιήσεις και σχετικές αντιστοιχίσεις ζευγαριών λέξεων, όπου απαιτείται. Εξωτερικά εργαλεία λογισμικού έχουν χρησιμοποιηθεί ως βοηθήματα με σκοπό τη μελέτη και εξαγωγή γραμματικών/γλωσσικών φαινομένων των σωμάτων μας, όπως για παράδειγμα το στατιστικό γλωσσικό εργαλείο Concordance, με σκοπό τη στατιστική ανάλυση των λέξεων μέσα σε προτάσεις.



Σχήμα 5.7 Χρονικό κόστος (σε λεπτά) για την ανάπτυξη κανόνων μεταφοράς από τα 9 υποσύνολα

Κεφάλαιο 6

Chunking with Regular Expressions

Για τις ανάγκες της παρούσας διατριβής χρησιμοποιήθηκαν τα εργαλεία RegexpChunk και RegexpParser από τη σουίτα NLP της Python, για τη δημιουργία προτύπων ετικετών (tag pattern) και κανόνων αναλυτή φράσεων (chunk rule), έτσι ώστε να κατασκευάσουμε έναν αναλυτή φράσεων.

6.1 Εφαρμογή ChunkRule και RegexpChunk

Το εργαλείο αναλυτή φράσεων "ChunkRule" της σουίτας NLP, χρησιμοποιεί πιο απλούς κανόνες για τον τεμαχισμό φράσεων, σε αντίθεση με το "ChunkParser". Ο αναλυτής αυτός χωρίζει την πρόταση σε φράσεις σύμφωνα με τα πρότυπα ετικετών. Ο "ChunkRule" δέχεται ως παραμέτρους εισόδου ένα πρότυπο ετικετών μαζί με μια περιγραφή. Παρακάτω βλέπουμε ένα παράδειγμα προτύπου το οποίο τεμαχίζει σε τμήματα ακολουθίες που αποτελούνται από μία ή περισσότερες λέξεις, χαρακτηρισμένες ως "At" ή "No" (δηλαδή άρθρα ή ουσιαστικά).

```
>>> rule = parse.ChunkRule('<At|No>+', ... 'Chunk sequences of DT and NN')
```

Τώρα μπορούμε να ορίσουμε τον αναλυτή κανονικών εκφράσεων με αυτό το πρότυπο ή κανόνα:

```
>>> chunkparser = parse.RegexpChunk([rule], chunk]_node='NP', top_node='S')
```

Ο "RegexpChunk" δέχεται, προαιρετικά, δεύτερη και τρίτη παράμετρο, οι οποίες προσδιορίζουν τις ετικέτες των κεφαλών των τεμαχίων-φράσεων, καθώς και την ονομασία της πρώτης κεφαλής.

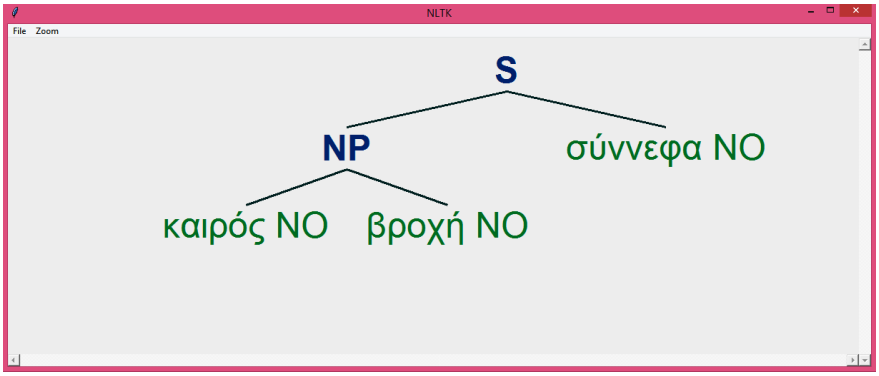
Μπορούμε, επίσης, να χρησιμοποιήσουμε πιο πολύπλοκα σχήματα ετικετών, όπως <At> ? <Aj.*> * <No.*>. Η πρακτική αυτή μπορεί να χρησιμοποιηθεί για τον τεμαχισμό οποιασδήποτε ακολουθίας ετικετών των λεκτικών μονάδων που αρχίζει με έναν προαιρε-

τικό προσδιοριστή “At”, που ακολουθείται από μηδέν ή περισσότερα επίθετα οποιουδήποτε τύπου “Aj.*” που ακολουθούνται από ουσιαστικό οποιουδήποτε τύπου “No.*”.

Εάν ένα πρότυπο ετικέτας αντιστοιχεί σε πολλαπλές θέσεις που επικαλύπτονται, τότε το πρώτο ταίριασμα προηγείται. Για παράδειγμα εάν εφαρμόσουμε έναν κανόνα σύμφωνα με τον οποίο θέλουμε να τεμαχίσουμε δύο συνεχόμενα ουσιαστικά και ο κανόνας αυτός τρέξει σε μια πρόταση που περιέχει τρία συνεχόμενα ουσιαστικά, τότε θα τεμαχίσει μόνο τα δύο πρώτα ουσιαστικά. Δείτε στο Σχήμα 6.1 τον κώδικα διπλού ταιριάσματος ουσιαστικού και στο Σχήμα 6.2 τη γραφική έξοδο με τη μορφή δέντρου.

```
from nltk.chunk.regexp import ChunkRule
from nltk import tag
from nltk.chunk import RegexpChunkParser
rule = ChunkRule('<NO><NO>', 'Τεμάχισε δυο συνεχόμενα
ουσιαστικά') #'Chunk two consecutive nouns'
text = "καιρός/NO βροχή/NO σύννεφα/NO"
sent = []
for t in text.split(): sent.append(tag.str2tuple(t))
print(sent)
chunker = RegexpChunkParser([rule])
print(chunker.parse(sent))
chunker.parse(sent).draw()
-----------------------
Εξοδος συστήματος :
-----------------------
(S: (NP: ('καιρός', 'No') ('βροχή', 'No')) ('σύννεφα', 'No'))
```

Σχήμα 6.1 Κώδικας κανόνα αναλυτή δύο συνεχόμενων ουσιαστικών



Σχήμα 6.2 Κανόνας αναλυτή δύο συνεχόμενων ουσιαστικών

Chinking			
Είσοδος	[Μια/At κρύα/Aj μέρα/No]	[Μια/At κρύα/Aj μέρα/No]	[Μια/At κρύα/Aj μέρα/No]
Κανόνας	[Chink "Μια/at κρύα/Aj μέρα/No"	Chink "κρύα/Aj"	Chink "μέρα/No"
Έξοδος	Μια/At κρύα/Aj μέρα/No	[Μια/At] κρύα/Aj μέρα/No	[Μια/At κρύα/Aj] μέρα/No

Πίνακας 6.1 Chinking

6.2 Κανόνας Chink Rule

Chinking είναι η διαδικασία αφαίρεσης μιας αλληλουχίας λεκτικών μονάδων από ένα τεμαχισμένο κείμενο (chunked text). Αν η αλληλουχία των λεκτικών μονάδων εκτείνεται σε μια ολόκληρη υποφράση (chunk), τότε ολόκληρη η υποφράση αφαιρείται. Αν η αλληλουχία των μονάδων εμφανίζεται στη μέση της υποφράσης, αυτές οι μονάδες (tokens) απομακρύνονται, αφήνοντας δύο νέες υποφράσεις εκεί που υπήρχε μόνο μία πριν. Αν η αλληλουχία είναι στην αρχή ή στο τέλος της υποφράσης (chunk), αυτές οι λεκτικές μονάδες απομακρύνονται και μια μικρότερη υποφράση παραμένει. Αυτές οι τρεις περιπτώσεις απεικονίζονται στον Πίνακα 6.1.

Ένας κανόνας ChinkRule κάνει "chink" οτιδήποτε ταιριάζει σε ένα συγκεκριμένο πρότυπο ετικέτας. Ο κανόνας ChinkRule δημιουργείται με τη βιβλιοθήκη του ChinkRule, η οποία λαμβάνει ένα πρότυπο ετικέτας και μια περιγραφή. Για παράδειγμα, ο κανόνας του Σχήματος 6.3 κάνει "chink" οποιαδήποτε ακολουθία των λεκτικών μονάδων των οποίων οι ετικέτες είναι "VB" ή "IN".

Ο αναλυτής RegexpChunk ξεκινά τη διαδικασία τεμαχισμού φράσεων με την προϋπόθεση ότι τίποτα δεν είναι κατατμημένο. Έτσι, πριν εφαρμόσει τον κανόνα του "chink", θα εφαρμόσει έναν άλλο κανόνα, τον "chunk all" (Σχήμα 6.4), ο οποίος θέτει το σύνολο της πρότασης σε μια ενιαία τεμαχισμένη υποφράση.

```
>>> chink_rule = parse.ChinkRule('<VBD|IN>+',
... 'Chink sequences of VBD and IN')
```

Σχήμα 6.3 Παράδειγμα κανόνα chinkrule "VBD" ή "IN"

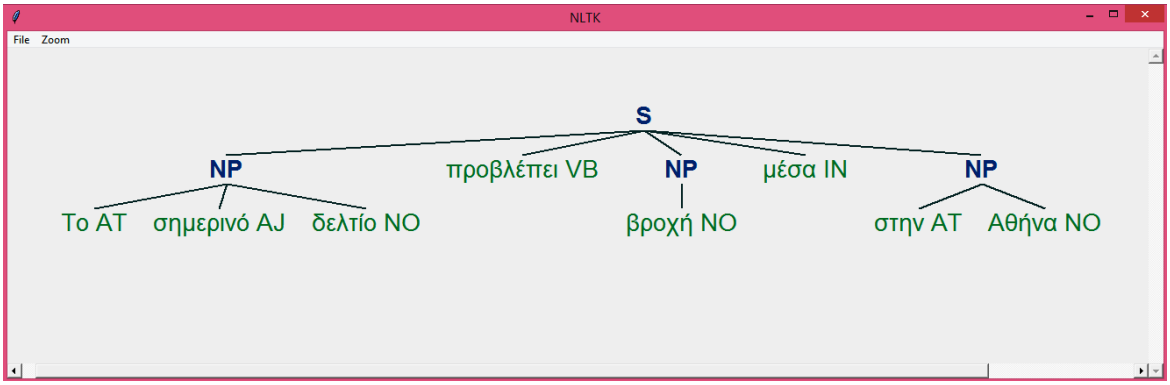
```
>>> chunkall_rule = parse.ChunkRule('<.*>+',
... 'Chunk everything')
```

Σχήμα 6.4 Κανόνας τεμαχισμού όλης φράσης (chunkall)

Τέλος, μπορούμε να συνδυάσουμε αυτούς τους δύο κανόνες στο παρακάτω παράδειγμα (Σχήμα 6.5) για τη δημιουργία ενός αναλυτή υποφράσεων.

```
from nltk.chunk.regexp import ChunkRule
from nltk.chunk.regexp import ChinkRule
from nltk.chunk import RegexpChunkParser
from nltk import tag
text = "Το/AT σημερινό/AJ δελτίο/NO προβλέπει/VB βροχή/NO μέσα/IN
στην/AT Αθήνα/NO"
sent = []
for t in text.split(): sent.append(tag.str2tuple(t))
print('sent:',sent)
chink_rule = ChinkRule("<VB|IN>+", "Chink sequences of VBD and
IN")
chunkall_rule = ChunkRule("<.*>+","Chunk everything")
chunkparser = RegexpChunkParser([chunkall_rule, chink_rule],
chunk_label="NP", root_label="S")
chunk_tree = chunkparser.parse(sent, trace=1)
chunk_tree.draw()
Έξοδος συστήματος :
sent: [('Το', 'AT'), ('σημερινό', 'AJ'), ('δελτίο', 'NO'),
('προβλέπει', 'VB'), ('βροχή', 'NO'), ('μέσα', 'IN'), ('στην',
'AT'), ('Αθήνα', 'NO')]
# Input:
 <AT>  <AJ>  <NO>  <VB>  <NO>  <IN>  <AT>  <NO>
# Chunk everything:
{<AT>  <AJ>  <NO>  <VB>  <NO>  <IN>  <AT>  <NO>}
# Chink sequences of VBD and IN:
{<AT>  <AJ>  <NO>} <VB> {<NO>} <IN> {<AT>  <NO>}
```

Σχήμα 6.5 Παράδειγμα κανόνα chunkall και Chinkrule μαζί με έξοδο αποτελεσμάτων



Σχήμα 6.6 Έξοδος αποτελεσμάτων του παραδείγματος της Εικόνας 7 υπό μορφή γραφικού δέντρου

Αν ένα πρότυπο ετικέτας (κανόνας) βρίσκει εφαρμογή σε περισσότερες από μία περιπτώσεις, τότε εφαρμόζεται στην πρώτη περίπτωση ταιριάσματος.

6.3 Κανόνας Unchunk Rule

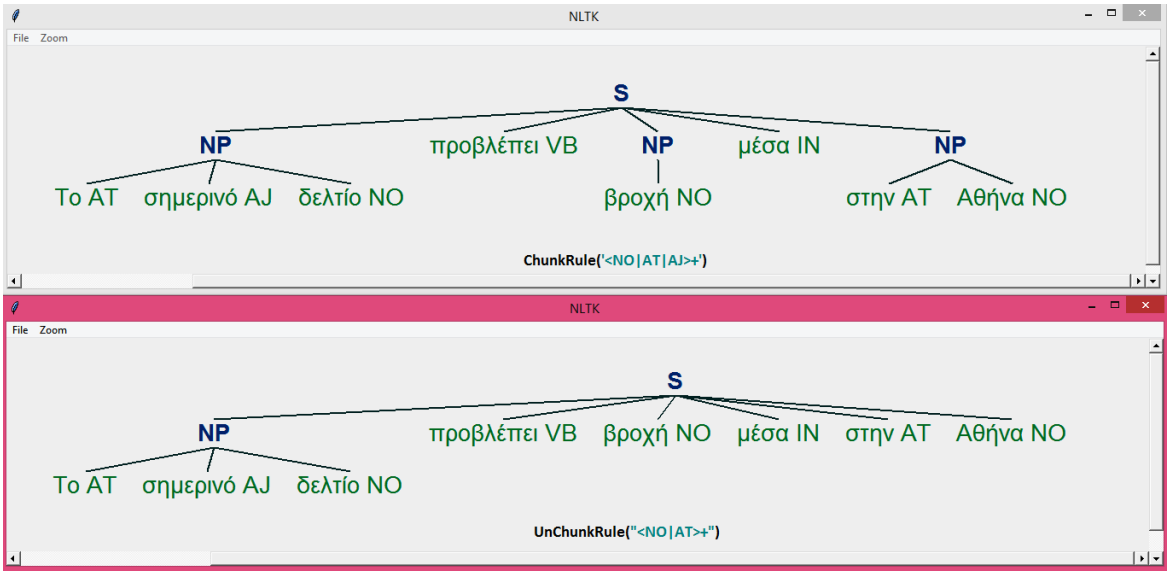
Ένας κανόνας "UnChunkRule" αφαιρεί κάθε χαρακτηρισμένη υποφράση (chunk) που ταιριάζει σε ένα συγκεκριμένο πρότυπο ετικέτας. Ο κανόνας "UnChunkRule" μοιάζει πολύ με τον κανόνα "ChinkRule", με τη διαφορά ότι όταν το πρότυπο ετικέτας ταιριάξει σε ολόκληρη την τεμαχισμένη φράση, ο κανόνας "UnChunkRule" θα αφαιρέσει μόνο ένα τμήμα (υποφράση). Σε αντίθεση, με τον κανόνα "ChinkRule" που μπορεί να αφαιρέσει ακολουθίες λεκτικών μονάδων από τη μέση μιας υποφράσης (chunk).

Στα Σχήματα 6.7 και 6.8 βλέπουμε ένα παράδειγμα κανόνα "UnChunkRule" μαζί με την έξοδο αποτελεσμάτων και το γραφικό δέντρο της τεμαχισμένης πρότασης σε υποφράσεις.

```
from nltk.chunk.regexp import ChunkRule
from nltk.chunk.regexp import ChinkRule
from nltk.chunk import RegexpChunkParser
from nltk.chunk.regexp import UnChunkRule
from nltk import tag
text = "Το/AT σημερινό/AJ δελτίο/NO προβλέπει/VB βροχή/NO μέσα/IN
στην/AT Αθήνα/NO"
sent = []
for t in text.split(): sent.append(tag.str2tuple(t))
print('sent:',sent)
unchunk_rule = UnChunkRule("<NO|AT>+","Unchunk sequences of NO and
AT")
chunk_rule = ChunkRule('<NO|AT|AJ>+', 'Τεμαχισμός αλληλουχίας
No,Aj και At')
chunkparser = RegexpChunkParser([chunk_rule, unchunk_rule],
chunk_label='NP', root_label='S')
chunk_tree = chunkparser.parse(sent, trace=1)
chunk_tree.draw()
Έξοδος συστήματος :
sent: [('Το', 'AT'), ('σημερινό', 'AJ'), ('δελτίο', 'NO'),
('προβλέπει', 'VB'), ('βροχή', 'NO'), ('μέσα', 'IN'), ('στην',
'AT'), ('Αθήνα', 'NO')]
# Input:
 <AT>  <AJ>  <NO>  <VB>  <NO>  <IN>  <AT>  <NO>
# Τεμαχισμός αλληλουχίας No,Aj και At:
{<AT>  <AJ>  <NO>} <VB> {<NO>} <IN> {<AT>  <NO>}
# Unchunk sequences of NO and AT:
{<AT>  <AJ>  <NO>} <VB>  <NO>  <IN>  <AT>  <NO>
```

Σχήμα 6.7 Παράδειγμα κανόνα "UnChunk" μαζί με έξοδο αποτελεσμάτων

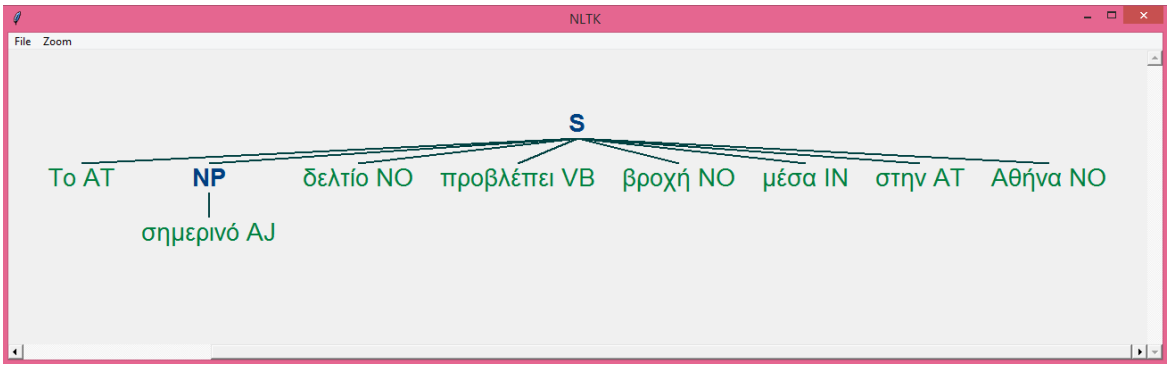
Μπορούμε να δούμε καλύτερα τη διαφορά μεταξύ των κανόνων "Chinkrule" και "UnChunkrule" αν ανατρέξουμε στο παρακάτω παράδειγμα, με το ίδιο πρότυπο ετικέτας του (<'NO|AT>+') στον κανόνα "ChinkRule" (<'NO|AT>+') (Σχήματα 6.9 και 6.10).



Σχήμα 6.8 Γραφική έξοδος σε μορφή δέντρου των αποτελεσμάτων του κανόνα “UnChunk”

```
...
chunk_rule = ChunkRule('<NO|AT|AJ>+', 'Τεμαχισμός αλληλουχίας
No,Aj και At')
chink_rule = ChinkRule('<NO|AT>+','Chink sequences of NO and AT')
chunkparser = RegexpChunkParser([chunk_rule, chink_rule],
chunk_label='NP', root_label='S')
chunk_tree = chunkparser.parse(sent, trace=1)
chunk_tree.draw()
Έξοδος συστήματος :
sent: [('Το', 'AT'), ('σημερινό', 'AJ'), ('δελτίο', 'NO'),
('προβλέπει', 'VB'), ('βροχή', 'NO'), ('μέσα', 'IN'), ('στην',
'AT'), ('Αθήνα', 'NO')]
# Input:
 <AT>  <AJ>  <NO>  <VB>  <NO>  <IN>  <AT>  <NO>
# Τεμαχισμός αλληλουχίας No,Aj και At:
{<AT>  <AJ>  <NO>} <VB> {<NO>} <IN> {<AT>  <NO>}
# Chink sequences of NO and AT:
 <AT> {<AJ>} <NO>  <VB>  <NO>  <IN>  <AT>  <NO>
```

Σχήμα 6.9 Παράδειγμα κανόνα Chinkrule (’<NO|AT>+’)



Σχήμα 6.10 Γραφικό δέντρο του κανόνα Chinkrule ('<NO|AT>+')

6.4 Κανόνας Merge Rule

Όταν κατασκευάζουμε έναν πολύπλοκο αναλυτή υποφράσεων (chunk parser), είναι συνήθως βολικό ο αναλυτής αυτός να "τρέχει" και άλλες εργασίες εκτός από τον κανονικό τεμαχισμό φράσεων (chunking), όπως τις εργασίες "chinking" και "unchunking" που περιγράψαμε προηγουμένως. Σε αυτή την ενότητα και στην επόμενη, θα μιλήσουμε για δύο πιο σύνθετους κανόνες που μπορούν να χρησιμοποιηθούν για τη συγχώνευση (merge) και τη διάσπαση (split) των υποφράσεων.

Οι κανόνες συγχώνευσης "MergeRules" χρησιμοποιούνται για να ενώσουν δύο συνεχόμενες κατατετμημένες φράσεις (υποφράσεις-chunk). Κάθε κανόνας συγχώνευσης αποτελείται από δύο πρότυπα ετικέτας, ένα αριστερό και ένα δεξί. Ο κανόνας συγχώνευσης ενώνει δύο συνεχόμενες υποφράσεις, C1 και C2, όταν το τέλος της C1 ταιριάζει στο αριστερό πρότυπο και η αρχή της C2 στο δεξί. Μόνο τότε έχουμε συγχώνευση των υποφράσεων C1 και C2 σε μια ενιαία υποφράση. Για παράδειγμα, ας δούμε την παρακάτω υπόθεση ανάλυσης υποφράσεων.

[Το/AT σημερινό/AJ] [δελτίο/NO]

Όπου η υποφράση C1 είναι η [Το/AT σημερινό/AJ] και η C2 είναι η [δελτίο/NO]. Αν το αριστερό πρότυπο είναι "AJ", και το δεξί "NO", τότε οι υποφράσεις C1 και C2 πρόκειται να συγχωνευτούν σε μια ενιαία υποφράση:

[Το/AT σημερινό/AJ δελτίο/NO].

Οι κανόνες συγχώνευσης δημιουργούνται από τη ρουτίνα MergeRule της βιβλιοθήκης NTLK-Python, η οποία δέχεται ένα αριστερό και ένα δεξί πρότυπο ετικέτας, καθώς και μια περιγραφή του κανόνα υπό μορφή συμβολοσειράς κειμένου. Για παράδειγμα, ο παρακάτω κανόνας (Σχήμα 6.11 πρόκειται να συγχωνεύσει δύο συνεχόμενες υποφράσεις όταν η πρώτη τελειώνει με "AT", "AJ", "NO" ή "IN" και η δεύτερη με "AJ", "IN", "AT" ή "NO".

```
merge_rule = MergeRule("<AT|AJ|NO|IN>", "<AJ|IN|AT|NO>",
... "Merge NOs + ATs + AJs + INs")
```

Σχήμα 6.11 Κανόνας συγχώνευσης "MergeRule"

Για να δούμε μια εφαρμογή αυτού του κανόνα, θα συνδυάσουμε τον προηγούμενο κανόνα "MergeRule" με έναν κανόνα "Chunkrule", ο οποίος θα τεμαχίζει όλες τις λεκτικές μονάδες σε τεμάχια υποφράσεων ανεξαρτήτως ετικέτας ("<.*>") και έναν κανόνα "UnChunkRule" με πρότυπο ετικέτας που θα αποχαρακτηρίζει τις υποφράσεις ρήματος ή επιρρήματος ("IN|VB.*>"). Βλέπουμε τον κώδικα των κανόνων στο Σχήμα 6.12, την έξοδο στο Σχήμα 6.13 και την έξοδο σε μορφή γραφικού δέντρου στο Σχήμα 6.14.

```
chunk_rule = ChunkRule("<.*>", "Chunk all individual tokens")
unchunk_rule = UnChunkRule("<IN|VB.*>", "Unchunk VBs and INs")
chunkparser = RegexpChunkParser([chunk_rule,
unchunk_rule,merge_rule], chunk_label='NP', root_label='S')
print('sent:',sent)
chunk_tree = chunkparser.parse(sent, trace=1)
chunk_tree.draw()
```

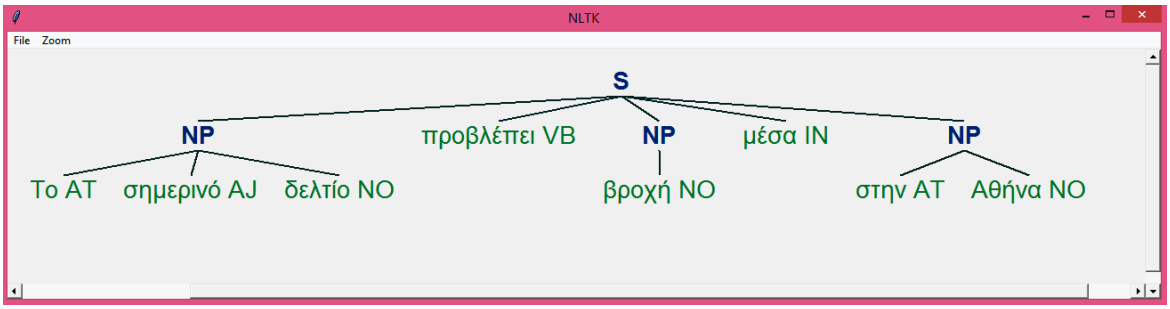
Σχήμα 6.12 Παράδειγμα συνδυασμού κανόνα συγχώνευσης "MergeRule" με "Chunkrule" και "UnChunkRule"

```
Έξοδος συστήματος :
sent: [('Το', 'AT'), ('σημερινό', 'AJ'), ('δελτίο', 'NO'),
('προβλέπει', 'VB'), ('βροχή', 'NO'), ('μέσα', 'IN'), ('στην',
'AT'), ('Αθήνα', 'NO')]
# Input:
 <AT>  <AJ>  <NO>  <VB>  <NO>  <IN>  <AT>  <NO>
# Chunk all individual tokens:
{<AT>}{<AJ>}{<NO>}{<VB>}{<NO>}{<IN>}{<AT>}{<NO>}
# Unchunk VBs and INs:
{<AT>}{<AJ>}{<NO>} <VB> {<NO>} <IN> {<AT>}{<NO>}
# Merge NOs + ATs + AJs + INs:
{<AT>  <AJ>  <NO>} <VB> {<NO>} <IN> {<AT>  <NO>}
```

Σχήμα 6.13 Έξοδος συστήματος του παραδείγματος

6.5 Κανόνας διαχωρισμού Split Rule

Οι κανόνες διαχωρισμού "SplitRules" χρησιμοποιούνται για να χωρίσουν μια χαρακτηρισμένη υποφράση (chunked) σε δύο μικρότερες υποφράσεις (chunks). Κάθε κανόνας



Σχήμα 6.14 Έξοδος συστήματος του παραδείγματος υπο μορφή δέντρου

διαχωρισμού (SplitRule) παραμετροποιείται από δύο πρότυπα ετικέτας, ένα αριστερό και ένα δεξί. Ένας κανόνας "SplitRule" θα χωρίσει μια υποφράση σε οποιοδήποτε σημείο "P", όπου το αριστερό πρότυπο ταιριάζει με την υποφράση προς τα αριστερά του "P" και το δεξί πρότυπο ταιριάζει με την υποφράση προς τα δεξιά του "P". Για παράδειγμα, μελετήστε την παρακάτω υπόθεση διαχωρισμού υποφράσεων:

Έστω ότι έχουμε την εξής ενιαία υποφράση:

[Το/AT σημερινό/AJ δελτίο/NO του/AT καιρού/NO]

Αν το αριστερό πρότυπο ετικέτας είναι "NO", και το δεξί είναι "AT", τότε η υποφράση θα χωριστεί μεταξύ των λεκτικών μονάδων "δελτίο" και "του", για να σχηματίσουν στη συνέχεια δύο μικρότερες υποφράσεις (chunks):

[Το/AT σημερινό/AJ δελτίο/NO] [του/AT καιρού/NO]

Ο κανόνας διαχωρισμού "SplitRule" δημιουργείται από τη βιβλιοθήκη εργαλείων NLTP-Python του κανόνα "SplitRule". Η δομή της ρουτίνας αυτής λαμβάνει ένα αριστερό πρότυπο ετικέτας, ένα δεξί πρότυπο ετικέτας και μια συμβολοσειρά για την περιγραφή του κανόνα. Για παράδειγμα, ο παρακάτω κανόνας θα χωρίσει κάθε υποφράση στο σημείο που έχει ετικέτα "NO" από τα αριστερά και "AT" από τα δεξιά:

split_rule = SplitRule(**"<NO>", "<AT>", "Split NO followed by AT"**)

Για να δούμε καλύτερα τη λειτουργία του παραπάνω κανόνα (SplitRule) θα τον χρησιμοποιήσουμε σε συνδυασμό με έναν κανόνα τεμαχισμού (Chunkrule), που θα τεμαχίζει σε υποφράσεις ουσιαστικών (NO), επιθέτων (AJ) και άρθρων (AT):

Παρακάτω βλέπουμε τον κώδικα "SplitRule" στο Σχήμα 6.15, την έξοδο του κώδικα στο Σχήμα 6.16 και την έξοδο σε μορφή γραφικού δέντρου στο Σχήμα 6.16.

6.6 Εφαρμογές Tree και chunkString

Η γλώσσα προγραμματισμού Python χρησιμοποιεί δύο εφαρμογές για τον χαρακτηρισμό υποφράσεων, την "ChunkString" και την "Tree", καθώς και τα αντίστοιχα εργαλεία

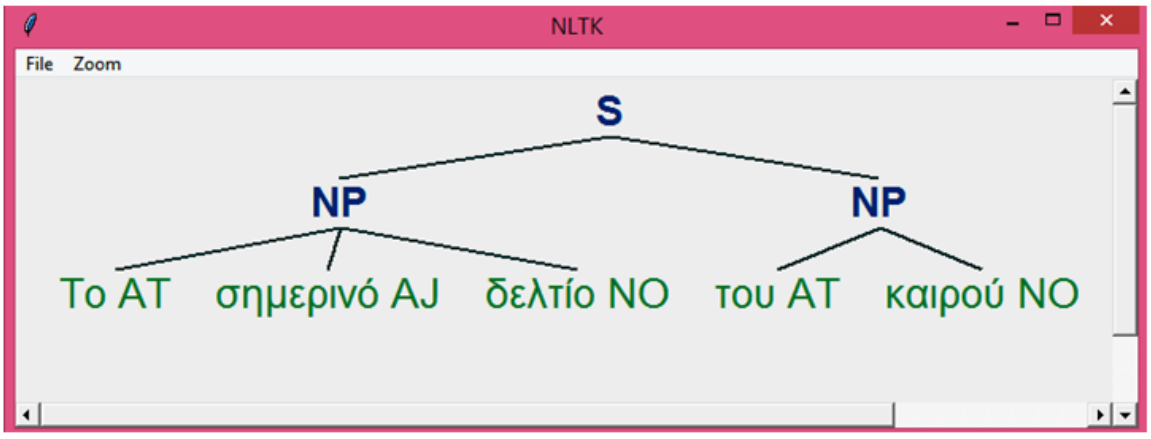
```
text = "Το/AT σημερινό/AJ δελτίο/NO του/AT καιρού/NO"
sent: [('Το', 'AT'), ('σημερινό', 'AJ'), ('δελτίο', 'NO'), ('του', 'AT'), ('καιρού', 'NO')]

chunk_rule = ChunkRule("<NO.*|AT|AJ>+", "Chunk sequences of NO,
AJ, and AT")
split_rule = SplitRule("<NO>", "<AT>", "Split NO followed by AT")
chunkparser = RegexpChunkParser([chunk_rule, split_rule],
chunk_label='NP', root_label='S')
chunk_tree = chunkparser.parse(sent, trace=1)
chunk_tree.draw()
```

Σχήμα 6.15 Επίδειξη κανόνα “SplitRule”

Έξοδος συστήματος:

```
sent: [('Το', 'AT'), ('σημερινό', 'AJ'), ('δελτίο', 'NO'), ('του', 'AT'), ('καιρού', 'NO')]
# Input:
 <AT> <AJ> <NO> <AT> <NO>
# Chunk sequences of NO, AJ, and AT:
{<AT> <AJ> <NO> <AT> <NO>}
# Split NO followed by AT:
{<AT> <AJ> <NO>}{<AT> <NO>}
```

Σχήμα 6.16 Έξοδος κώδικα κανόνα “SplitRule”



Σχήμα 6.17 Γραφικό δέντρο εξόδου κανόνα “SplitRule”

για τη μετατροπή μεταξύ αυτών των εφαρμογών. Συγκεκριμένα η εφαρμογή “ChunkString” μετατρέπει αντικείμενα μορφής δέντρου και μορφής συμβολοσειράς, οριοθετημένα με αγκύλες “” και έπειτα μέσω της εφαρμογής “to_chunkstruct” μετατρέπει το “chunkstring” σε αντικείμενο μορφής δέντρου (Σχήμα 6.10).

```
>>> from nltk.chunk.regexp import ChunkString, ChunkRule,
ChinkRule
>>> from nltk.tree import Tree
>>> t = Tree('S', [('Ο', 'AT'), ('καιρός', 'NO'), ('προβλέπεται', 'VB'),
('πολύ', 'AJ'), ('βροχερός', 'NO')])
>>> cstring = ChunkString(t)
>>>print(cstring)

<ChunkString: ' <AT>  <NO>  <VB>  <AJ>  <NO> '>

>>>  ur = ChunkRule('<AT><NO.*><.*>*<NO.*>', 'chunk
determiners and nouns')

      ur.apply(cstring)
>>> print(cstring)

<ChunkString: '{<AT><NO><VB><AJ><NO>}'>

>>> ir = ChinkRule('<VB.*>', 'chink verbs')
>>> ir.apply(cstring)
>>> print(cstring)

<ChunkString: '{<AT><NO>}<VB>{<AJ><NO>}'>

>>> cstring.to_chunkstruct()

Tree('S', [Tree('CHUNK', [('Ο', 'DT'), ('καιρός', 'NO')]), ('προβλέπεται',
'VB'), Tree('CHUNK', [('πολύ', 'AJ'), ('βροχερός', 'NO')])])
```

Σχήμα 6.18 Παράδειγμα εφαρμογής “ChunkString” και “chunkstruct”

6.7 Κανόνας επέκτασης και αφαίρεσης φράσεων με κανονικές εκφράσεις

Υπάρχουν τρεις υποκατηγορίες κλάσης εφαρμογής “RegexpChunkRule” που δεν υποστηρίζονται από τις κλάσεις “RegexpChunkRule.fromstring” ή “RegexpParser”. Ως εκ

τούτου, θα πρέπει να δημιουργηθούν χειρονακτικά, αν θέλουμε να τις χρησιμοποιήσουμε. Οι υποκατηγορίες αυτές είναι οι εξής:

- **ExpandLeftRule** (Επέκταση αριστερού κανόνα): ο κανόνας αυτός προσθέτει αποχαρακτηρισμένη λεκτική μονάδα ή ζεύγος μονάδων (add unchunked/chink) στο αριστερό μέρος μιας χαρακτηρισμένης φράσης (chunk).

- **ExpandRightRule** (Επέκταση δεξιού κανόνα): ο κανόνας αυτός προσθέτει μη χαρακτηρισμένες λεκτικές μονάδες στο δεξί μέρος μια χαρακτηρισμένης φράσης (chunk).

- **UnChunkRule** (Κανόνας αποχαρακτηρισμού φράσης): ο κανόνας αυτός αποχαρακτηρίζει τις επισημειωμένες υποφράσεις (chunk) σε φράση, εφόσον ικανοποιούν τη συνθήκη του προτύπου ετικέτας (tag pattern).

Οι κανόνες "ExpandLeftRule" και "ExpandRightRule" δέχονται και οι δυο ως παραμέτρους εισόδου πρότυπα ετικέτας μαζί με μια περιγραφή μορφής κειμένου. Στον κανόνα "ExpandLeftRule", το πρώτο πρότυπο που θέλουμε να προσθέσουμε στην αρχή μια χαρακτηρισμένης φράσης (chunk) είναι το chink (αποχαρακτηρισμένη μονάδα), καθώς το δεξί πρότυπο πρόκειται να βρει τα "ταίρια" του στην αρχή της χαρακτηρισμένης φράσης (chunk) που θέλουμε να επεκτείνουμε. Στον κανόνα "ExpandRightRule", το πρότυπο πρόκειται να βρει τα "ταίρια" του στο τέλος της φράσης που θέλουμε να επεκτείνουμε. Η ιδέα είναι ίδια με τον κανόνα "MergeRule class", με τη διαφορά ότι στην περίπτωση αυτή συγχωνεύουμε μόνο μη χαρακτηρισμένες μονάδες (chink words) και όχι χαρακτηρισμένες μονάδες. Ο κανόνας "UnChunkRule" είναι ο αντίθετος του "ChunkRule". Κάθε μονάδα ή συνδυασμός μονάδων που ταιριάζει στο πρότυπο του κανόνα "UnChunkRule" πρόκειται να αποχαρακτηριστεί και να μετατραπεί σε ένα "chink". Στο Σχήμα 6.24 παρουσιάζουμε έναν κώδικα επίδειξης της χρήσης των παραπάνω κανόνων της κλάσης εφαρμογής "RegexpChunkParser class":

Ας δούμε βήμα-βήμα την εκτέλεση των κανόνων που περιγράψαμε παραπάνω:

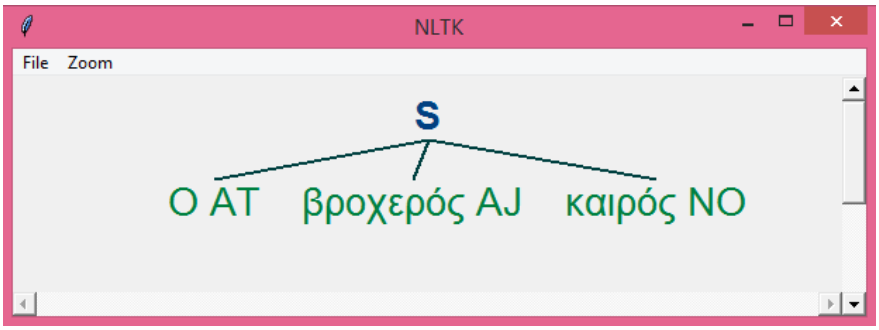
1. Δημιουργία μονού επιθέτου (single Adjective-AJ) σε φράση (Σχήμα 6.20).

2. Επέκταση προς τα αριστερά, ενσωματώνοντας μονάδες με ετικέτες άρθρου (AJ) σε επιθετικές φράσεις (AJ) (Σχήμα 6.21)

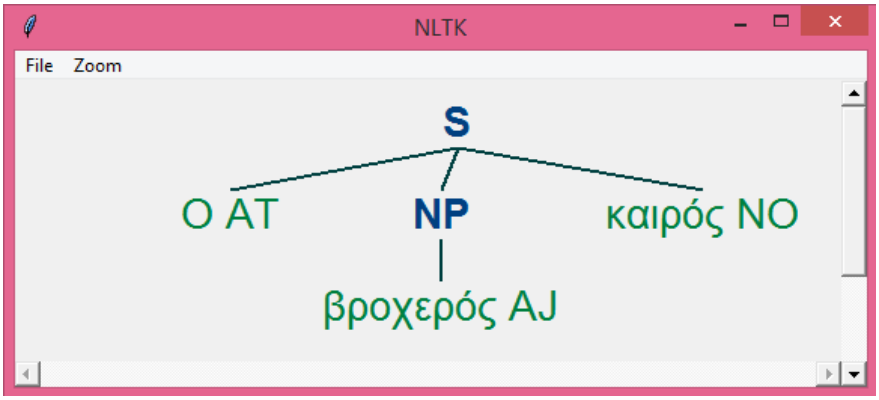
3. Επέκταση προς τα δεξιά, ενσωματώνοντας μονάδες με ετικέτες ουσιαστικού (NO) σε επιθετικές φράσεις (AJ) (Σχήμα 6.22)

4. Τέλος, κάνουμε πλήρη αποχαρακτηρισμό (Unchunk) για κάθε μονάδα που είναι άρθρο, επίθετο και ουσιαστικό και βλέπουμε το αποτέλεσμα του Σχήματος 6.23)

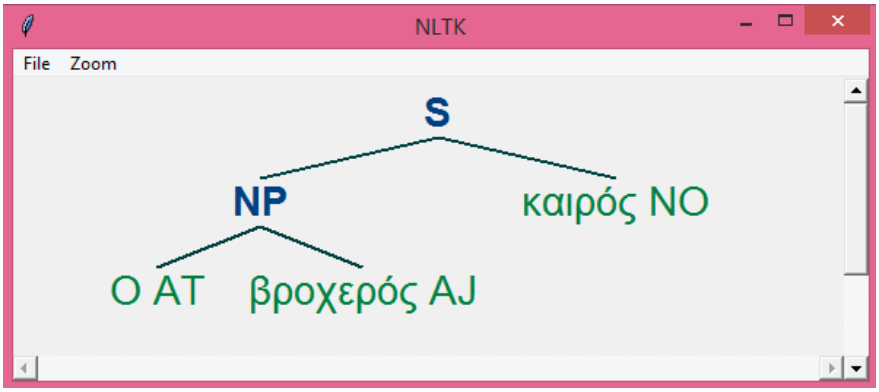
Οι κανόνες που περιγράψαμε παραπάνω εφαρμόζονται στην παρακάτω πρόταση υπό μορφή δέντρου, όπως στο Σχήμα 6.19.



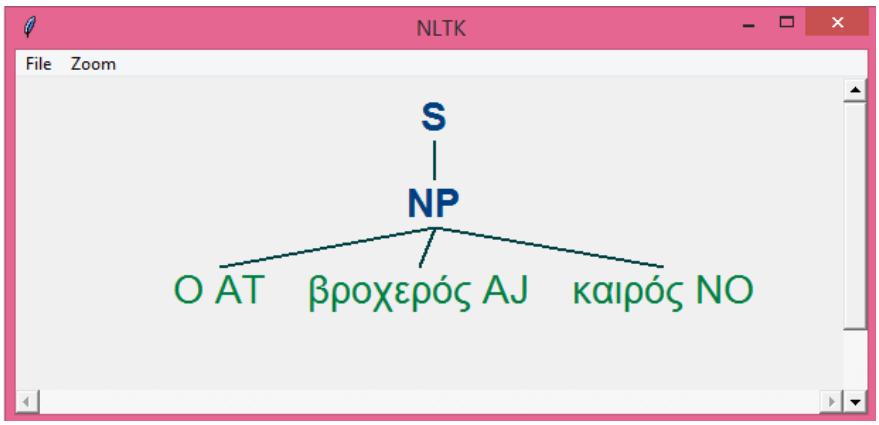
Σχήμα 6.19 Η πρόταση "Ο/At βροχερός/Aj καιρός/No" σε γραφικό δέντρο



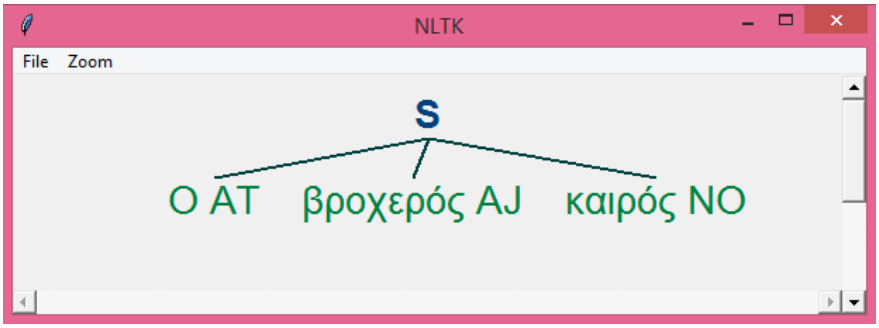
Σχήμα 6.20 ChunkRule('<AJ>', 'single Adjective')



Σχήμα 6.21 ExpandLeftRule('<AT>', '<AJ>', 'get left article')



Σχήμα 6.22 ExpandRightRule('<AJ>', '<NO>', 'get right noun')



Σχήμα 6.23 UnChunkRule('<AT><.*>*', 'unchunk everything')

```
from nltk.chunk.regexp import ChunkRule, ExpandLeftRule,
ExpandRightRule, UnChunkRule
from nltk.chunk import RegexpChunkParser

sent = [('Ο', 'AT'), ('βροχερός', 'AJ'), ('καιρός', 'NO')]
ch = ChunkRule('<AJ>', 'single Adjective')
exL = ExpandLeftRule('<AT>', '<AJ>', 'get left article')
exR = ExpandRightRule('<AJ>', '<NO>', 'get right noun')
unR = UnChunkRule('<AT><.*>*', 'unchunk everything')
chunk = RegexpChunkParser([ch, exL, exR, unR])
print(chunk.parse(sent))
chunk.parse(sent).draw()
```

Σχήμα 6.24 Κανόνες επέκτασης φράσεων

Κεφάλαιο 7

Transforming Chunks and Trees

Στην παρούσα διατριβή χρησιμοποιούνται κανόνες μεταφοράς σε επίπεδο υποφράσεων (chunk-tree). Στις επόμενες ενότητες θα δούμε μερικά παραδείγματα κανόνων.

7.1 Εφαρμογή κανόνων Μεταφοράς Σειράς Φράσεων

7.1.1 Τεμαχισμός πρότασης εισόδου

Θα χρησιμοποιήσουμε μια παραλλαγή του κώδικα αναλυτή φράσης από το Κεφάλαιο 6 και θα πάρουμε ως παράδειγμα πρότασης εισόδου μια μικρή παραλλαγή της ίδιας πρότασης, προσθέτοντας στο τέλος τη χρονική φράση "Τα Χριστούγεννα", ώστε να έχουμε μια πρόταση σειράς "SVOT" (Σχήμα 39). Μπορούμε να δούμε τη γραμματική του αναλυτή φράσης στο Σχήμα 7.1 που ακολουθεί.

```
sent = [('Βροχές' ,'NoCmFePlAc'), ('και','CjCo'), ('καταιγίδες' ,'NoCmFePlAc'),
('θα','PtFu'), ('εκδηλωθούν','VbMnIdXx03PlXxPePvXx'), ('κατά','AsPpSp'),
('τόπους','NoCmMaPlAc'), ('στη','AsPpPaFeSgAc'), ('Δυτική','AjBaFeSgAc'),
('Ελλάδα','NoPrFeSgAc'), ('Τα','AtDfNePlNm'), ('Χριστούγεννα','NoPrNePlAc')]
```

Σχήμα 7.1 Επισημειωμένη πρόταση εισόδου

Η παραπάνω γραμματική του Chunker (Σχήμα 7.2) εξάγει την επισημειωμένη πρόταση (Σχήμα 7.1) στο δέντρο φράσεων (Σχήμα 7.9).

```
ch = RegexpParser(r'''
NP:
    {<.*>*<No.*>}    # chunk as noun everything ends with Nouns
    }<Vb.*|Pt.*>+{               # chink verbs
    }<AtDfNePlNm><NoPrNePlAc>{  # chink nouns Common Plural
NP-CM:
    {<AtDfNePlNm><NoPrNePlAc>} # chunk nouns Common Plural
VB:
    {<Vb.*|Pt.*>+}              # chunk verbs
''')
```

Σχήμα 7.2 Γραμματική κανονικών εκφράσεων του RegexpChunker

7.1.2 Κανόνας μεταφοράς χρονικής φράσης (ChunkTime Rule)

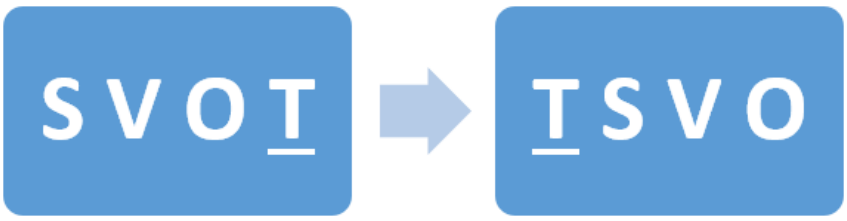
Ο κανόνας μεταφοράς χρονικής φράσης μεταφέρει οποιαδήποτε χρονική φράση στην αρχή της πρότασης. Χρησιμοποιώντας την κλάση εφαρμογής δέντρων της βιβλιοθήκης NLTK-Tree, που αποτελεί και την έξοδο δεδομένων του ChunkerParser της ίδιας κλάσης NLTK, έχουμε δημιουργήσει τη ρουτίνα "MoveChunkTime(input-tree)", η οποία δέχεται ως παράμετρο εισόδου δεδομένα μορφής δέντρου-φράσεων και επιστρέφει πάλι δέντρο-πρόταση ίδιου τύπου δεδομένων. Παρακάτω μπορείτε να δείτε τον κώδικα της ρουτίνας αυτής, ο οποίος μπορεί να εισαχθεί ως εφαρμογή κλάσης σε οποιονδήποτε κώδικα Python, με χρήση της εισαγωγής κλάσης εφαρμογής με τον κώδικα "from GSL_Tree_Transforms import MoveChunkTime". Οπου "GSL_Tree_Transforms" είναι το αρχείο Python που φιλοξενεί τον κώδικα της ρουτίνας του Σχήματος 7.3.

```
#KANONAS ORISMOS XRONIKOU SHMEIOU PANTA PAEI STHN ARXH. OSO AFORA
MONO GIA XRONIKA SHMEIA.
def MoveChunkTime(tree):
    newTree = tree.copy() #Tree('S',[])
    for i in range(len(newTree)):
        if newTree[i].label()=='NP-CM' and
        newTree[i][1][0]=='Χριστούγεννα':
            newTree.insert(0,tree[i])
            newTree.pop(i+1)
    return newTree
```

Σχήμα 7.3 Ρουτίνα "MoveChunkTime(tree)"

Η παραπάνω ρουτίνα κανόνα "MoveChunkTime" (Σχήμα 7.3) μετατρέπει ένα δέντρο-πρόταση σε φράση, ψάχνει ένα-ένα τα chunks και ελέγχει αν η ετικέτα κεφαλής κάθε chunk είναι επισημειωμένη με "NP-CM", που δηλώνει ότι έχουμε φράση ουσιαστικού ο-

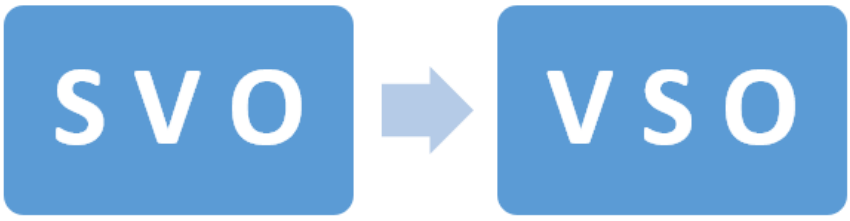
νομασίας (Noun Comment Type). Αν η φράση περιέχει τη λέξη “Χριστούγεννα”, η συνθήκη έχει ικανοποιηθεί πλήρως και τότε η συγκεκριμένη υποφράση μεταφέρεται στην αρχή της πρότασης (“newTree.insert(0,tree[i])”) και έχουμε μεταφορά της σειράς από “SVOT” σε “TSVO” (Σχήμα 7.4).



Σχήμα 7.4 Μεταφορά “SVOT” σε “TSVO”

7.1.3 Κανόνας μεταφοράς ρήματος φράσης (ChunkVerbSwap Rule)

Με τον ίδιο τρόπο δημιουργούμε και τον κανόνα αντιμετάθεσης του ρήματος “SVO->VSO” (Σχήμα 7.5), όπου το ρήμα μεταφέρεται πίσω από τη λεκτική μονάδα τύπου “Subject”.



Σχήμα 7.5 Κανόνας μεταφοράς “SVO -> VSO”

Ο κανόνας μεταφοράς ρήματος φράσης μεταφέρει οποιαδήποτε ρηματική φράση πίσω από τη φράση “Np-Subject”. Χρησιμοποιώντας την κλάση εφαρμογής δέντρων της βιβλιοθήκης NLTK-Tree, που αποτελεί και την έξοδο δεδομένων του ChunkerParser πάλι της ίδιας κλάσης NLTK, έχουμε δημιουργήσει τη ρουτίνα “VerbChunkSwap(input-tree)” (Σχήμα 7.6), η οποία δέχεται ως παράμετρο εισόδου δεδομένα τύπου δέντρου που αφορούν μια πρόταση τεμαχισμένη και χαρακτηρισμένη σε υποφράσεις και επιστρέφει πάλι δέντρο-πρόταση ίδιου τύπου δεδομένων. Παρακάτω μπορείτε να δείτε τον κώδικα της ρουτίνας αυτής, ο οποίος μπορεί να εισαχθεί ως εφαρμογή κλάσης σε οποιονδήποτε κώδικα Python, με χρήση της εισαγωγής κλάσης εφαρμογής με τον κώδικα “from GSL_Tree_Transforms import VerbChunkSwap”. ’Οπου “GSL_Tree_Transforms” είναι το αρχείο Python που φιλοξενεί τον κώδικα της ρουτίνας του Σχήματος 7.6.

```
#KANONAS ANTIMETATHESIS RHMATOS
def VerbChunkSwap(tree):
    newTree = tree.copy()
    for i in range(len(newTree)):
        #print(i,newTree[i].label())
        if newTree[i].label()=='VB':
            newTree.insert(i-1,tree[i])
            newTree.pop(i+1)
    return newTree
```

Σχήμα 7.6 Κώδικας κανόνα "VerbChunkSwap"

Η παραπάνω ρουτίνα κανόνα "VerbChunkSwap" δέχεται ως παράμετρο εισόδου δεδομένα τύπου δέντρου που αφορούν μια πρόταση τεμαχισμένη και χαρακτηρισμένη σε υποφράσεις, ψάχνει ένα-ένα τα chunks και ελέγχει αν η ετικέτα κεφαλής κάθε chunk είναι επισημειωμένη με "VB", που δηλώνει ότι έχουμε ρηματική φράση. Εφόσον έχει ικανοποιηθεί η συνθήκη, η συγκεκριμένη ρηματική υποφράση μεταφέρεται μία θέση ακριβώς πριν από την πρώτη αριστερή φράση ("newTree.insert(i-1,tree[i])") στην προκειμένη περίπτωση και έχουμε μεταφορά της σειράς από "SVO" σε "VSO" (Σχήμα 7.5).

7.1.4 Εφαρμογή διπλού κανόνα VerbChunSwap και MoveChunkTime

Εφαρμόζοντας τις ρουτίνες "VerbChunkSwap" και "MoveChunkTime" έχουμε εφαρμογή διπλού κανόνα και την μετατροπή "SVOT->TVSO" (Σχήμα 7.7).

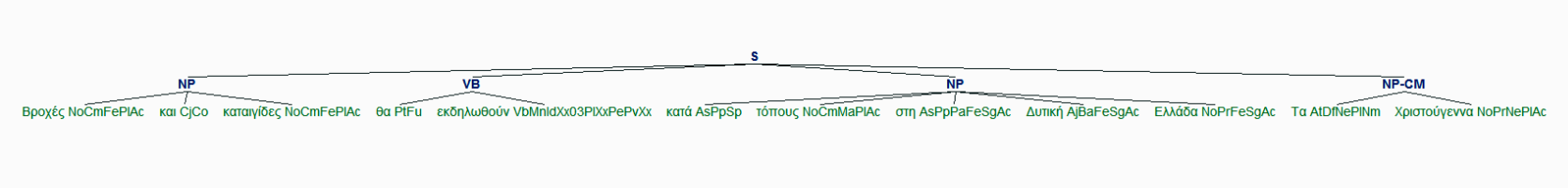
Σχήμα 7.7 Εφαρμογή διπλού κανόνα "MoveChunkTime" και "VerbChunkSwap"

Ο κώδικας του παραδείγματος, όπως μπορείτε να δείτε στο Σχήμα 7.8, εξάγει τα ακόλουθα αποτελέσματα ανά στάδιο. Καταρχήν, γίνεται ανάλυση της πρότασης σε chunks (Σχήμα 7.9). Η τεμαχισμένη σε φράσεις πρόταση, σε μορφή δέντρου πλέον, περνάει ως είσοδος στον κανόνα χρονικής μετατόπισης και στη συνέχεια έχουμε την έξοδο του Σχήματος 7.10. Τέλος, η έξοδος του κανόνα χρονικής μεταφοράς αποτελεί στην συνέχεια είσοδο στον δεύτερο κανόνα της ρηματικής μετατόπισης προς τα πίσω, η οποία μας οδηγεί στην έξοδο του Σχήματος 7.11.

```
from nltk.chunk import RegexpParser
from nltk import Tree
from GSL_Tree_Transforms import MoveChunkTime
from GSL_Tree_Transforms import VerbChunkSwap
sent = [('Βροχές' ,'NoCmFePlAc'),('και','CjCo'),('καταιγίδες'
,'NoCmFePlAc'),       ('θα','PtFu'),
('εκδηλωθούν','VbMnIdXx03PlXxPePvXx'),          ('κατά','AsPpSp'),
('τόπους','NoCmMaPlAc'), ('στη','AsPpPaFeSgAc'),
('Δυτική','AjBaFeSgAc'), ('Ελλάδα','NoPrFeSgAc'),
('Τα','AtDfNePlNm'), ('Χριστούγεννα','NoPrNePlAc')]
ch = RegexpParser(r'''
NP:
    {<.*>*<No.*>} # chunk as noun everything start with  Article
                  # and ends with Nouns
    }<Vb.*|Pt.*>+{              # chink verbs
    }<AtDfNePlNm><NoPrNePlAc>{  # chink nouns Common Plural
NP-CM:
    {<AtDfNePlNm><NoPrNePlAc>} # chunk nouns Common Plural
VB:
    {<Vb.*|Pt.*>+}             # chunk verbs
''')
tree = ch.parse(sent)
print(tree)
tree.draw()

#MOVES SPECIAL TIMES_CHUNK TO THE START OF SENT. AS THE CHRISTMAS
TIME.
newTree = MoveChunkTime(tree)
newTree.draw()

#SWAP VERB_CHUNK WITH NP_CHUNK.
newTree = VerbChunkSwap(newTree)
newTree.draw()
```

Σχήμα 7.8 Κώδικας γραμματικής και εφαρμογής διπλού κανόνα

```
S
 (NP Βροχές/NoCmFePlAc και/CjCo καταιγίδες/NoCmFePlAc)
 (VB θα/PtFu εκδηλωθούν/VbMnIdXx03PlXxPePvXx)
 (NP
  κατά/AsPpSp
  τόπους/NoCmMaPlAc
  στη/AsPpPaFeSgAc
  Δυτική/AjBaFeSgAc
  Ελλάδα/NoPrFeSgAc)
 (NP-CM Τα/AtDfNePlNm Χριστούγεννα/NoPrNePlAc)
)
```

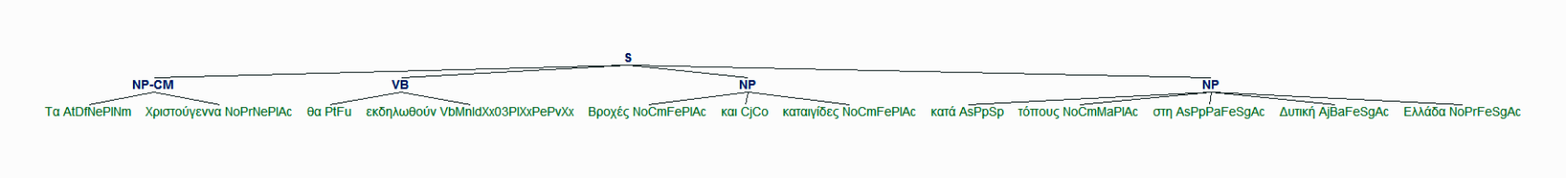


Σχήμα 7.9 Έξοδος τεμαχισμένης πρότασης από αναλυτή φράσεων

```
(S
 (NP-CM      Τα/AtDfNePlNm      Χριστούγεννα/NoPrNePlAc)
 (NP  Βροχές/NoCmFePlAc  και/CjCo  καταιγίδες/NoCmFePlAc)
 (VB      θα/PtFu      εκδηλωθούν/VbMnIdXx03PlXxPePvXx)
 (NP
  κατά/AsPpSp
  τόπους/NoCmMaPlAc
  στη/AsPpPaFeSgAc
  Δυτική/AjBaFeSgAc
  Ελλάδα/NoPrFeSgAc)
)
```


Σχήμα 7.10 Έξοδος κανόνα χρονικής μεταφοράς

```
(S
 (NP-CM Τα/AtDfNePlNm Χριστούγεννα/NoPrNePlAc)
 (VB θα/PtFu εκδηλωθούν/VbMnIdXx03PlXxPePvXx)
 (NP Βροχές/NoCmFePlAc και/CjCo καταιγίδες/NoCmFePlAc)
 (NP
  κατά/AsPpSp
  τόπους/NoCmMaPlAc
  στη/AsPpPaFeSgAc
  Δυτική/AjBaFeSgAc
  Ελλάδα/NoPrFeSgAc)
)
```



Σχήμα 7.11 Έξοδος κανόνα ρηματικής μετατόπισης

Κεφάλαιο 8

Στατιστικά Γλωσσικά Μοντέλα (ΣΓΜ)

Τα στατιστικά γλωσσικά μοντέλα χρησιμοποιούν τεχνικές στατιστικής εκτίμησης γλωσσικών δεδομένων εκπαίδευσης που εφαρμόζονται σε εκτεταμένα κείμενα με σκοπό τη μοντελοποίηση της γλώσσας και ανικατοπτρίζουν επιτυχώς τη γλώσσα που αντιπροσωπεύουν. Τα μοντέλα N-grams αποτελούν μία από τις πιο δημοφιλείς τεχνικές στατιστικής εκτίμησης. Ο ρόλος τους είναι πολύ σημαντικός για μια σειρά από εφαρμογές της γλωσσικής τεχνολογίας, όπως η αναγνώριση φωνής, η οπτική αναγνώριση χαρακτήρων, η μηχανική μετάφραση, ακόμη και η ορθογραφική διόρθωση. Σε αυτή την ενότητα παρουσιάζουμε τα αποτελέσματα αξιολόγησης διαφόρων στατιστικών γλωσσικών μοντέλων που δημιουργήσαμε από ποικίλους τύπους σωμάτων σύντομων μεταγραφών της ΕΝΓ, με σκοπό να αντλήσουμε χρήσιμα συμπεράσματα. Χρήσιμες πληροφορίες για τα στατιστικά γλωσσικά μοντέλα, αλλά και τα συστήματα στατιστικής μηχανικής μετάφρασης, παρέχει το βιβλίο των Jurafsky και Martin (Martin and Jurafsky, 2009), καθώς το βιβλίο του Koehn (Koehn, 2009).

8.1 Υπολογισμός πιθανοτήτων απλών μοντέλων N-grams

Ο σκοπός ενός γλωσσικού μοντέλου είναι να ορισθεί η πιθανότητα $P\left(\mathrm{w}_1^{\mathrm{n}}\right)$ μιας ακολουθίας λέξεων $\mathrm{w}_1^{\mathrm{n}} = \mathrm{w}_1, \mathrm{w}_2, \mathrm{w}_3 ... \mathrm{w}_n$. Μπορούμε να χρησιμοποιήσουμε τον κανόνα της αλυσίδας των πιθανοτήτων για να υπολογίσουμε την πιθανότητα:

$$P\left(\mathrm{w}_1^{\mathrm{n}}\right) = \mathrm{P}\left(\mathrm{w}_1\right) P\left(\mathrm{w}_2|\mathrm{w}_1\right) P\left(\mathrm{w}_3|\mathrm{w}_1^2\right) \cdots P\left(\mathrm{w}_{\mathrm{n}}|\mathrm{w}_1^{\mathrm{n}-1}\right) = \prod_{k=1}^{n} \mathrm{P}\left(\mathrm{w}_{\mathrm{k}}|w_1^{k-1}\right) \qquad (8.1)$$

Για μεγάλα σύνολα n, μπορούμε να υπολογίσουμε την πιθανότητα μιας λέξης που εξαρτάται μόνο από δύο προγενέστερες λέξεις, βάσει της θεωρίας των τριγραμμάτων, ό-

που ισχύει ότι $P\left(\mathrm{w}_n|\mathrm{w}_1^{n-1}\right) \cong P\left(\mathrm{w}_n|\mathrm{w}_{\mathrm{n}-2}, \mathrm{w}_{\mathrm{n}-1}\right)$ και η οποία στην πράξη εφαρμόζεται σε ικανοποιητικό βαθμό. Η υπόθεση αυτή, που θέλει την πιθανότητα της λέξης να εξαρτάται μόνο από τις προηγούμενες δύο λέξεις, καλείται υπόθεση Markov. Η εν λόγω υπόθεση δέχεται ότι η πιθανότητα ενός μελλοντικού γεγονότος μπορεί να προβλεφθεί κοιτώντας το άμεσο και όχι το πολύ μακρινό παρελθόν του. Το γλωσσικό μοντέλο N-grams χρησιμοποιεί τις προηγούμενες N -1 λέξεις (τυπικά μία ή δύο). Συνεπώς, ένα δίγραμμα καλείται ως "μοντέλο Markov πρώτης τάξης" (κοιτάζει μία λέξη στο παρελθόν), ενώ ένα τρίγραμμα καλείται ως "μοντέλο Markov δεύτερης τάξης". Σε γενικές γραμμές κάθε μοντέλο N-gram καλείται ως "μοντέλο Markov N -1 τάξης".

Άρα, η γενική εξίσωση για κάθε N-gram, υπολογίζοντας την υπό συνθήκη πιθανότητα μιας επόμενης λέξης σε μια ακολουθία, είναι:

$$P\left(\mathrm{w}_n|\mathrm{w}_1^{n-1}\right) \approx P\left(\mathrm{w}_n|\mathrm{w}_{\mathrm{v}-\mathrm{N}+1}^{n-1}\right) \tag{8.2}$$

Η γενική αυτή εξίσωση δείχνει ότι, η πιθανότητα μιας λέξης w_n βάσει όλων των προγενέστερων λέξεων μπορεί να προσεγγιστεί με την πιθανότητα βάσει των N τελευταίων λέξεων. Βάσει ενός διγραμμικού μοντέλου η συνολική πιθανότητα του string μέσω της προηγούμενης εξίσωσης, είναι ίση με:

$$P\left(\mathrm{w}_1^{\mathrm{n}}\right) \approx \prod_{k=1}^{n} \mathrm{P}\left(\mathrm{w}_\mathrm{k}|w_{k-1}\right) \tag{8.3}$$

Τα μοντέλα N-grams μπορούν να εκπαιδευτούν μετρώντας και κανονικοποιώντας τις εμφανίσεις τους σε ένα σώμα κειμένων εκπαίδευσης. Παίρνουμε ένα σώμα κειμένων, υπολογίζουμε τον αριθμό ενός συγκεκριμένου διγράμματος και τον διαιρούμε με το άθροισμα όλων των διγραμμάτων που μοιράζονται την ίδια πρώτη λέξη:

$$P\left(\mathrm{w}_n|\mathrm{w}_{\mathrm{n}-1}\right) = \frac{c\left(w_{n-1}w_n\right)}{\sum_w c\left(w_{n-1}w\right)} \tag{8.4}$$

Το άθροισμα όλων των διγραμμάτων που ξεκινούν με αυτή τη λέξη ισούται με τον αριθμό εμφάνισης της λέξης. Άρα μπορούμε να γράψουμε ότι ισχύει:

$$P\left(\mathrm{w}_n|\mathrm{w}_{\mathrm{n}-1}\right) = \frac{C\left(w_{n-1}w_n\right)}{C\left(w_{n-1}\right)} \tag{8.5}$$

Και στη γενικευμένη μορφή της, στην περίπτωση των N-grams:

$$P\left(\mathrm{w}_n|\mathrm{w}_{\mathrm{n-N+1}}^{n-1}\right) = \frac{c\left(w_{n-N+1}^{n-1}w_n\right)}{c\left(w_{n-N+1}^{n-1}\right)} \tag{8.6}$$

Για τα τριγράμματα η πιθανότητα αυτή γράφεται ως εξής:

$$P\left(\mathrm{w}_3|\mathrm{w}_1,\mathrm{w}_2\right) = \frac{C\left(w_1,w_2,w_3\right)}{C\left(w_1,w_2\right)} \tag{8.7}$$

Η τελευταία εξίσωση υπολογίζει την πιθανότητα του N-gram διαιρώντας τον αριθμό εμφάνισης μιας ακολουθίας λέξεων με τον αριθμό εμφάνισης των προηγούμενων λέξεων. Ο λόγος αυτός καλείται σχετική συχνότητα. Οι εκτιμήσεις της πιθανότητας N-gram μπορούν να υπολογισθούν χρησιμοποιώντας τις σχετικές πιθανότητες, που καλούνται εκτιμήσεις μέγιστης πιθανότητας, δηλαδή οι κανονικοποιημένοι αριθμοί εμφάνισης των N-grams σε ένα συγκεκριμένο σώμα κειμένων εκπαίδευσης.

8.2 Μοντέλα Εξομάλυνσης

Ένα από τα σημαντικότερα προβλήματα των N-grams είναι η ίδια η εκπαίδευσή τους. Η απόδοσή τους εξαρτάται άμεσα από την ποιότητα των σωμάτων κειμένων που χρησιμοποιούνται για την εκπαίδευσή τους. Κάθε σώμα κειμένων είναι πεπερασμένο και μερικά από τα πιο αποδεκτά N-grams δεν βρίσκονται σε αυτό. Ένα άλλο πρόβλημα που αντιμετωπίζουν τα N-grams είναι η αδυναμία τους να χρησιμοποιήσουν μακράς απόστασης συμφραζόμενα, με αποτέλεσμα να υπάρχει εμφανής τάση υποεκτίμησης της πιθανότητας των strings (συμβολοσειρές/λέξεις) που τυγχάνει να μην συνυπάρχουν στα δεδομένα εκπαίδευσης. Για αυτόν τον λόγο αναπτύχθηκαν τεχνικές που αναθέτουν μη μηδενικές πιθανότητες σε N-grams που εμφανίζουν μηδενικό αριθμό εμφανίσεων στα δεδομένα εκπαίδευσης. Η διαδικασία της επαναξιολόγησης των N-grams με μηδενική ή μικρή πιθανότητα καλείται εξομάλυνση. Στις επόμενες παραγράφους θα περιγράψουμε μερικές από τις πιο δημοφιλές τεχνικές εξομάλυνσης.

8.2.1 Τεχνική Good-Turing

Μεταξύ των πιο διαδεδομένων τεχνικών στα γλωσσικά μοντέλα είναι η εξομάλυνση Good-Turing. Αποτελεί μια βασική προσέγγιση ανάμεσα σε όλες τις μεθόδους "έκπτωσης". Περιγράφηκε για πρώτη φορά από τον Good (Good, 1953) ο οποίος βασίστηκε στην

ιδέα του "Turing", ενώ η πλήρης απόδειξή της παρουσιάστηκε από τους Church et al. (Church and Gale, 1991).

Η βασική ιδέα αυτής της τεχνικής είναι ότι εκτιμάται το μέγεθος της μάζας πιθανότητας που θα ανατεθεί στα N-grams με μηδενική ή μικρή πιθανότητα, βάσει των N-grams με μεγαλύτερο αριθμό εμφανίσεων. Με άλλα λόγια εξετάζεται το N_c, δηλαδή ο αριθμός των N-grams που εμφανίζονται c φορές. Με τον αριθμό των N-grams που εμφανίζονται c φορές αναφερόμαστε στη συχνότητα της συχνότητας c . Έτσι, εφαρμόζοντας την ιδέα της εξομάλυνσης των διγραμμάτων, με N_0 αναφερόμαστε στα διγράμματα b με αριθμό εμφανίσεων το 0, με N_1 αναφερόμαστε στα διγράμματα b με αριθμό εμφανίσεων το 1, και ούτω καθεξής:

$$N_c = \sum_{b:c(b)=c} 1 \tag{8.8}$$

Με αυτή την τεχνική ο μη μηδενικός αριθμός εμφάνισης N-grams υπολογίζεται ως εξής:

$$c^* = (c+1)\frac{N_{c+1}}{N_c} \tag{8.9}$$

Όπου c* συμβολίζεται ο αριθμός των εμφανίσεων με τη χρήση της εξομάλυνσης και Nc ο αριθμός εμφάνισης των N-grams που εμφανίζονται c φορές. Ο υπολογισμός βάσει της τεχνικής Good-Turing για τα N-grams με μηδενική συχνότητα εμφάνισης είναι

$$c^* = \frac{N_1}{N_0} \tag{8.10}$$

και μπορεί να μεταφραστεί σε πιθανότητα κανονικοποιώντας τον αρχικό αριθμό με την κατανομή N:

$$P_{\text{Good-Turing}} = \frac{c^*}{N} \tag{8.11}$$

Όπως γίνεται αντιληπτό η εκτίμηση Good-Turing για τα διγράμματα που δεν έχουν εμφανιστεί στα δεδομένα εκπαίδευσης πραγματοποιείται με τον λόγο του αριθμού των διγραμμάτων που εμφανίστηκαν μία φορά προς τον αριθμό εκείνων που δεν έχουν εμφανιστεί ακόμα. Το να πραγματοποιούμε εκτιμήσεις γεγονότων που δεν έχουν συμβεί βάσει γεγονότων που συνέβησαν μία φορά είναι μια ιδέα που χρησιμοποιεί όχι μόνο ο αλγόριθμος Good-Turing αλλά και η τεχνική Witten-Bell που θα παρουσιάσουμε παρακάτω.

8.2.2 Τεχνική Witten-Bell

Μια ακόμη συχνά εφαρμοζόμενη τεχνική έκπτωσης είναι αυτή των Witten-Bell (Witten and Bell, 1991), η οποία εκτιμά την πιθανότητα των μηδενικών N-grams, λαμβάνοντας υπόψη τα N-grams που εμφανίζονται για πρώτη φορά και όχι αυτά που εμφανίζονται ακριβώς μία φορά, όπως γίνεται με την μέθοδο Good-Turing. Η ιδέα πίσω από αυτή την τεχνική είναι ότι τα N-grams που εμφανίζουν μηδενική συχνότητα μπορούν να μοντελοποιηθούν με την πιθανότητα ενός N-gram που εμφανίζεται για πρώτη φορά. Το ερώτημα όμως είναι πώς θα υπολογίσουμε την πιθανότητα εμφάνισης για πρώτη φορά ενός N-gram. Η απάντηση δίνεται με τον υπολογισμό του αριθμού εμφάνισης των N-grams για πρώτη φορά στα δεδομένα εκπαίδευσης. Ο υπολογισμός αυτός είναι πολύ εύκολος από τη στιγμή που ο αριθμός των πρώτο-εμφανιζόμενων N-grams είναι ίσος με τον αριθμό των ειδών N-grams που υπάρχουν στα δεδομένα εκπαίδευσης.

Συνεπώς, η συνολική πιθανότητα όλων των μηδενικών N-grams μπορεί να υπολογιστεί από τον λόγο του αριθμού των ειδών προς το άθροισμα του αριθμού των ειδών με τον συνολικό αριθμό στοιχείων στα δεδομένα εκπαίδευσης.

$$\sum_{i:c_i} p_i = \frac{T}{N+T} \tag{8.12}$$

Η προηγούμενη σχέση δίνει την εκτίμηση μέγιστης πιθανότητας για γεγονότα εμφάνισης νέου είδους. Ας σημειωθεί εδώ ότι ο αριθμός των παρατηρούμενων ειδών T είναι διαφορετικός από τον συνολικό αριθμό ειδών (N) που χρησιμοποιείται στην τεχνική προσθήκης ενός. Ο αριθμός T συμβολίζει είδη που έχουμε ήδη δει στο κείμενο ενώ N είναι ο συνολικός αριθμός των ειδών που θα δούμε εντέλει. Αυτή είναι η συνολική πιθανότητα των μη εμφανιζόμενων N-grams. Ας υποθέσουμε ότι Z είναι ο συνολικός αριθμός ειδών των N-grams με αριθμό εμφάνισης το μηδέν, άρα η πιθανότητα του κάθε μονογράμματος μετά τον ισοκαταμερισμό της μάζας πιθανότητας θα είναι ίση με

$$p_i = \frac{T}{Z(N+T)} \tag{8.13}$$

και

$$Z = \sum_{i:c_i} 1 \tag{8.14}$$

Εάν η συνολική πιθανότητα των μηδενικών N-grams υπολογίζεται από την Εξίσωση 8.7 τότε η παραπάνω πιθανότητα θα πρέπει να απορρέει από την έκπτωση των πιθανοτήτων όλων των εμφανιζόμενων N-grams. Άρα:

$$p_i = \frac{c_i}{\mathrm{Z}(N+T)}, if(c_i > 0) \tag{8.15}$$

Χρησιμοποιώντας σαν λόγο κανονικοποίησης, μετά τη μείωση, τον $\frac{N}{N+T}$, όπου N είναι ο αρχικός αριθμός εμφάνισης του N-gram, ο αριθμός των εμφανίσεων διαμορφώνεται ως εξής:

$$c_i \begin{cases} \frac{T}{Z}\frac{N}{N+T} \text{ if } c_i = 0 \\ c_i\frac{N}{N+T} \text{ if } c_i > 0 \end{cases} \tag{8.16}$$

Ο αλγόριθμος Witten-Bell για τα μονογράμματα μοιάζει κατά πολύ με την εξομάλυνση προσθήκης ενός, αλλά αν επεκταθούμε στα διγράμματα τότε υπάρχει διαφορά. Σε αυτή την περίπτωση, για να υπολογίσουμε την πιθανότητα ενός διγράμματος $w_{i-1}w_i$ που δεν έχει εμφανιστεί, θα χρησιμοποιηθεί η πιθανότητα εμφάνισης ενός νέου διγράμματος που ξεκινάει από την λέξη w_{i-1}.

Μπορούμε λοιπόν να χρησιμοποιήσουμε την παρακάτω σχέση για να αποτυπώσουμε τον αλγόριθμο Witten-Bell για τα διγράμματα:

$$\sum_{i:c(w_x w_i)} \mathrm{p_i}\left(w_i|w_{i-1}\right) = \frac{T\left(w_x\right)}{N\left(w_x\right)+T\left(w_x\right)} \tag{8.17}$$

Όπου ο όρος $T(w_x)$ αντιστοιχεί στον αριθμό των ειδών των διγραμμάτων και ο $N(w_x)$ στον συνολικό αριθμό διγραμμάτων με πρώτη λέξη την w_x. Συνεπώς, κάθε πρώην μηδενικό δίγραμμα μοιράζεται εξίσου τη μάζα πιθανότητας που απέμεινε από την έκπτωση και έτσι έχουμε:

$$\mathrm{p_i}\left(w_i|w_{i-1}\right) = \frac{T\left(w_{i-1}\right)}{Z\left(w_{i-1}\right)\left(N+T\left(w_{i-1}\right)\right)} \text{ if } \left(c_{w_{i-1}w_i} = 0\right) \tag{8.18}$$

Έστω N ο αριθμός των διγραμμάτων που αρχίζουν με w_{i-1} και Z ο συνολικός αριθμός των διγραμμάτων με την πρώτη λέξη να έχει αριθμό εμφάνισης το μηδέν, δηλαδή $Z(w_{i-1}) = V - T(w_{i-1})$. Όπου V είναι ο αριθμός του λεξιλογίου και συνεπώς ο αριθμός των πιθανών ειδών των διγραμμάτων. Όσον αφορά τα μη μηδενικά διγράμματα ισχύει ότι:

$$\mathrm{p_i}\left(w_i|w_{i-1}\right) = \frac{c\left(w_{i-1}w_i\right)}{c\left(w_{i-1}\right)+T\left(w_{i-1}\right)} if\left(c_{w_{i-1}w_i} > 0\right) \tag{8.19}$$

8.2.3 Τεχνική Katz's backing off

Η εξομάλυνση Katz (Katz, 1987) είναι αποκλειστικά μια μορφή μη γραμμικής εξομάλυνσης Backoff. Βάσει αυτής της τεχνικής πραγματοποιείται διαδικασία υποχώρησης σε μικρότερης τάξης μοντέλα, π.χ (N-1)-grams, όταν απαιτείται να προσδιορισθούν μηδενικά εμφανιζόμενα N-grams. Η αρχική φάση της εξομάλυνσης περιλαμβάνει και την τεχνική Good Turing. Η ουσιαστική διαφορά μεταξύ της παρεμβολής και της υποχώρησης είναι ότι για τα N-grams με μη μηδενικές εμφανίσεις η παρεμβολή χρησιμοποιεί πάντα την πληροφορία των μικρότερης τάξης N-grams, ενώ η υποχώρηση όχι. Στην περίπτωση της υποχώρησης Katz, όταν ένα N-gram έχει μηδενική πιθανότητα, τότε χρησιμοποιείται ένα (N-1)-gram και ενδεχομένως ένα μοντέλο μικρότερης τάξης, μέχρι να φθάσει στο σημείο να μπορεί να προσδιοριστεί. Έτσι, στη γενική περίπτωση μπορούμε να ισχυριστούμε ότι:

$$p_{\text{katz}}\left(w_n|w_{n-N+1}^{n-1}\right) = \begin{cases} P\left(w_n|w_{n-N+1}^{n-1}\right) & \varepsilon\grave{\alpha}\upsilon \quad C\left(w_{n-N+1}^{n-1}\right) > 0 \\ \alpha\left(w_{n-N+1}^{n-1}\right) p_{katz}\left(w_n|w_{n-N+2}^{n-1}\right) & \alpha\lambda\lambda\iota\omega\varsigma \end{cases} \tag{8.20}$$

8.3 Εντροπία και διασταυρωμένη εντροπία

Η εντροπία μπορεί να χρησιμοποιηθεί σαν ένα εργαλείο μέτρησης του πλήθους πληροφοριών που υπάρχουν σε μια ειδική γραμματική, της ποιότητας μοντελοποίησης της γλώσσας από μια γραμματική και του βαθμού προβλεψιμότητας της επόμενης λέξης βάσει των μοντέλων N-grams.

Ο υπολογισμός της εντροπίας απαιτεί να διασφαλιστεί μια τυχαία μεταβλητή x, η οποία αντιστοιχεί σε αυτό που θα περιγράψουμε (λέξεις, γράμματα, μέρη του λόγου, που ανήκουν σε ένα σύνολο X) και η οποία έχει μια ιδιαίτερη συνάρτηση πιθανότητας p(x). Η εντροπία, λοιπόν, της τυχαίας μεταβλητής αυτής είναι:

$$\mathrm{H(X)} = -\sum_{x \in X} p(x) \cdot \log_2 p(x) \tag{8.21}$$

Το αποτέλεσμα της εντροπία μετριέται σε bits. Ο πιο διαισθητικός τρόπος για να οριστεί η εντροπία είναι να θεωρηθεί σαν το κατώτερο άκρο του αριθμού των bits που απαιτούνται για να κωδικοποιηθεί ένα κομμάτι πληροφορίας.

Οι Cover και Thomas (Cover and Thomas, 1991) πρότειναν το ακόλουθο παράδειγμα. Ας υποθέσουμε ότι θέλουμε να στείλουμε ένα μικρό μήνυμα σε έναν booker ώστε να στοι-

Άλογο 1	1/2	Άλογο 5	1/64
Άλογο 2	1/4	Άλογο 6	1/64
Άλογο 3	1/8	Άλογο 7	1/64
Άλογο 4	1/16	Άλογο 8	1/64

Πίνακας 8.1 Η πιθανότητα νίκης του κάθε αλόγου στον αγώνα

χηματίσει σε ένα αγώνα ιπποδρομίας 8 αλόγων που πραγματοποιείται στην Αγγλία. Ένας τρόπος είναι να στείλουμε τη δυαδική αναπαράσταση του αριθμού του συγκεκριμένου αλόγου. Έτσι το άλογο νούμερο 1 θα έχει σαν κωδικό το 001, το άλογο νούμερο 2 το 010, το άλογο νούμερο 3 το 011 και το άλογο νούμερο 8 το 000. Κατά συνέπεια, αντιλαμβανόμαστε ότι θα απαιτηθούν κατά μέσο όρο 3 bits/αγώνα ώστε να στοιχηματίζουμε καθ' όλη τη διάρκεια της ημέρας. Το ερώτημα που τίθεται είναι αν μπορούμε να στείλουμε μήνυμα με λιγότερα bits. Για αυτόν τον λόγο θα χρησιμοποιήσουμε την πληροφορία που έχουμε για τα προγνωστικά του κάθε αγώνα, αναλογιζόμενοι ότι κάθε άλογο έχει μια προϊστορία νικών στους αγώνες σύμφωνα με τον Πίνακα 8.1.

Η εντροπία της τυχαίας μεταβλητής X δίνει ένα κατώτατο άκρο αριθμού bits και ισούται με:

$$\begin{aligned} \mathrm{H(X)} &= -\sum_{i=1}^{i=8} p(i) \cdot \log_2 p(i) \\ &= -\tfrac{1}{2}\log_2 \tfrac{1}{2} - \tfrac{1}{4}\log_2 \tfrac{1}{4} - \tfrac{1}{8}\log_2 \tfrac{1}{8} - \tfrac{1}{16}\log_2 \tfrac{1}{16} - 4\left(\tfrac{1}{64}\log_2 \tfrac{1}{64}\right) \\ &= -2\mathrm{bits} \end{aligned} \tag{8.22}$$

Άρα ο κώδικας που μπορεί να χρησιμοποιηθεί με μέσο όρο 2 bits/αγώνα μπορεί να υλοποιηθεί με τη χρήση λιγότερων bits για τα πιο πιθανά άλογα και με περισσότερα bits για τα λιγότερο πιθανά. Στην περίπτωση όπου όλα τα άλογα είναι εξίσου πιθανά να κερδίσουν τον κάθε αγώνα, το μήκος του δυαδικού κώδικα κάθε αλόγου θα έχει τιμή ίση με 3 bits, λόγω του ότι η πιθανότητα κάθε αλόγου ισούται με 1/8 και συνεπώς από την Εξίσωση 8.21 εξάγεται το συμπέρασμα ότι η εντροπία επιλογής του κάθε αλόγου θα ισούται με 3 bits.

Μέχρι τώρα υπολογίστηκε μόνο η εντροπία μιας τυχαίας μεταβλητής. Το ερώτημα που γεννάται είναι τι θα συμβεί στην περίπτωση που θέλουμε να υπολογίσουμε την εντροπία για μια ακολουθία λέξεων και όχι μιας μόνο λέξης. Για παράδειγμα, θέλουμε να υπολογίσουμε την εντροπία της ακόλουθης σειράς λέξεων $W = w_1, w_2, \ldots, w_N$. Μπορούμε λοιπόν να υπολογίσουμε την εντροπία μια τυχαίας μεταβλητής που κυμαίνεται σε όλες τις πεπερασμένες ακολουθίες λέξεων μήκους b, για κάποια γλώσσα L, με την παρα-

κάτω σχέση:

$$\mathrm{H}\left(\mathrm{w}_1, \mathrm{w}_2 \ldots \mathrm{w}_\mathrm{N}\right) = -\sum_{\mathrm{w}_1^\mathrm{N} \in L} P\left(\mathrm{w}_1^\mathrm{N}\right) \cdot \log_2 P\left(\mathrm{w}_1^\mathrm{N}\right) \tag{8.23}$$

Επιπλέον, μπορούμε να ορίσουμε τον ρυθμό της εντροπίας, ή πιο σωστά την ανά λέξη εντροπία, ως την εντροπία της ακολουθίας διαιρούμενη με τον αριθμό των λέξεων:

$$\frac{1}{N}\mathrm{H}\left(\mathrm{w}_1^\mathrm{N}\right) = -\frac{1}{N}\sum_{\mathrm{w}_1^\mathrm{N} \in L} P\left(\mathrm{w}_1^\mathrm{N}\right) \cdot \log_2 P\left(\mathrm{w}_1^\mathrm{N}\right) \tag{8.24}$$

Αν θέλουμε να υπολογίσουμε την πραγματική εντροπία θα πρέπει να θεωρήσουμε ακολουθίες λέξεων μη πεπερασμένες. Έτσι, αν θεωρήσουμε τη γλώσσα σαν μια στοχαστική διεργασία L η οποία παράγει ακολουθίες λέξεων, τότε η ανά λέξη εντροπία μπορεί να γραφτεί ως εξής:

$$H(L) = \lim_{N \to \infty} \frac{1}{N} H\left(\mathrm{w}_1, \mathrm{w}_2 \ldots \mathrm{w}_\mathrm{N}\right) \tag{8.25}$$

$$H(L) = \lim_{N \to \infty} \frac{1}{N} \sum_{\mathbf{w}_1^\mathrm{N} \in L} P\left(\mathrm{w}_1, \mathrm{w}_2 \ldots \mathrm{w}_\mathrm{N}\right) \cdot \log_2 P\left(\mathrm{w}_1, \mathrm{w}_2 \ldots \mathrm{w}_\mathrm{N}\right) \tag{8.26}$$

Το θεώρημα των Shannon-McMillan-Breiman (Jurafsky and Martin, 2000) αποδεικνύει ότι όταν μια γλώσσα είναι στάσιμη, ισχύει ότι:

$$H(L) = \lim_{N \to \infty} \frac{1}{N} \log_2 P\left(\mathrm{w}_1, \mathrm{w}_2 \ldots \mathrm{w}_\mathrm{N}\right) \tag{8.27}$$

Μπορούμε να ερμηνεύσουμε το θεώρημα αν αναλογιστούμε ότι μια ακολουθία λέξεων είναι τόσο μεγάλη όσο ένα άθροισμα μικρότερων προτάσεων. Η ιδέα του θεωρήματος είναι ότι μια μεγάλη ακολουθία λέξεων περιλαμβάνει πολλές μικρές και κάθε μικρή επαναλαμβάνεται στη μεγαλύτερη ακολουθία ανάλογα με την συχνότητα εμφάνισής της.

Για να ανακεφαλαιώσουμε, μπορούμε να πούμε ότι υπολογίζουμε την εντροπία μιας οποιασδήποτε στοχαστικής διεργασίας παίρνοντας ένα μεγάλο δείγμα της εξόδου και υπολογίζοντας τον μέσο όρο της λογαριθμικής της πιθανότητας. Στη θεωρία πληροφοριών αυτό που πραγματικά μετράμε είναι η διασταυρωμένη εντροπία (cross entropy) των δεδομένων εκπαίδευσης για ένα συγκεκριμένο μοντέλο. Επειδή η διασταυρωμένη εντροπία είναι το πιο γνωστό είδος εντροπίας, από εδώ και πέρα με τον όρο εντροπία θα αναφερόμαστε στη διασταυρωμένη εντροπία.

Η διασταυρωμένη εντροπία ορίζεται ως το ανώτατο όριο της εντροπίας, καθώς το μήκος μιας ακολουθίας λέξεων πάει προς το άπειρο. Άρα λοιπόν, χρειάζεται μια προσέγγιση

της διασταυρωμένης εντροπίας που θα σχετίζεται με μια ακολουθία λέξεων πολύ μεγάλη, αλλά με πεπερασμένο μήκος. Αυτή η προσέγγιση της διασταυρωμένης εντροπίας ενός μοντέλου $P(w_i|w_{i-N+1}, ..., w_{i-1})$, για μια ακολουθία λεξεων W είναι:

$$H(W) = -\frac{1}{N} \log P\left(\mathrm{w}_i|\mathrm{w}_{i-N+1} \ldots \mathrm{w}_{\mathrm{i-1}}\right) \tag{8.28}$$

Όταν συγκρίνονται δύο γλωσσικά μοντέλα ως προς ένα συγκεκριμένο σύνολο δεδομένων δοκιμής, το γλωσσικό μοντέλο που δίνει τη μεγαλύτερη πιθανότητα P(W), θεωρείται ως το πιο αντιπροσωπευτικό για να το μοντελοποιήσει. Παράλληλα, το μοντέλο αυτό παρουσιάζει και τη μικρότερη διασταυρωμένη εντροπία. Η διασταυρωμένη εντροπία H(W) ενός γλωσσικού μοντέλου για ένα σύνολο δεδομένων δοκιμής W που περιλαμβάνει N λέξεις, ορίζεται ως εξής:

$$H(W) = -\frac{1}{N} \log P(W) \tag{8.29}$$

και μπορεί να ερμηνευτεί ως ο μέσος όρος των bits που χρειάζονται για να κωδικοποιηθεί κάθε λέξη στο σύνολο δεδομένων δοκιμής, με τη χρήση του γλωσσικού μοντέλου.

8.4 Περιπλοκή – perplexity

Σκοπός της παρούσας εργασίας ήταν να διερευνήσει τις επιδόσεις των στατιστικών γλωσσικών μοντέλων, κυρίως των μοντέλων n-γραμμάτων, σε γλώσσες με μορφολογικές διαφορές και ιδιαίτερα γλώσσες με περίπλοκη μορφολογία, όπως η ΕΝΓ.

Η περιπλοκή ενός μοντέλου είναι το αντίστροφο του μέσου όρου της πιθανότητας που αντιστοιχεί στην κάθε λέξη στο σύνολο των δεδομένων δοκιμής και σχετίζεται με τη διασταυρωμένη εντροπία βάσει της εξίσωσης:

$$PP(W) = 2^{H(W)} \tag{8.30}$$

Αναδιατυπώνοντας την Εξίσωση 8.29 προκύπτει ότι:

$$H(W) = -\frac{1}{N} \log_2 P\left(w_1 w_2 \ldots w_N\right) \tag{8.31}$$

και κατά συνέπεια ότι:

$$2^{H(W)} = P\left(w_1 w_2 \ldots w_N\right)^{-\frac{1}{N}} \tag{8.32}$$

Από την Εξίσωση 8.30 προκύπτει ότι:

$$PP(W) = P\left(w_1 w_2 \ldots w_N\right)^{-\frac{1}{N}} \tag{8.33}$$

και παίρνοντας την N-ιοστή ρίζα ότι:

$$PP(W) = \sqrt[N]{\frac{1}{P\left(w_1 w_2 \ldots w_N\right)}} \tag{8.34}$$

Χρησιμοποιώντας τον κανόνα της αλυσίδας μπορούμε να επεκτείνουμε την πιθανότητα της ακολουθίας W:

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P\left(w_i | w_1 \ldots w_{i-1}\right)}} \tag{8.35}$$

Στην περίπτωση που θέλουμε να υπολογίσουμε την περιπλοκή μιας ακολουθίας W με ένα διγραμμικό γλωσσικό μοντέλο, τότε έχουμε:

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P\left(w_i | w_{i-1}\right)}} \tag{8.36}$$

Για κάθε γλωσσικό μοντέλο, είναι δυνατό να υπολογιστεί η περιπλοκή για κάποιο σώμα κειμένων που θα αναγνωριστεί. Για τεχνητά, περιορισμένα θέματα, με αυστηρή σύνταξη, η περιπλοκή μπορεί να μεταφραστεί ως ισοδύναμη με τον μέσο όρο των διαφορετικών λέξεων που απαιτείται να διαχωριστούν σε κάθε σημείο της ακολουθίας, εάν όλες οι λέξεις σε κάθε σημείο είναι εξίσου πιθανές. Η περιπλοκή πολλές φορές παριστάνει έναν μέσο συντελεστή διακλάδωσης εναλλακτικών λέξεων. Η μικρότερη τιμή περιπλοκής είναι ίση με το 1, αλλά αυτή θα μπορούσε να προκύψει μόνο στην περίπτωση που όλες οι πιθανές λέξεις είχαν πιθανότητα 1, με αποτέλεσμα μόνο αυτή η ακολουθία λέξεων να μπορεί να αναγνωριστεί.

Όσο μικρότερη είναι η περιπλοκή, τόσο καλύτερο είναι το γλωσσικό μοντέλο που εξετάζουμε. Σε ένα μοντέλο n-γραμμάτων, μπορεί επίσης να θεωρήσει κανείς την περιπλοκή και τη διασταυρωμένη εντροπία ως ένα μέτρο του κατά πόσον το γλωσσικό μοντέλο προβλέπει με βεβαιότητα την κάθε λέξη, δεδομένων των προηγούμενων N–1.

Η περιπλοκή είναι ένα χρήσιμο εργαλείο για τη σύγκριση διαφορετικών γλωσσικών μοντέλων. Ωστόσο, δεν πρέπει να ξεχνάμε ότι το τελικό πείραμα πρέπει να γίνει βάσει του ποσοστού επιτυχίας όλου του συστήματος μηχανικής μετάφρασης.

8.5 Δημιουργία Γλωσσικών Μοντέλων

Προκειμένου να δημιουργήσουμε Γλωσσικά Μοντέλα, χρησιμοποιήσαμε την εργαλειοθήκη SRILM[1] (Stolcke, 2002). Η εργαλειοθήκη SRILM χρησιμοποιεί εξ ορισμού εξομάλυνση οπισθοδρόμησης χωρίς παρεμβολή (backoff smoothing without interpolation) (Chen and Goodman, 1999, Katz, 1987), ενώ υποστηρίζει και διάφορες μεθόδους έκπτωσης (discount methods). Στα πειράματά μας, τα καλύτερα αποτελέσματα έδωσε η μέθοδος έκπτωσης των Witten-Bell (Witten and Bell, 1991). Οι παρεμβαλλόμενες (Interpolated) εκδόσεις των περισσότερων μοντέλων έκπτωσης είναι επίσης διαθέσιμες, χρησιμοποιώντας επιπλέον την επιλογή -interpolate. Οι εξισώσεις για όλες τις μεθόδους εξομάλυνσης (οπισθοδρόμησης ή παρεμβολής) μπορούν να αναζητηθούν στην ιστοσελίδα του SRILM. Συγκεκριμένα, εξετάζονται τα ακόλουθα σώματα της ΕΝΓ:

- Greek plain text (input plain text)
- GSL glosses annotated with NmCs tags
- GSL glosses annotated with NmCs tags (factored mode)
- GSL glosses without any tags
- GSL glosses without any tags (factored mode)

Στο σημείο αυτό θα πρέπει να διευκρινιστεί ότι σε όλες τις καταστάσεις ομαδοποίησης (factory mode), όλα τα σχετικά λήμματα ομαδοποιούνται, δηλαδή αντικαθιστούμε όλα τα ψηφία/αριθμούς με το λήμμα “DIGIT”, όλες τις ημερομηνίες με το λήμμα “DATE”, όλους τους μήνες με το λήμμα “MONTH” κλπ.

Τα γλωσσικά μοντέλα (ΓΜ) με τις χαμηλότερες τιμές περιπλοκής (perplexity) (Koehn et al., 2003) (Ενότητα 8.4) θεωρούνται καλύτερα από εκείνα που εμφανίζουν υψηλότερες τιμές περιπλοκής. Στην παρούσα εργασία εξετάζουμε τα γλωσσικά μοντέλα που δημιουργήσαμε από τα παραπάνω είδη σωμάτων της ΕΝΓ, για να μελετήσουμε ποιος τύπος σώματος παρέχει γλωσσικό μοντέλο με τις καλύτερες τιμές περιπλοκής και επομένως την καλύτερη πρόβλεψη n-γραμμάτων. Για τη δημιουργία των γλωσσικών μοντέλων και τον υπολογισμό της περιπλοκής τους, χρησιμοποιήσαμε την εργαλειοθήκη SRILM. Επειδή η περιπλοκή εξαρτάται από το λεξιλόγιο του σώματος κατάρτισης γλωσσικού μοντέλου καθώς και το σύνολο των σωμάτων δοκιμής, δεν θεωρείται καλή πρακτική να συγκρίνουμε διαφορετικούς τύπους σωμάτων με τη μέθοδο της περιπλοκής, γιατί τα γλωσσικά μας

[1] http://www.speech.sri.com/projects/srilm/

μοντέλα αναφέρονται κάθε φορά σε διαφορετικούς τύπους σωμάτων και κατά συνέπεια σε διαφορετικά λεξιλόγια.

Για να αντιμετωπιστεί αυτό το ζήτημα, διερευνήθηκαν δύο κατευθύνσεις. Αρχικά χρησιμοποιήσαμε μια κανονικοποιημένη μορφή περιπλοκής (normalized perplexity - P_{norm}) (Bunke and Caelli (2001)), η οποία εκφράζεται ως

$$P_{norm} = \frac{P}{n} \tag{8.37}$$

όπου P είναι η περιπλοκή του ΓΜ και n είναι ο αριθμός των λέξεων του υπό εξέταση λεξιλογίου. Στη συνέχεια, εκτιμήθηκαν οι συντελεστές διάκρισης (discrimination coefficients) (Delić et al. (2013), Ostrogonac et al. (2012)). Ένας συντελεστής διάκρισης αντιπροσωπεύει μια ποσοτική περιγραφή (quantitative description - KD) της ικανότητας του γλωσσικού μοντέλου να διακρίνει μεταξύ του αυθεντικού κειμενικού πλαισίου και του περιεχομένου κειμένου που δεν φέρει σημασιολογικές πληροφορίες. Υπολογίζεται ως ο αντίστροφος λόγος της περιπλοκής που υπολογίστηκε στο αρχικό κείμενο και της περιπλοκής που υπολογίστηκε στο κείμενο που δημιουργήθηκε αφού ανακατέψαμε τυχαία τη σειρά των λέξεων σε κάθε πρόταση του αρχικού κειμένου. Στην Εξίσωση 8.38 το ppl είναι η τιμή της περιπλοκής, ενώ ο KD είναι ο συντελεστής διάκρισης (discrimination coefficient). Οι τιμές KD είναι πρακτικά ασυσχέτιστες με το σύνολο δεδομένων δοκιμών που χρησιμοποιήθηκε στην αξιολόγηση, γεγονός που τις καθιστά πιο κατάλληλες για τη σύγκριση των γλωσσικών μοντέλων (LM).

$$KD = \frac{ppl(randomizedtext)}{ppl(originaltext)} \tag{8.38}$$

Υπάρχουν επίσης και άλλες τεχνικές (Chen et al., 2004, Lin et al., 1997, Moore and Lewis, 2010) που προσπαθούν να επιλύσουν το πρόβλημα των διαφόρων μεγεθών λεξιλογίου. Οι περισσότερες από αυτές εστιάζουν στη μείωση του μεγέθους των σωμάτων ή στην αύξηση της περιπλοκής ενός ΓΜ σώματος εντός θέματος (In-topic), παρεμβάλλοντας (interpolating) τα καλύτερα μέρη σωμάτων εκτός θέματος (out-of-topic).

8.6 Αποτελέσματα ΓΜ από άλλες ΝΓ

Σε αυτήν την υποενότητα θα παρουσιάσουμε τα αποτελέσματα γλωσσικών μοντέλων από άλλες ΝΓ. Στη μελέτη του Stein (Stein et al., 2006), οι συγγραφείς δημιούργησαν ΓΜ από τα σώματα του Phoenix, χρησιμοποιώντας την εξομάλυνση εκπτώσεων Kneser-Ney. Σε αυτή την περίπτωση, η περιπλοκή τους εκτιμήθηκε στο 53,3 για 3-γράμματα (3-grams) και στο 30,5 για την περίπτωση του γερμανικού σώματος. Στη μελέτη του Morrissey

(2008), εξετάστηκε το σώμα ATIS. Σε αυτή την περίπτωση τα αποτελέσματα περιπλοκής που βρέθηκαν ήταν (α) 15,7 για το αγγλικό κείμενο (3-grams), (β) 12,4 για το γερμανικό σώμα, (γ) 28,3 για το σώμα της Ιρλανδικής ΝΓ (Irish SL corpus) και (δ) 11,39 για το σώμα της Γερμανικής ΝΓ (DGS corpus). Στη μελέτη του Dreuw et al. (2008), εξετάστηκε το σώμα ATIS. Σε αυτή την περίπτωση τα αποτελέσματα περιπλοκής που βρέθηκαν ήταν (α) 15,7 για το αγγλικό κείμενο (3-grams), (β) 12,4 για το γερμανικό σώμα, (γ) 28,3 για το σώμα Ιρλανδικής ΝΓ (Irish SL corpus) και (δ) 11,39 για το σώμα της Γερμανικής ΝΓ (DGS corpus). Στη μελέτη του Dreuw et al. (2008), εξετάστηκαν 3-γράμματα στην περίπτωση των σωμάτων RWTH-BOSTON. Η εργαλειοθήκη SRILM ενσωματώθηκε και εξετάστηκε μια τροποποιημένη έκπτωση της Kneser-Ney με παρεμβολή (interpolation). Για τα 3-γράμματα, η περιπλοκή που βρέθηκε για τα σώματα RWTH-BOSTON ήταν 30,1 για το σώμα ανάπτυξης και 25,1 για το σώμα δοκιμής. Τέλος, οι συγγραφείς San-Segundo et al. (2008) δημιούργησαν ΓΜ από σώματα της ισπανικής γλώσσας και Ισπανικής ΝΓ (LSE). Κατά τη διάρκεια της εκπαίδευσης, χρησιμοποιήθηκαν 226 προτάσεις, εκ των οποίων οι 150 χρησιμοποιήθηκαν για τη δοκιμή. Στην περίπτωση των 3-γραμμάτων, η τιμή περιπλοκλής ppl ήταν 15,4 για την ισπανική γλώσσα και 10,7 για την Ισπανική ΝΓ (LSE). Τα αποτελέσματα περιπλοκής (ppl), συνοψίζονται στον Πίνακα 8.2.

Corpus	Language	PPL 3-grams	Reference
RWTH-Phoenix	DGS German	53.3 30.5	Stein et al. (2006)
ATIS	English German ISL DGS	15.7 12.4 28.3 11.39	Morrissey et al. (2007)
RWTH-BOSTON	DGS Development set DGS Test set	30.1 25.1	Dreuw et al. (2008)
-	Spanish LSE	15.4 10.7	San-Segundo et al. (2008)

Πίνακας 8.2 PPL results of LMs from other corpora

8.7 Αξιολόγηση, αποτελέσματα και συζήτηση

Για τα πειράματα της εργασίας μας δημιουργήσαμε 10 υποσώματα για κάθε τύπο λέξης (token). Σε κάθε φάση εκπαίδευσης προσθέτουμε κάθε φόρα ένα υποσώμα (π.χ.

στην φάση Train1 χρησιμοποιήθηκε το 1ο υποσώμα, στο Train2 χρησιμοποιήθηκαν το 1ο και το 2ο υποσώμα, στο Train3 το 1ο, 2ο και 3ο υποσώμα κλπ.).

Στον Πίνακα 8.3 συνοψίζονται τα στατιστικά στοιχεία σχετικά με τα μεγέθη των σωμάτων για κάθε τύπο κορμού.

Επιπλέον, δημιουργήσαμε γλωσσικά μοντέλα n-γραμμάτων (n-gram LMs), χρησιμοποιώντας τα ακόλουθα σύνολα σωμάτων εκπαίδευσης:

- ΕΝΓ glosses (σύντομες γραπτές μεταγραφές-ΣΓΜ) χαρακτηρισμένες μόνο με ετικέτες NmCs (Σχήμα 8.1) και σε κατάσταση ομαδοποίησης (factored mode) (Σχήμα 8.2).

- ΕΝΓ glosses μη χαρακτηρισμένες με ετικέτες (Σχήμα 8.3) και σε κατάσταση ομαδοποίησης (factored mode) (Σχήμα 8.4).

ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΧΡΙΣΤΟΥΓΕΝΝΑ ΕΧΕΙ ΒΡΟΧΗ ΚΑΙ ΚΑΤΑΙΓΙΔΑ/ΜΧ(ΕΝΤΑΣΗ)/ΜΓΛ(ΦΟΥΣΚΩΜΕΝΑ)

Σχήμα 8.1 ΕΝΓ glosses με ετικέτες NmCs

ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΘΕΡΜΟΚΡΑΣΙΑ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΦΤΑΝΩ ΜΕΓΑΛΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΗΠΕΙΡΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) **DIGIT** ΕΩΣ **DIGIT** ΒΑΘΜΟΣ ΛΙΓΟΣ ΒΡΟΧΗ ΤΟΠΙΚΟΣ

Σχήμα 8.2 ΕΝΓ glosses με ετικέτες NmCs (factored mode)

ΚΑΚΟΚΑΙΡΙΑ ΜΕΤΑ ΔΙΑΡΚΩ ΕΩΣ ΑΝΤ_3 **ΔΕΥΤΕΡΑ** ΑΝΤ_3

Σχήμα 8.3 ΕΝΓ glosses χωρίς ετικέτες

ΚΑΚΟΚΑΙΡΙΑ ΜΕΤΑ ΔΙΑΡΚΩ ΕΩΣ ΑΝΤ_3 **ΔΑΥ** ΑΝΤ_3

Σχήμα 8.4 ΕΝΓ glosses χωρίς ετικέτες (factored mode)

Οι τιμές περιπλοκής για όλα τα σώματα παρουσιάζονται στον Πίνακα 8.4. Ο Πίνακας 8.5 παρέχει τις ίδιες πληροφορίες με τον Πίνακα 8.4, με τη διαφορά ότι στην περίπτωση αυτή έχουμε τυχαιοποίηση της σειράς των λέξεις του κειμένου σε επίπεδο προτάσεων. Στον Πίνακα 8.6 παρουσιάζονται οι συντελεστές διάκρισης (discrimination coefficients), ενώ στον Πίνακα 8.7 παρουσιάζεται η κανονικοποιημένη περιπλοκή (normalized perplexity)

	Source Greek Plain Text	Source Greek Plain Text	GSL glossed with NmCs	GSL glossed with NmCs	GSL glossed with NmCs (factored)	GSL glossed with NmCs (factored)	GSL glossed without NmCs	GSL glossed without NmCs	GSL glossed without NmCs (factored)	GSL glossed without NmCs (factored)
	Indivi-dual words	Words	Indivi-dual words	Words	Individual words	Words	Indivi-dual words	Words	Individual	Words
Cor-pus parts	448	2471	368	2298	317	2243	347	2283	297	224
Eval	262	1,582	211	2,165	178	2,076	21	1,515	176	1,433
Train1	2501	3,945	366	5,483	327	533	365	3,757	325	3,634
Train2	621	6149	467	8449	43	8,217	466	5,859	427	5,689
Train3	1,030	8,637	770	12,219	713	11,863	769	8,611	711	8,349
Train4	1,124	10,886	851	14,329	783	13,953	850	10,143	783	9,862
Train5	1,447	138	866	15,947	798	15,571	865	11,301	796	11,020
Train6	1,567	15,473	874	17,439	806	17,063	873	12,385	804	12,104
Train7	1,578	16,704	881	18,760	811	18,382	880	13,362	822	13,029
Train8	1,588	17,855	913	21,086	845	20,700	912	14,998	841	14,710
Train9										

Πίνακας 8.3 Statistics Regarding Corpora Sizes

Corpus parts	**GSL glossed with NmCs**	**GSL glossed with NmCs (factored)**	**GSL glossed without NmCs**	**GSL glossed without NmCs (factored)**
Train1	44.28	35.55	40.49	29.02
Train2	46.42	37.06	39.55	27.35
Train3	43.19	37.04	34.42	24.98
Train4	37.89	32.49	28.69	21.06
Train5	36.81	32.93	27.67	20.93
Train6	36.11	33.17	27.12	20.93
Train7	36.21	34.04	27.05	21.34
Train8	36.74	34.70	27.43	21.80
Train9	31.10	30.00	22.43	18.08

Πίνακας 8.4 Perplexity values on authentic text

Corpus parts	**GSL glossed with NmCs**	**GSL glossed with NmCs (factored)**	**GSL glossed without NmCs**	**GSL glossed without NmCs (factored)**
Train1	91.81	80.36	94.17	66.33
Train2	108.22	87.81	108.23	84.46
Train3	115.30	90.97	116.89	84.95
Train4	135.68	106.41	130.27	100.69
Train5	133.20	105.55	124.82	100.86
Train6	125.52	95.09	134.90	91.06
Train7	125.18	97.53	125.49	94.16
Train8	122.44	102.46	129.24	93.75
Train9	121.59	101.19	124.85	95.76

Πίνακας 8.5 Perplexity values on text with random word order on a sentence level

Corpus parts	**GSL glossed with NmCs**	**GSL glossed with NmCs (factored)**	**GSL glossed without NmCs**	**GSL glossed without NmCs (factored)**
Train1	2.07	2.26	2.33	2.29
Train2	2.33	2.37	2.74	3.09
Train3	2.67	2.46	3.40	3.40
Train4	3.58	3.27	4.54	4.78
Train5	3.62	3.21	4.51	4.82
Train6	3.48	2.87	4.97	4.35
Train7	3.46	2.87	4.64	4.41
Train8	3.33	2.95	4.71	4.30
Train9	3.91	3.37	5.57	5.30

Πίνακας 8.6 Discrimination Coefficient values

Corpus parts	GSL glossed with NmCs	GSL glossed with NmCs (factored)	GSL glossed without NmCs	GSL glossed without NmCs (factored)
Train1	20.45	17.13	26.73	20.25
Train2	8.47	6.95	10.53	7.53
Train3	5.11	4.51	5.87	4.39
Train4	3.10	2.74	3.33	2.52
Train5	2.57	2.36	2.73	2.12
Train6	2.26	2.13	2.40	1.90
Train7	2.08	1.99	2.18	1.76
Train8	1.96	1.89	2.05	1.67
Train9	1.47	1.45	1.50	1.23

Πίνακας 8.7 Normalized perplexity (P_{norm})

Corpus parts	Greek source plain text
Train1	39.23
Train2	34.55
Train3	31.31
Train4	31.93
Train5	28.48
Train6	23.68
Train7	23.00
Train8	22.85
Train9	22.47

Πίνακας 8.8 Greek plain text Perplexity values

για όλα τα ΓΜ. Τέλος, στον Πίνακα 8.8 παρουσιάζονται οι περιπλοκές των ελληνικών σωμάτων κειμένου που χρησιμοποιήσαμε ως είσοδο για το σύστημα ΜΜ κανόνων (RBMT).

Όπως μπορεί να παρατηρηθεί από τους παραπάνω πίνακες, όλα τα ΓΜ αντικατοπτρίζουν επιτυχώς τη γλώσσα που αντιπροσωπεύουν, ενώ ταυτόχρονα αποδεικνύεται ότι μπορούν να χρησιμοποιηθούν σε γλωσσικές εφαρμογές όπως ένα σύστημα στατιστικής μηχανικής μετάφρασης (SMTS). Επιπλέον, οι δείκτες του συντελεστή διάκρισης και Pnorm παρέχουν μια πιο αντικειμενική αξιολόγηση της ποιότητας των παραπάνω ΓΜ. Όπως φαίνεται από τους Πίνακες 8.6 και 8.5, το ΓΜ που βασίζεται στο σώμα χωρίς καμία ετικέτα δίνει καλύτερα αποτελέσματα. Αυτό οφείλεται κυρίως στο γεγονός ότι το συγκεκριμένο ΓΜ χρησιμοποιεί πιο περιορισμένο λεξιλόγιο και μπορεί να επιτύχει καλύτερες εκτιμήσεις πιθανότητας ακολουθίας λέξεων. Σε κάθε περίπτωση, πρέπει να τονιστεί ότι οι ετικέτες NMCs είναι αναμφισβήτητα πολύ σημαντικά χαρακτηριστικά της ΕΝΓ (Efthimiou et al., 2016, Fotinea et al., 2005, Kouremenos et al., 2010). Από την άλλη πλευρά, μπορεί επίσης να παρατηρηθεί ότι τα ΓΜ στα οποία έχουμε ομαδοποιήσεις λέξεων (factory mode) στα σώματα, δίνουν καλύτερες τιμές περιπλοκής, ειδικότερα σε μικρά μεγέθη σωμάτων. Αυτή η βελτίωση περιορίζεται όταν αυξάνεται το μέγεθος του σώματος, γεγονός που ερμηνεύεται από το ότι ο τρόπος εφαρμογής της ομαδοποίησης των λέξεων μειώνει το μέγεθος του λεξιλογίου, καθώς αντικαθιστούμε τις θεματικές λέξεις με ομάδες (DIGITS, DATE, MONTH, PLACE κ.λπ.).

Κεφάλαιο 9

Στατιστική Μηχανική Μετάφραση (SMT - MOSES)

9.1 Στατιστική μηχανική μετάφραση

Η έρευνα της στατιστική μηχανική μετάφραση ξεκίνησε στα τέλη της δεκαετίας του 1980 από την IBM με το πρόγραμμα Candide. Η αρχική προσέγγιση της IBM ήταν η χαρτογράφηση μεμονωμένων λέξεων, επιτρέποντας τη διαγραφή και εισαγωγή λέξεων.

Οι παράλληλες προτάσεις ευθυγραμμίζονται σε επίπεδο λέξης, χρησιμοποιώντας συνήθως το λογισμικό GIZA++[1], το οποίο υλοποιεί ένα σύνολο στατιστικών μοντέλων που αναπτύχθηκαν στην IBM στη δεκαετία του '80. Αυτές οι ευθυγραμμίσεις λέξεων χρησιμοποιούνται για την εξαγωγή μεταφράσεων φράσης-φράσης ή ιεραρχικών κανόνων, όπως απαιτείται, και οι στατιστικές αναλύσεις σε ολόκληρα τα σώματα εφαρμόζονται για την εκτίμηση πιθανοτήτων. Τα στατιστικά μοντέλα που βασίζονται σε φράσεις μεταφράζουν φράσεις ως ατομικές μονάδες.

Ένα σημαντικό μέρος του συστήματος της στατιστικής μετάφρασης είναι τα γλωσσικά μοντέλα. Πρόκειται για στατιστικά μοντέλα που κατασκευάζονται με τη χρήση μονογλωσσικών δεδομένων της γλώσσας-στόχου και χρησιμοποιούνται από τον αποκωδικοποιητή της ΜΜ για να εξασφαλιστεί η ευχέρεια/ορθότητα της παραγωγής.

Τον τελευταίο καιρό, διάφοροι ερευνητές έχουν παρουσιάσει συστήματα μετάφρασης καλύτερης ποιότητας με χρήση της μηχανικής μετάφρασης που βασίζεται σε φράσεις. Το σύστημα ΜΜ που βασίζεται σε φράσεις μπορεί να ανιχνευθεί στο μοντέλο προτύπου ευθυγράμμισης φράσεων του Och (Och and Ney, 2000), το οποίο μπορεί να ανα-

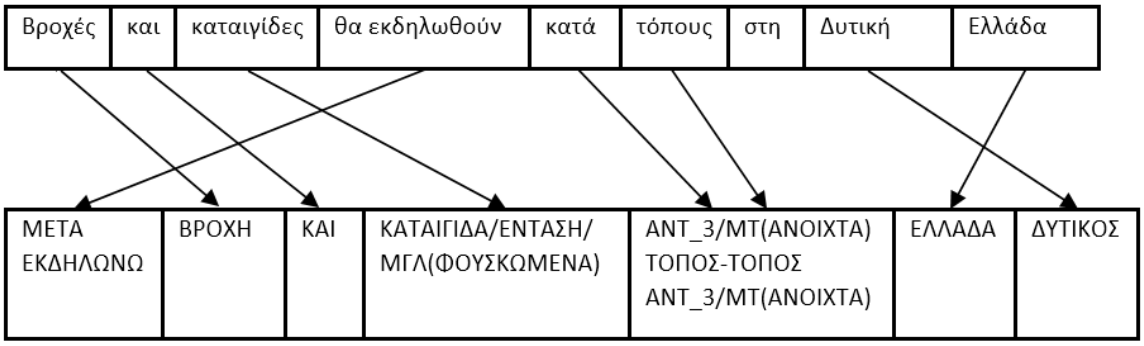
[1] http://www.isi.edu/~och/GIZA++.html

διαμορφωθεί ως σύστημα μετάφρασης φράσεων. Άλλοι ερευνητές χρησιμοποίησαν συστήματα με φωνητική μετάφραση, όπως το Yamada, που βασίζονται στη σύνταξη.

Υπάρχουν επίσης και άλλοι αλγόριθμοι για τη στατιστική μηχανική μετάφραση. Γενικώς, υπάρχει μια προσπάθεια να αναπτυχθούν μοντέλα που βασίζονται στη σύνταξη και που είτε χρησιμοποιούν πραγματικά δέντρα σύνταξης, τα οποία παράγονται από συντακτικούς αναλυτές, είτε χρησιμοποιούν μεθόδους μεταφοράς δέντρων που καθοδηγούνται από συντακτικά σχέδια αναδιάταξης (syntactic reordering patterns). Το μοντέλο στατιστικής μηχανικής μετάφρασης με βάση τη φράση που παρουσιάζουμε εδώ καθορίστηκε από τους Koehn (2004), ενώ μπορείτε να δείτε επίσης και την περιγραφή του Zens et al. (2002). Οι εναλλακτικές μέθοδοι που βασίζονται σε φράσεις διαφέρουν ως προς τον τρόπο με τον οποίο δημιουργείται ο πίνακας μετάφρασης φράσεων, τον οποίο συζητούμε λεπτομερώς παρακάτω.

9.1.1 Το μοντέλο φράσεων (Phrase Model)

Στο παρακάτω Σχήμα 9.1 απεικονίζεται η διαδικασία της μετάφρασης που βασίζεται στη φράση. Η είσοδος χωρίζεται σε διάφορες ακολουθίες διαδοχικών λέξεων (λεγόμενες φράσεις). Κάθε φράση μεταφράζεται σε μια μεταγραφή της ΕΝΓ, η οποία στην έξοδο μπορεί να αναδιαταχθεί.



Σχήμα 9.1 Illustration of the process of phrase-based translation

Σε αυτή την ενότητα, θα καθορίσουμε τυπικά το μοντέλο μηχανικής μετάφρασης με βάση τη φράση. Το μοντέλο μετάφρασης φράσης βασίζεται στο μοντέλο καναλιού-θορύβου (noisy channel model). Χρησιμοποιούμε τον κανόνα Bayes για να επαναδιατυπώσουμε την πιθανότητα μετάφρασης για την μετάφραση μιας ελληνικής πρότασης **f** σε πρόταση μεταγραφής της ΕΝΓ **e** ως:

$$argmax_e p(\mathbf{e}|\mathbf{f}) = argmax_e p(\mathbf{f}|\mathbf{e}) p(\mathbf{e})$$

Αυτό επιτρέπει ένα γλωσσικό μοντέλο $p(\mathbf{e})$ και ένα ξεχωριστό μοντέλο μετάφρασης $p(\mathbf{f}|\mathbf{e})$.

Κατά τη διάρκεια της αποκωδικοποίησης, η ξένη πρόταση $\mathbf{f}$ εισάγεται σε μια ακολουθία από I φράσεις $\mathbf{f}_1^I$. Υποθέτουμε μια ομοιόμορφη κατανομή πιθανότητας σε όλες τις πιθανές κατατμήσεις.

Κάθε ελληνική φράση $\mathbf{f}_i$ σε $\mathbf{f}_1^I$ μεταφράζεται σε μια φράση της ΕΝΓ $\mathbf{e}_i$. Η φράση της ΕΝΓ μπορεί να αναδιαταχθεί. Η μετάφραση φράσης διαμορφώνεται από κατανομή πιθανότητας $\boldsymbol{\varphi}(f_i|e_i)$. Να υπενθυμίσουμε ότι λόγω του κανόνα της Bayes, η κατεύθυνση της μετάφρασης είναι ανεστραμμένη από άποψη μοντελοποίησης.

Η αναδιάταξη των φράσεων εξόδου μεταγραφής της ΕΝΓ διαμορφώνεται από μια σχετική κατανομή πιθανότητας παραμόρφωσης (relative distortion probability distribution) $\mathbf{d}(start_i, end_{i-1})$, όπου το $start_i$ δηλώνει την αρχική θέση της Ελληνικής πρότασης η οποία μεταφράζεται στην ith φράση ΕΝΓ, και το end_{i-1} δηλώνει την τελευταία θέση της φράσης της Ελληνικής πρότασης η οποία μεταφράζεται στην $(i-1)$th φράση της ΕΝΓ.

Χρησιμοποιούμε ένα απλό μοντέλο παραμόρφωσης $\mathbf{d}(start_i, end_{i-1}) = \boldsymbol{\alpha}^{|start_i - end_{i-1} - 1|}$ με την κατάλληλη τιμή για την παράμετρο $\boldsymbol{\alpha}$.

Για να βαθμονομήσουμε το μήκος εξόδου, εισάγουμε έναν παράγοντα $\boldsymbol{\omega}$, τον οποίο καλούμε ως κόστος λέξης, για κάθε παραγόμενη μεταγραφή της ΕΝΓ εκτός από το γλωσσικό μοντέλο τρι-γραμμάτων (trigram language model) $\mathbf{p}_{LM}$. Αυτό είναι ένα απλό μέσο για τη βελτιστοποίηση της απόδοσης. Συνήθως, αυτός ο παράγοντας είναι μεγαλύτερος από 1 και επηρεάζει τη μεγαλύτερη παραγωγή.

Εν ολίγοις, η καλύτερη πρόταση εξόδου της μεταγραφής της ΕΝΓ $\mathbf{e}_{best}$, δεδομένης της ελληνικής πρότασης $\mathbf{f}$ σύμφωνα με το μοντέλο μας είναι:

$$\mathbf{e}_{best} = argmax_e p(\mathbf{e}|\mathbf{f}) = argmax_e p(\mathbf{f}|\mathbf{e}) p_{LM}(\mathbf{e}) \boldsymbol{\omega}^{length(\mathbf{e})}$$

όπου το $p(\mathbf{f}|\mathbf{e})$ αποσυντίθεται σε:

$$p(\mathbf{f}_1^I|\mathbf{e}_1^I) = \Phi_{i=1}^I \boldsymbol{\varphi}(f_i|e_i) \mathbf{d}(start_i, end_{i-1})$$

9.1.2 Ευθυγράμμιση λέξεων (Word Alignment)

Κατά τη μέχρι τώρα περιγραφή του μοντέλου μετάφρασης με φράση, δεν αναφερθήκαμε στον τρόπο απόκτησης των παραμέτρων του μοντέλου, ειδικά του πίνακα μεταφράσεων πιθανότητας φράσης που χαρτογραφεί ξένες φράσεις στις φράσεις της ΕΝΓ.

Οι πιο πρόσφατες δημοσιευμένες μέθοδοι για την εξαγωγή ενός πίνακα μετάφρασης φράσεων από ένα παράλληλο σώμα ξεκινούν με την ευθυγράμμιση λέξεων. Η ευθυγράμ-

μιση λέξεων είναι ένα ενεργό ερευνητικό θέμα. Βλέπουμε για παράδειγμα, ότι το θέμα αυτό ήταν το επίκεντρο ως κοινό έργο σε ένα πρόσφατο συνέδριο μεταφραστικής μηχανικής μετάφρασης. Μπορείτε επίσης να ανατρέξετε τις συστημικές συγκρίσεις μεταξύ των Och και Ney (Och and Ney, 2003).

Το πιο συνηθισμένο εργαλείο για τη ευθυγράμμιση λέξεων είναι η χρήση της εργαλειοθήκης GIZA++[1]. Αυτή η εργαλειοθήκη είναι μια εφαρμογή των αρχικών μοντέλων της IBM, που ξεκίνησε με την έρευνα στατιστικής μηχανικής μετάφρασης. Ωστόσο, αυτά τα μοντέλα έχουν κάποιες σοβαρές ελλείψεις. Η πιο σημαντική είναι ότι επιτρέπουν μόνο μία λέξη της γλώσσας-πηγή να ευθυγραμμίζεται με κάθε λέξη της γλώσσας-στόχου. Για να επιλυθεί το πρόβλημα αυτό, εφαρμόζονται μερικοί μετασχηματισμοί.

Αρχικά, το παράλληλο σώμα είναι ευθυγραμμισμένο αμφίδρομα, π.χ. η ελληνική μεταγραφή της ΕΝΓ προς τα ελληνικά. Το γεγονός αυτό δημιουργεί την ανάγκη να συμφωνηθούν δύο ευθυγραμμίσεις λέξεων. Εάν διασταυρώσουμε τις δύο ευθυγραμμίσεις, έχουμε μια ευθυγράμμιση υψηλής ακρίβειας των σημείων και πετυχαίνουμε μια ευθυγράμμιση υψηλής εμπιστοσύνης. Εάν λάβουμε την ένωση των δύο ευθυγραμμίσεων, έχουμε μια ευθυγράμμιση με υψηλή ανάκληση (high-recall) με επιπλέον σημεία ευθυγράμμισης (alignment points).

9.1.3 Εκμάθηση πίνακα μετάφρασης φράσεων (Learning a Phrase Translation Table)

Δεδομένης αυτής της ευθυγράμμισης, είναι αρκετά εύκολο να εκτιμηθεί ένας πίνακας λεξικών μεταφράσεων μέγιστης πιθανότητας (maximum likelihood). Υπολογίζουμε τον πίνακα μεταφράσεων λέξεων $w(e|f)$ καθώς και τον αντίστροφο $w(f|e)$. Στο παρακάτω σχήμα θα δείτε μερικές μεταφράσεις φράσεων ελληνικού κειμένου προς την μεταγραφή της ΕΝΓ (GSL Gloss) (Σχήμα 9.2).

9.1.4 Φράσεις εξαγωγής και βαθμολόγησης (Extract and Score Phrases)

Στο στάδιο εξαγωγής φράσης όλες οι φράσεις τοποθετούνται/εξάγονται (dumped) σε ένα μεγάλο αρχείο. Στο παρακάτω σχήμα βλέπουμε την αρχή του αρχείου (Σχήμα 9.3).

Το περιεχόμενο αυτού του αρχείου είναι για κάθε γραμμή το εξής: ελληνική φράση, φράση μεταγραφής ΕΝΓ (GSL Gloss) και σημεία ευθυγράμμισης. Τα σημεία ευθυγράμμισης είναι ζεύγη (ελληνικής, ΕΝΓ gloss). Επίσης, δημιουργείται ένα αρχείο ανεστραμμένης ευθυγράμμισης στο αρχείο extract.inv και αν το (προεπιλεγμένο) μοντέλο λεξικής

[1] http://www.isi.edu/~och/GIZA++.html

$ grep 'θερμοκρα' lex.e2f | sort -nrk 4 | head
θερμοκρασίες ΨΗΛΟΣ 0.3076923
θερμοκρασίες ΘΕΡΜΟΚΡΑΣΙΑ 0.0644391
θερμοκρασίες ΕΠΟΧΗ 0.2222222
θερμοκρασίες ΑΡΝΗΤΙΚΟΣ 0.2500000
θερμοκρασίας ΘΕΡΜΟΚΡΑΣΙΑ 0.0357995
θερμοκρασίας ΑΡΝΗΤΙΚΟΣ 0.2500000
θερμοκρασία ΘΕΡΜΟΚΡΑΣΙΑ 0.4415274
$ grep 'της' lex.e2f | sort -nrk 4 | head
Τρίτης ΤΡΙΤΗ 0.1333333
της ΚΤΗΤ_3 0.6073826
της ΚΤΗΤ_2 0.0800000
της ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) 0.0016474
Τετάρτης ΤΕΤΑΡΤΗ 0.0909091
Πέμπτης ΠΕΜΠΤΗ 0.1034483
Κρήτης ΚΡΗΤΗ 0.0689655
ανοχύρωτης ΑΝΟΧΥΡΩΤΟΣ 0.5000000

Σχήμα 9.2 word translation table

> head model/extract.sorted
! ! περισσότερες πληροφορίες ||| ! ! ΠΛΗΡΟΦΟΡΙΑ ΠΟΛΥΣ ||| 0-0 1-1 3-2 2-3
! ! περισσότερες πληροφορίες στις ||| ! ! ΠΛΗΡΟΦΟΡΙΑ ΠΟΛΥΣ ΚΤΗΤ_3 ||| 0-0 1-1 3-2 2-3 4-4
! ! περισσότερες πληροφορίες στις προγνώσεις ||| ! ! ΠΛΗΡΟΦΟΡΙΑ ΠΟΛΥΣ ΚΤΗΤ_3 ΠΡΟΓΝΩΣΗ ||| 0-0 1-1 3-2 2-3 4-4 5-5

Σχήμα 9.3 extract.sorted file

αναδιάταξης εκπαιδευτεί (lexicalized reordering) τότε δημιουργείται και ένα αρχείο αναδιάταξης extract.o.

Στη συνέχεια, δημιουργείται ένας πίνακας μετάφρασης από τα αποθηκευμένα ζεύγη μετάφρασης φράσεων. Τα δύο βήματα διαχωρίζονται επειδή, για μεγαλύτερα μοντέλα μετάφρασης, ο πίνακας μετάφρασης φράσης δεν ταιριάζει στη μνήμη. Δεν πρέπει ποτέ να αποθηκεύουμε τους πίνακες των φράσεων μετάφρασης στη μνήμη.

Για να υπολογίσουμε την πιθανότητα μετάφρασης φράσης φ(e|f) προχωρούμε ως εξής: Αρχικά, ταξινομείται το αρχείο εξόδου για να διασφαλίσουμε ότι όλες οι μεταφράσεις της ΕΝΓ για μια ελληνική φράση βρίσκονται στο άλλο αρχείο. Έτσι, μπορούμε να επεξεργαστούμε από το αρχείο μία ελληνική φράση κάθε φορά, να συλλέξουμε μετρήσεις και να υπολογίσουμε τη πιθανότητα φ (e|f) για αυτή την ελληνική φράση f. Για να υπολογίσουμε την πιθανότητα φ (f|e), το ανεστραμμένο αρχείο ταξινομείται και στη συνέχεια η πιθανότητα φ (f|e) υπολογίζεται για μία φράση της ΕΝΓ κάθε φορά.

Δίπλα από τις κατανομές πιθανότητας μεταφράσεων φ(f|e) και φ(e|f), μπορούν να υπολογιστούν πρόσθετες λειτουργίες μέτρησης (βαθμολόγησης) της μετάφρασης φράσης, π.χ. λεκτικό βάρος (lexical weighting), λεκτική ποινή (word penalty), ποινή φράσης (phrase penalty) κλπ. Στις προαναφερθείσες λειτουργίες μέτρησης προστίθεται η λεκτική στάθμιση (lexical weighting) και για τις δύο κατευθύνσεις, καθώς και η ποινή φράσης (phrase penalty).

Στην παρούσα μελέτη υπολογίζονται τέσσερις διαφορετικές μετρήσεις μετάφρασης φράσης:

1. αντίστροφη πιθανότητα μετάφρασης φράσης (inverse phrase translation probability) φ(f|e)

2. αντίστροφη λεξική στάθμιση (inverse lexical weighting) lex(f|e)

3. άμεση πιθανότητα μετάφρασης φράσης (direct phrase translation probability) φ(e|f)

4. άμεση λεκτική στάθμιση (direct lexical weighting) lex(e|f)

9.1.5 Αναδιάταξη και δημιουργία μοντέλου (Reordering and generate model)

Από προεπιλογή, στην τελική διαμόρφωση περιλαμβάνεται μόνο ένα μοντέλο αναδιάταξης που βασίζεται στην απόσταση (distance-based reordering model). Αυτό το μοντέλο δίνει ένα γραμμικό κόστος στην απόσταση αναδιάταξης. Για παράδειγμα, η παράλειψη περισσότερων από δύο λέξεων κοστίζει διπλάσια από την παράλειψη μιας λέξης.

```
$ grep 'πληροφορίες |' ./phrase-table/phrase-table | sort -nrk 7 -t | head
πληροφορίες ||| ΠΛΗΡΟΦΟΡΙΑ ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 ||| |||
περισσότερες πληροφορίες ||| ΠΛΗΡΟΦΟΡΙΑ ΠΟΛΥΣ ||| 1 0.625 1 1 ||| 1-0 0-1 ||| 1 1 1 ||| |||
! περισσότερες πληροφορίες ||| ! ΠΛΗΡΟΦΟΡΙΑ ΠΟΛΥΣ ||| 1 0.511364 1 0.9 ||| 0-0 2-1 1-2 ||| 1 1 1 ||| |||
! ! περισσότερες πληροφορίες ||| ! ! ΠΛΗΡΟΦΟΡΙΑ ΠΟΛΥΣ ||| 1 0.418388 1 0.81 ||| 0-0 1-1 3-2 2-3 ||| 1 1 1 ||| |||
```

```
$ grep 'σύμφωνα με |' ./phrase-table/phrase-table | sort -nrk 7 -t | head
20 04 , σύμφωνα με ||| 20 04 , ΣΥΜΦΩΝΟ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ||| 1 0.121289 1 0.149511 ||| 0-0 1-1 2-2 3-3 4-4 ||| 1 1 1 ||| |||
σύμφωνα με ||| ΣΥΜΦΩΝΟ ΜΕ ||| 1 0.990698 0.0769231 0.150235 ||| 0-0 1-1 ||| 2 26 2 ||| |||
σύμφωνα με ||| ΣΥΜΦΩΝΟ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ||| 1 0.123009 0.230769 0.157994 ||| 0-0 1-1 ||| 6 26 6 ||| |||
σύμφωνα με ||| ΣΥΜΦΩΝΑ ΜΕ ||| 1 0.941163 0.269231 0.317163 ||| 0-0 1-1 ||| 7 26 7 ||| |||
σύμφωνα με ||| ΣΥΜΦΩΝΑ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ||| 1 0.116859 0.230769 0.333542 ||| 0-0 1-1 ||| 6 26 6 ||| |||
! σύμφωνα με ||| ! ||| 0.0833333 0.00676183 0.142857 0.312252 ||| 0-0 1-0 2-0 ||| 12 7 1 ||| |||
χώρας σύμφωνα με ||| ΧΩΡΑ ΣΥΜΦΩΝΟ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ||| 1 0.0471099 1 0.126395 ||| 0-0 1-1 2-2 ||| 1 1 1 ||| |||
του Πάσχα 20 04 , σύμφωνα με ||| ΚΤΗΤ_2 ΠΑΣΧΑ 20 04 , ΣΥΜΦΩΝΟ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ||| 1 0.0246988 1 0.0089451 ||| 0-0 1-1 2-2 3-3 4-4 5-5 6-6 ||| 1 1 1 ||| |||
του καιρού , σύμφωνα με ||| ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΚΑΙΡΟΣ , ΣΥΜΦΩΝΟ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ||| 1 0.000433275 1 0.129065 ||| 0-0 1-1 2-2 3-3 4-4 ||| 2 2 2 ||| |||
τις πρωινές ώρες ! σύμφωνα με ||| ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΩΡΑ ΠΡΩΙΝΟΣ ! ΣΥΜΦΩΝΑ ΜΕ ΔΕΛΤΙΟ ||| 1 0.0318885 0.5 0.00021461 ||| 0-0 2-1 1-2 3-3 4-4 5-5 ||| 1 2 1 ||| |||
```

Σχήμα 9.4 phrase table

Εντούτοις, μπορούν να δημιουργηθούν πρόσθετα μοντέλα ανακατανομής υπό όρους, τα αποκαλούμενα λεξικοποιημένα μοντέλα αναδιοργάνωσης. Υπάρχουν τρεις τύποι λεξικοποιημένων μοντέλων αναδιοργάνωσης στη εργαλειοθήκη Moses που βασίζονται στους Koehn et al. (2005) και Galley and Manning (2008).

Το μοντέλο παραγωγής κατασκευάζεται από την πλευρά-στόχο του παράλληλου σώματος. Από προεπιλογή, υπολογίζονται οι πιθανότητες προς τα εμπρός και προς τα πίσω (forward and backward). Εάν χρησιμοποιήσουμε την επιλογή τύπου απλής παραγωγής (generation single), υπολογίζονται μόνο οι πιθανότητες κατεύθυνσης.

9.1.6 Το στάδιο εξαγωγής του Moses (Moses export stage)

Το σύστημα στατιστικής ΜΜ της εργαλειοθήκης Moses εξάγει τα αποτελέσματα σε μορφή κειμένου. Συγκεκριμένα εξάγονται τρεις τύποι σώματος κειμένων: (α) το σώμα κειμένων-πηγή (Σχήμα 9.5), (β) το σώμα κειμένων αναφοράς (Σχήμα 9.6), το οποίο είναι το σώμα κειμένων που έχει μεταφραστεί από έναν έμπειρο μεταφραστή της ΕΝΓ, και (γ) το σώμα κειμένων-στόχος, δηλαδή η μετάφραση (Σχήμα 9.7) του προτεινόμενου συστήματος ΣΜΜ Moses. Τα τρία αυτά αρχεία εξόδου τα χρησιμοποιούμε αργότερα για να αξιολογήσουμε την ποιότητα μετάφρασης μέσω της μέτρησης ακρίβειας της μετάφρασης του αλγόριθμου BLUE της Papineni et al. (2002).

9.1.7 Αξιολόγηση του ΣΜΜ (SMT MOSES)

Η εργαλειοθήκη Moses περιέχει έτοιμο κώδικα σεναρίου (scripts) για την εκτέλεση του αλγορίθμου αξιολόγησης μετάφρασης BLUE της Papineni et al. (2002). Για αυτή τη διαδικασία, χρησιμοποιήσαμε ως σώμα αξιολόγησης το 10% του συνολικού σώματος της εργασίας μας, το οποίο εξαιρέθηκε πλήρως από την όλη διαδικασία της εκπαίδευσης. Το σώμα αυτό αποτελείται από το σώμα της γλώσσας-πηγής, δηλαδή της Ελληνικής, πάντα σε μορφή γραπτού κειμένου. Το σώμα αυτό αποτέλεσε είσοδο στο ήδη εκπαιδευμένο σύστημα Moses και μεταφράστηκε αυτόματα δημιουργώντας το σώμα μετάφρασης μηχανής. Επίσης, για τις ανάγκες της αξιολόγησης και της εφαρμογής του αλγόριθμου BLUE μεταφράστηκε και χειρωνακτικά από έμπειρο επαγγελματία μεταφραστή και χρησιμοποιήθηκε ως σώμα αναφοράς. Έχουμε δηλαδή το σώμα αξιολόγησης που αποτελείται από τρία σώματα: το σώμα της γλώσσας-πηγής (της Ελληνικής), το σώμα αναφοράς του μεταφραστή και το σώμα μετάφρασης της μηχανής ΣΜΜ-Moses. Υπενθυμίζουμε ότι η αξιολόγηση του προτεινόμενου σχεδίου πραγματοποιείται στο πεδίο της πρόβλεψης καιρικών συνθηκών, δηλαδή των μετεωρολογικών δελτίων. Στο πεδίο αυτό έχουμε δημιουργήσει 20.284 λέξεις και 1.000 προτάσεις. Χρησιμοποιώντας το σώμα αξιολόγησης και τον

Source text:

1. Βροχές και καταιγίδες θα εκδηλωθούν κατά τόπους στη Δυτική Ελλάδα τα Χριστούγεννα.
2. Χιόνια θα πέσουν στα ορεινά της Δυτικής και Βόρειας Ελλάδας σε υψόμετρα μεγαλύτερα των 1400 μέτρων.
3. Στην υπόλοιπη χώρα αναμένονται τοπικές νεφώσεις, ενώ πρόσκαιρες βροχές δεν αποκλείεται να εκδηλωθούν στα κεντρικά ηπειρωτικά.
4. Λόγω των υψηλών ποσοστών υγρασίας και των αυξημένων συγκεντρώσεων σκόνης η ορατότητα θα είναι και πάλι τοπικά περιορισμένη.
5. Οι νότιοι άνεμοι που αρχικά θα πνέουν στα πελάγη με εντάσεις έως 8 μπ.
6. Σταδιακά μετά το μεσημέρι θα στραφούν σε δυτικούς νοτιοδυτικούς και θα παρουσιάσουν εξασθένηση στα επίπεδα των 6 μπ.
7. Οι θερμοκρασίες θα σημειώσουν πτώση κατά 4 έως 5 βαθμούς κυρίως στα δυτικά και νότια ηπειρωτικά διατηρούμενες σε ελαφρώς υψηλότερα από τα κανονικά για την εποχή επίπεδα κυρίως στη Β. Κρήτη.
8. Την Κυριακή 26/12 θα επικρατήσουν ανάλογες καιρικές συνθήκες, ωστόσο οι βροχοπτώσεις στη Δυτική και Βόρεια Ελλάδα και στα νησιά του ανατολικού Αιγαίο θα είναι εντονότερες και μεγαλύτερης διάρκειας, με τάσεις επέκτασης σε αρκετές περιοχές της χώρας.
9. Χιόνια θα πέσουν στα ορεινά της Δυτικής και Βόρειας Ελλάδας σε υψόμετρα μεγαλύτερα των 1500 μέτρων.
10. Παράλληλα οι συγκεντρώσεις σκόνης στην ατμόσφαιρα θα ελαττωθούν σημαντικά με εξαίρεση το Νότιο Αιγαίο και την Κρήτη όπου θα αυξηθούν.

Σχήμα 9.5 Σώμα κειμένων-πηγή (Evaluation Source text)

Source text:

1. ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΧΡΙΣΤΟΥΓΕΝΝΑ ΕΧΕΙ ΒΡΟΧΗ ΚΑΙ ΚΑΤΑΙ-ΓΙΔΑ/ΜΧ(ΕΝΤΑΣΗ)/ΜΓΛ(ΦΟΥΣΚΩΜΕΝΑ) ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΤΟΠΟΣ/ΤΠΘ(Χ1)-ΤΟΠΟΣ/ΤΠΘ(Χ2) ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΛΛΑΔΑ ΔΥΤΙ-ΚΟΣ .

2. ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΧΙΟΝΙ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΟΡΕΙΝΑ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΛΛΑΔΑ ΒΟΡΕΙΟΣ ΔΥΤΙΚΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΥΨΟΜΕΤΡΟ ΜΕΓΑΛΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) 1400 ΜΕΤΡΟ .

3. ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΧΩΡΑ ΥΠΟΛΟΙΠΟΣ ΑΝΑΜΕΝΩ ΣΥΝΝΕΦΑ-/ΜΤ(ΜΙΣΑΝΟΙΧΤΑ ΤΟΠΙΚΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΒΡΟΧΗ ΝΩΡΙΣ ΔΕΝ ΑΠΟΚΛΕΙΕΤΑΙ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΚΔΗΛΩΝΩ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΗΠΕΙΡΟΣ ΚΕΝΤΡΙΚΑ .

4. ΛΟΓΩ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΠΟΣΟΣΤΟ ΨΗΛΟΣ ΥΓΡΑΣΙΑ ΚΑΙ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΣΚΟΝΗ ΜΑΖΕΥΕΙ ΠΟΛΥ/ΧΛ(ΠΟΛΥ) ΟΡΑΤΟΤΗ-ΤΑ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΚΑΙ ΠΑΛΙ ΤΟΠΙΚΑ ΠΕΡΙΟΡΙΣΜΕΝΟΣ .

5. ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΑΝΕΜΟΣ ΝΟΤΙΟΣ ΠΟΥ ΑΡΧΙΚΑ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΠΝΕΩ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΠΕΛΑΓΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΝΤΑΣΗ Ε-ΩΣ 8 ΜΠΟΦΩΡ .

6. ΣΤΑΔΙΑΚΑ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΜΕΣΗΜΕΡΙ ΣΤΡΕΦΩ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΔΥΤΙΚΟΣ ΝΟΤΙΟΔΥΤΙΚΟΣ ΚΑΙ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΜΕΙΩΣΗ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΠΙΠΕΔΑ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) 6 ΜΠΟΦΩΡ .

7. ΘΕΡΜΟΚΡΑΣΙΑ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΠΤΩΣΗ ΑΠΟ 4 ΕΩΣ 5 ΒΑΘΜΟΣ ΚΥΡΙΩΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΔΥΤΙΚΑ ΚΑΙ ΝΟΤΙΑ ΗΠΕΙΡΟΣ ΔΙΑΤΗΡΟΥΜΕΝΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΛΑΦΡΩΣ ΨΗΛΟΣ ΑΠΟ ΚΑΝΟΝΙΚΑ ΚΤΗΤ_3 ΕΠΟ-ΧΗ ΕΠΙΠΕΔΟ ΚΥΡΙΩΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΒΟΡΕΙΑ ΚΡΗΤΗ .

8. ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΚΥΡΙΑΚΗ 26 12 ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΕΧΕΙ ΑΝΑΛΟ-ΓΑ/ΧΛ(ΜΑΓΚΩΜΕΝΑ) ΚΑΙΡΟΣ ΩΣΤΟΣΟ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΒΡΟ-ΧΟΠΤΩΣΗ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΛΛΑΔΑ ΒΟΡΕΙΟΣ ΔΥΤΙΚΟΣ ΚΑΙ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΝΗΣΙ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΑΝΑΤΟΛΙΚΟΣ ΑΙ-ΓΑΙΟ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΓΙΝΕΙ ΕΝΤΟΝΟΣ ΚΑΙ ΜΕΓΑΛΥΤΕΡΟΣ ΔΙΑΡΚΕΙΑ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΣΤΟΧΟ ΕΠΕΚΤΑΣΗ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΠΕΡΙΟΧΗ ΑΡΚΕΤΟΣ ΚΤΗΤ_3 ΧΩΡΑ .

9. ΜΕΤΑ ΧΙΟΝΙ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΟΡΕΙΝΑ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΛ-ΛΑΔΑ ΒΟΡΕΙΟΣ ΔΥΤΙΚΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΥΨΟΜΕΤΡΟ ΜΕΓΑΛΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) 1500 ΜΕΤΡΟ .

10. ΠΑΡΑΛΛΗΛΑ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΜΑΖΕΥΕΙ ΣΚΟΝΗ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΑΤΜΟΣΦΑΙΡΑ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΕΛΑΤΤΩΝΩ ΣΗΜΑΝΤΙΚΑ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΞΑΙΡΕΣΗ ΝΟΤΙΟΣ ΑΙΓΑΙΟ ΚΑΙ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΚΡΗΤΗ ΟΠΟΥ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΑΥΞΑΝΩ .

Σχήμα 9.6 Σώμα κειμένων αναφοράς (Evaluation Reference text)

Source text:

1. ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΧΡΙΣΤΟΥΓΕΝΝΑ ΕΧΕΙ ΒΡΟΧΗ ΚΑΙ ΚΑΤΑΙΓΙΔΑ/ΜΧ(ΕΝΤΑΣΗ)/ΜΓΛ(ΦΟΥΣΚΩΜΕΝΑ) ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΤΟΠΟΣ/ΤΠΘ(Χ1)-ΤΟΠΟΣ/ΤΠΘ(Χ2) ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΛΛΑΔΑ ΔΥΤΙΚΟΣ.

2. ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΧΙΟΝΙ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΟΡΕΙΝΑ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΛΛΑΔΑ ΒΟΡΕΙΟΣ ΔΥΤΙΚΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΥΨΟΜΕΤΡΟ ΜΕΓΑΛΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) 1400 ΜΕΤΡΟ.

3. ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΧΩΡΑ ΥΠΟΛΟΙΠΟΣ ΑΝΑΜΕΝΩ ΣΥΝΝΕΦΑ-/ΜΤ(ΜΙΣΑΝΟΙΧΤΑ ΤΟΠΙΚΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΒΡΟΧΗ ΝΩΡΙΣ ΔΕΝ ΑΠΟΚΛΕΙΕΤΑΙ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΚΔΗΛΩΝΩ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΗΠΕΙΡΟΣ ΚΕΝΤΡΙΚΑ.

4. ΛΟΓΩ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΠΟΣΟΣΤΟ ΨΗΛΟΣ ΥΓΡΑΣΙΑ ΚΑΙ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΣΚΟΝΗ ΜΑΖΕΥΕΙ ΠΟΛΥ/ΧΛ(ΠΟΛΥ) ΟΡΑΤΟΤΗΤΑ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΚΑΙ ΠΑΛΙ ΤΟΠΙΚΑ ΠΕΡΙΟΡΙΣΜΕΝΟΣ .

5. ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΑΝΕΜΟΣ ΝΟΤΙΟΣ ΠΟΥ ΑΡΧΙΚΑ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΠΝΕΩ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΠΕΛΑΓΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΝΤΑΣΗ ΕΩΣ 8 ΜΠΟΦΩΡ.

6. ΣΤΑΔΙΑΚΑ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΜΕΣΗΜΕΡΙ ΣΤΡΕΦΩ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΔΥΤΙΚΟΣ ΝΟΤΙΟΔΥΤΙΚΟΣ ΚΑΙ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΜΕΙΩΣΗ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΠΙΠΕΔΑ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) 6 ΜΠΟΦΩΡ .

7. ΘΕΡΜΟΚΡΑΣΙΑ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΠΤΩΣΗ ΑΠΟ 4 ΕΩΣ 5 ΒΑΘΜΟΣ ΚΥΡΙΩΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΔΥΤΙΚΑ ΚΑΙ ΝΟΤΙΑ ΗΠΕΙΡΟΣ ΔΙΑΤΗΡΟΥΜΕΝΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΛΑΦΡΩΣ ΨΗΛΟΣ ΑΠΟ ΚΑΝΟΝΙΚΑ ΚΤΗΤ_3 ΕΠΟΧΗ ΕΠΙΠΕΔΟ ΚΥΡΙΩΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΒΟΡΕΙΑ ΚΡΗΤΗ .

8. ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΚΥΡΙΑΚΗ 26 12 ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΕΧΕΙ ΑΝΑΛΟΓΑ/ΧΛ(ΜΑΓΚΩΜΕΝΑ) ΚΑΙΡΟΣ ΩΣΤΟΣΟ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΒΡΟΧΟΠΤΩΣΗ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΛΛΑΔΑ ΒΟΡΕΙΟΣ ΔΥΤΙΚΟΣ ΚΑΙ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΝΗΣΙ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΑΝΑΤΟΛΙΚΟΣ ΑΙΓΑΙΟ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΓΙΝΕΙ ΕΝΤΟΝΟΣ ΚΑΙ ΜΕΓΑΛΥΤΕΡΟΣ ΔΙΑΡΚΕΙΑ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΣΤΟΧΟ ΕΠΕΚΤΑΣΗ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΠΕΡΙΟΧΗ ΑΡΚΕΤΟΣ ΚΤΗΤ_3 ΧΩΡΑ .

9. ΜΕΤΑ ΧΙΟΝΙ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΟΡΕΙΝΑ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΛΛΑΔΑ ΒΟΡΕΙΟΣ ΔΥΤΙΚΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΥΨΟΜΕΤΡΟ ΜΕΓΑΛΟΣ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) 1500 ΜΕΤΡΟ.

10. ΠΑΡΑΛΛΗΛΑ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΜΑΖΕΥΕΙ ΣΚΟΝΗ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΑΤΜΟΣΦΑΙΡΑ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΕΛΑΤΤΩΝΩ ΣΗΜΑΝΤΙΚΑ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΕΞΑΙΡΕΣΗ ΝΟΤΙΟΣ ΑΙΓΑΙΟ ΚΑΙ ΑΝΤ_3/ΜΤ(ΑΝΟΙΧΤΑ) ΚΡΗΤΗ ΟΠΟΥ ΜΕΤΑ/ΧΛ(ΜΕΤΑ) ΑΥΞΑΝΩ .

Σχήμα 9.7 Τελικό κείμενο μετάφρασης για αξιολόγηση (Evaluation Target text)

Μέση βαθμολογία (average score)	60.53%
1-gram	85.1%
2-gram	65.5%
3-gram	53.8%
4-gram	44.8%

Πίνακας 9.1 Αποτελέσματα αξιολόγησης ΜΜ BLEU για το SMT Moses

αλγόριθμο αξιολόγησης ΜΜ BLEU της Papineni (Papineni et al., 2002), το προτεινόμενο σύστημα στατιστικής ΜΜ που βασίστηκε στο Moses και εκπαιδεύτηκε από παράλληλα σώματα κειμένου που δημιουργήθηκαν χειρωνακτικά με την βοήθεια ενός ΜΜ-κανόνων, επιτυγχάνει μια μέση βαθμολογία 60,53% και συγκεκριμένα 85,1% / 65,5% / 53,8% / 44,8% για 1-gram / 2 -gram / 3-gram / 4-gram (Πίνακας 9.1).

9.1.8 Αποτελέσματα παρόμοιων συστημάτων ΣΜΜ για άλλες ΝΓ

Για λόγους σύγκρισης των αποτελεσμάτων, αξίζει να σας παρουσιάσουμε μερικές παρόμοιες εργασίες. Στη μελέτη του Kanis (Krňoul et al., 2007), το σώμα κατάρτισης περιελάμβανε 12.616 προτάσεις και αφορούσε ένα προτεινόμενο σύστημα ΜΜ της Τσεχικής γλώσσας προς την Τσεχική Νοηματική Γλώσσα. Σε αυτά τα πειράματα το προτεινόμενο σύστημα του Kanis πέτυχε την ακόλουθη βαθμολογία με τη χρήση των αλγορίθμων BLEU 0,81, WER 13,14% και PER 11,64%. Ομοίως, στη μελέτη του Stein (Stein et al., 2010) που αφορά την περίπτωση συστήματος ΣΜΜ της Γερμανικής γλώσσας προς Γερμανική Νοηματική γλώσσα, όπου πραγματοποιήθηκαν δύο πειράματα. Σε αυτές τις περιπτώσεις, τα BLEU και PER που ελήφθησαν ήταν 0,021 και 85,7% για το πρώτο πείραμα και 0,026 και 81,1% για το δεύτερο πείραμα αντίστοιχα. Ωστόσο, η αναφερθείσα βασική γραμμή με το εργαλείο ανοικτού λογισμικού για τη στατιστική μηχανική μετάφραση Moses (Koehn and Hoang, 2007) ήταν 0,181 BLEU και 71,0% TER, με εκπαιδευτικό σώμα 2.565 προτάσεων και σώμα δοκιμής 512 προτάσεων. Συνδυάζοντας διάφορα συστήματα, έφτασαν τελικά στις τιμές BLEU 0,234 και TER 65,5%. Εδώ θα πρέπει να σημειωθεί ότι η διαφορά μεταξύ αυτών των αποτελεσμάτων οφείλεται στο γεγονός ότι η Τσεχική γλώσσα και η Τσεχική ΝΓ έχουν την ίδια επιφανειακή δομή, ενώ η Γερμανική γλώσσα και η Γερμανική ΝΓ διαφέρουν. Επιπλέον, τα αποτελέσματα επιβεβαιώνουν ότι η έλλειψη δεδομένων και ο περιορισμός του εύρους των πεδίων έχουν ως αποτέλεσμα οι προσπάθειες και οι προσεγγίσεις των συστημάτων ΜΜ που βασίζονται σε δεδομένα να εμφανίζουν χειρότερες αποδόσεις από ότι τα συστήματα ΜΜ που βασίζονται σε κανόνες.

Κεφάλαιο 10

Συμπεράσματα και μελλοντικές επεκτάσεις

10.1 Σύνοψη και συμπεράσματα

Στην παρούσα διατριβή εστιάσαμε σε προβλήματα και προκλήσεις του κλάδου της ΜΜ για την ΝΓ. Αρχικά πραγματοποιήσαμε μια εκτενή βιβλιογραφική μελέτη για όλες τις σχετικές ερευνητικές εργασίες. Όπως είναι φανερό και από τη διάρθρωσή της, η παρούσα διατριβή αναπτύσσει και προτείνει έναν αριθμό μεθόδων εφαρμοσμένης γλωσσολογίας που βασίζονται στη φυσική γλώσσα των κωφών, τη Νοηματική Γλώσσα.

Συγκεκριμένα, στην παρούσα διατριβή ερευνώνται τα συστήματα μηχανικής μετάφρασης και γλωσσικών μοντέλων στις ΝΓ και ιδιαίτερα στην ΕΝΓ, όπου εφαρμόζουμε και πραγματοποιούμε πειραματικά αποτελέσματα.

Η επιλογή τεχνολογίας συγκεκριμένου τύπου για την επεξεργασία μιας γλώσσας (είτε πρόκειται για προφορική γλώσσα είτε για νοηματική) επηρεάζεται σε μεγάλο βαθμό από τη διαθεσιμότητα γλωσσικών πόρων της εν λόγω γλώσσας και πιο συγκεκριμένα από τη διαθεσιμότητα ψηφιακά αποθηκευμένων πόρων. Η εμπορική έρευνα και ανάπτυξη εφαρμογών επικεντρώθηκε κυρίως σε γλώσσες που παρουσιάζουν υψηλή διαθεσιμότητα ψηφιακών πόρων. Σήμερα, η ΕΝΓ, όπως και κάθε άλλη ΝΓ στο κόσμο, ανήκει στις γλώσσες με χαμηλή διαθεσιμότητα γλωσσικών πόρων και περιορισμένη χρηματοδότηση. Η δημιουργία γλωσσικών πόρων και δεδομένων στις ΝΓ είναι μια χρονοβόρα και δαπανηρή διαδικασία σε σύγκριση με τις προφορικές γλώσσες. Κατά την περίοδο συγγραφής της παρούσας εργασίας, δεν υπήρχε παράλληλο σώμα επαρκούς μεγέθους για την ΕΝΓ με το οποίο θα μπορούσε κανείς να πραγματοποιήσει προσεγγίσεις και δοκιμαστικές εφαρμογές για μηχανική μετάφραση σε διάφορους τομείς. Για τον λόγο αυτό, κρίνεται

απαραίτητο να αναπτυχθούν εργαλεία παρόμοια με αυτά που προτείνονται, έτσι ώστε οποιοσδήποτε να μπορεί εύκολα και γρήγορα να δημιουργήσει μεγάλα και υψηλής ποιότητας σώματα της ΕΝΓ σε σύντομη γραπτή μορφή (glosses) ή άλλη μορφή. Τα σύντομα αυτά glosses, όπως τα δικά μας στη παρούσα εργασία (ΣΜΕΝΓ), θα μπορούν στο μέλλον να τροφοδοτήσουν εφαρμογές κινούμενων ανθρωποειδών και να αναπαραστήσουν εικονικά τα νοήματα της εξόδου του συστήματός μας ΜΜ σε μορφή 3D avatars.

Από την άλλη πλευρά, η ΕΝΓ, όπως και όλες οι άλλες ΝΓ στον κόσμο, δεν είναι τυποποιημένη, ενώ δεν έχει δημοσιευθεί προς το παρόν καμία πλήρης γραμματική της ΕΝΓ. Μόνο μερικά πρόσφατα έργα επισημαίνουν κάποια σημαντικά γραμματικά σημεία, αναφορές και γλωσσικά φαινόμενα της ΕΝΓ (Efthimiou et al., 2016, Fotinea et al., 2005, Kouremenos et al., 2010). Όλα αυτά τα προβλήματα καθιστούν ως μόνη βιώσιμη λύση την ανάπτυξη ενός ειδικού συστήματος ΜΜ κανόνων για την ΕΝΓ που θα λειτουργεί ως υποστηρικτικό μεταφραστικό εργαλείο, υπό την επίβλεψη ενός επαγγελματία μεταφραστή. Σε αυτή την περίπτωση ο μεταφραστής με τη βοήθεια ενός συστήματος ΜΜ κανόνων θα είναι σε θέση να δημιουργήσει μεγάλα, παράλληλα, υψηλής ποιότητας σώματα ελληνικών κειμένων σε σύντομες γραπτές μεταγραφές της ΕΝΓ, από οποιαδήποτε θεματική περιοχή. Στην παρούσα εργασία, με την προτεινόμενη μεθοδολογία δημιουργήσαμε μεγάλα και υψηλής ποιότητας σώματα της ΕΝΓ, από τα οποία προέκυψαν γλωσσικά μοντέλα με καλά αποτελέσματα περιπλοκής. Στη συνέχεια δοκιμάστηκαν και σε σύστημα ΣΜΜ, συγκεκριμένα στην πλατφόρμα MOSES, όπου έδωσαν και εκεί ικανοποιητικά αποτελέσματα. Θεωρούμε ότι η προτεινόμενη μεθοδολογία θα μπορούσε να εξελιχθεί και ότι δείχνει τον δρόμο προς τη βελτίωση και επέκταση και σε άλλα πεδία σωμάτων της ΕΝΓ.

Είναι σημαντικό να τονίσουμε ότι είναι η πρώτη φορά που δοκιμάζεται ένα σύστημα στατιστικής μηχανικής μετάφρασης στην ΕΝΓ, όπως και η μοντελοποίηση της ΕΝΓ. Θεωρούμε ότι τα δεδομένα δεν επαρκούν ακόμη για να εκφράσουμε γενικά συμπεράσματα για όλο το φάσμα της ΕΝΓ, πέρα από το θεματικό πεδίο που ερευνήθηκε. Όσον αφορά το σύστημα ΜΜ με κανόνες, το πιο σημαντικό συμπέρασμα που μπορεί να εξαχθεί από τα παραπάνω πειράματα είναι ότι σε αρκετές περιπτώσεις η σειρά των νοημάτων είναι παρόμοια με τη σειρά των λέξεων του σώματος της ελληνικής γλώσσας. Εντούτοις, δεν πρέπει να θεωρηθεί ότι η ΕΝΓ και η ελληνική γλώσσα έχουν παρόμοιες δομές/σειρές λέξεων ή ότι η σειρά που παράγεται από το σύστημα δεν είναι έγκυρη. Θεωρείται ότι η δομή της ΕΝΓ αποδέχεται ως έναν βαθμό μια σχετική ελευθερία και ότι η σειρά των λέξεων/νοημάτων στο σώμα της κατάρτισης είναι έγκυρη και ικανοποιεί τους σκοπούς της επικοινωνίας. Εκτός αυτού θα πρέπει να ληφθεί υπόψη ότι χρησιμοποιήθηκε μια συγκεκριμένη θεματική περιοχή για τη συλλογή των σωμάτων κειμένων για τις ανάγκες της εργασίας μας. Προτείνεται στο μέλλον να επεκταθεί η έρευνα σε μεγαλύτερα σώματα κει-

μένων από διαφορετικές θεματικές ενότητες και από διαφορετικούς μεταφραστές, έτσι ώστε ληφθεί υπόψη και η κατανόηση από τον ανθρώπινο παράγοντα, ώστε να εξαχθούν περαιτέρω συμπεράσματα.

Επιπλέον, η εργασία αυτή αφορά μόνο τη μετάφραση γραπτών γλωσσών Greek-to-GSL. Το σύστημα κειμένου ΣΜΕΝΓ που χρησιμοποιήθηκε δεν αποτελεί κάποιο επίσημο σύστημα γραφής για την ΕΝΓ. Έτσι, δεν μπορεί να γίνει κατανοητό από τους κωφούς που δεν μπορούν να κατανοήσουν τον ελληνικό γραπτό λόγο. Ένα ολοκληρωμένο σύστημα ΜΜ για την ΕΝΓ θα πρέπει να παράγει κινούμενα σχέδια (3D avatar), ενώ μια πραγματική και σωστή αξιολόγηση, εκτός από τους σχετικούς αλγορίθμους που υπάρχουν, θα πρέπει να περιλαμβάνει και κωφούς νοηματιστές, που θα αξιολογούν τα αποτελέσματα σχετικά με την κινούμενη παραγωγή του εικονικού ανθρωποειδούς. Εντούτοις, ένα κείμενο της ΣΜΕΝΓ χωρίς τις ετικέτες των NmCs και POS θα μπορούσε να γίνει κατανοητό από τους "δίγλωσσους κωφούς" [1], αφού είναι πιο κοντά στη δομή και τη σύνταξη της ΕΝΓ. Επιπλέον, η διάδοση των κοινωνικών δικτύων έχει κάνει τους δίγλωσσους κωφούς να χρησιμοποιούν γραπτά μηνύματα της ελληνικής. Έτσι, κάτω από αυτές τις συνθήκες, δημιούργησαν με κάποιο τρόπο μια ειδική γλώσσα "γραφής" της ΕΝΓ, κάνοντας χρήση κεφαλαίων γραμμάτων, με εξάλειψη των άρθρων και ακολουθώντας μια δομή παρόμοια με αυτή της ΕΝΓ.

10.2 Μελλοντικές επεκτάσεις

Κατά την εκπόνηση της παρούσας διατριβής, διαπιστώθηκαν τα παρακάτω θέματα τα οποία προτείνονται για μελλοντική εργασία.

Στην κατεύθυνση των συστημάτων ΜΜ για την ΕΝΓ προτείνονται τα εξής:

- η επέκταση των σωμάτων και σε άλλες θεματικές περιοχές σωμάτων, έτσι ώστε να αξιολογηθεί περαιτέρω η απόδοση του συνολικού συστήματος.

- να εξαντληθούν και να αξιολογηθούν διάφορες λειτουργικές επιλογές της πλατφόρμας Moses.

- να συνδεθεί η έξοδος του συστήματός μας με μια γεννήτρια σύνθεσης της ΕΝΓ και να αποδίδονται τα αποτελέσματα της μετάφρασης από ένα τρισδιάστατο εικονικό ανθρωποειδές.

[1] Ο όρος "δίγλωσσος κωφός" αναφέρεται στα άτομα με προβλήματα ακοής που χειρίζονται με άνεση και τις δυο γλώσσες, την ΕΝΓ και την ελληνική (ομιλούμενη και γραπτή)

- να γίνει χρήση βελτιωμένων αλγορίθμων, καθώς και της τεχνητής νοημοσύνης στη μηχανική εκπαίδευση, με σκοπό τη βελτίωση της απόδοσης της μεταφραστικής μηχανής με λιγότερα σώματα κειμένων ανά θεματική περιοχή.
- η δοκιμή των σωμάτων κειμένων της εργασίας μας και σε άλλες ανοιχτές πλατφόρμες ΜΜ.

Βιβλιογραφία

Jordi Atserias, Bernardino Casas, Elisabet Comelles, Maritxell González, Lluís Padró, and Muntsa Padró. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *LREC*, volume 6, pages 48–55, 2006.

RA Augustus, E Ritchie, and S Stecker. The official american sign language writing textbook. *Los Angeles, CA: ASLized*, 2013.

Charlotte Baker and Robbin Battison. *Sign language and the Deaf community: Essays in honor of William C. Stokoe*. National Association of the deaf Silver Spring, MD, 1980.

Sandra Baldassarri and Francisco Royo-Santas. An automatic rule-based translation system to spanish sign language (lse). In *New Trends on Human–Computer Interaction*, pages 1–11. Springer, 2009.

JA Bangham, SJ Cox, R Elliott, JRW Glauert, I Marshall, S Rankov, and M Wells. An overview of visicast. In *IEEE Seminar on Speech and language processing for disabled and elderly people*, 2000.

Britta Bauer, Sonja Nießen, and Hermann Hienz. Towards an automatic sign language translation system. In *In 1st international*. Citeseer, 1999.

Adam Boretz and A Adam. Apptek launches hybrid machine translation software. *SpeechTechMag. com*, 2, 2009.

Annelies Braffort, Annick Choisier, Christophe Collet, Patrice Dalle, Frédérick Gianni, F Lenseigne, and Jérémie Segouat. Toward an annotation software for video of sign language, including image processing tools and signing space modelling. In *LREC*, 2004.

Peter Brown, John Cocke, S Della Pietra, V Della Pietra, Frederick Jelinek, Robert Mercer, and Paul Roossin. A statistical approach to language translation. In *Proceedings of the 12th conference on Computational linguistics-Volume 1*, pages 71–76. Association for Computational Linguistics, 1988.

Peter F Brown, John Cocke Stephen, A Della Pietra, J Della Vincent, Pietra Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. A statistical approach to machine translation. 1991.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.

Hennie Brugman, Peter Wittenburg, Stephen C Levinson, and Sotaro Kita. Multimodal annotations in gesture and sign language studies. In *The 3rd International Conference on Language Resources and Evaluation (LREC)*, pages 176–182. European Language Resources Association., 2002.

Jan Bungeroth, Daniel Stein, Philippe Dreuw, Morteza Zahedi, and Hermann Ney. A german sign language corpus of the domain weather report. In *Fifth International Conference on Language Resources and Evaluation*, pages 2000–2003, 2006.

Jan Bungeroth, Daniel Stein, Philippe Dreuw, Hermann Ney, Sara Morrissey, Andy Way, and Lynette van Zijl. The atis sign language corpus. 2008.

H Bunke and T Caelli. Hidden markov models applied in computer vision, em "machine perception and artificial intelligence", 2001.

Bob Carpenter. *The logic of typed feature structures: with applications to unification grammars, logic programs and constraint resolution*, volume 32. Cambridge University Press, 2005.

Robert L Carpenter. The logic of typed feature structures: With applications to unification grammars, logic programs and constraint resolution (cambridge tracts in theoretical computer science). 1992.

Violetta Cavalli-Sforza, Jaime G Carbonell, and Peter J Jansen. Developing language resources for a transnational digital government system. 2004.

John Chandioux. Meteo: an operational system, for the translation of public weather forecasts. In *FBIS Seminar on Machine Translation. American Journal of Computational Linguistics, microfiche*, volume 46, pages 27–36, 1976.

John Chandioux. Météo: 100 million words later. In *American Translators Association Conference*, pages 449–453, 1989.

DI Chen, Ming Zhao, and Gregory R Mundy. Bone morphogenetic proteins. *Growth factors*, 22(4):233–241, 2004.

Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394, 1999.

David Chiang. Learning to translate with source and target syntax. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1443–1452. Association for Computational Linguistics, 2010.

Yu-Hsien Chiu, Chung-Hsien Wu, Hung-Yu Su, and Chih-Jen Cheng. Joint optimization of word alignment and epenthesis generation for chinese to taiwanese sign synthesis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (1):28–39, 2007.

Kenneth W Church and William A Gale. Probability scoring for spelling correction. *Statistics and Computing*, 1(2):93–103, 1991.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

Thomas M Cover and Joy A Thomas. Entropy, relative entropy and mutual information. *Elements of information theory*, 2:1–55, 1991.

Robert Dale, Hermann Moisl, and Harold Somers. *Handbook of natural language processing*. CRC Press, 2000.

Donald M Decker et al. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

Vlado Delić, Milan Sečujski, Nikša Jakovljević, Darko Pekar, Dragiša Mišković, Branislav Popović, Stevan Ostrogonac, Milana Bojanić, and Dragan Knežević. Speech and language resources within speech recognition and synthesis systems for serbian and kindred south slavic languages. In *International Conference on Speech and Computer*, pages 319–326. Springer, 2013.

Steve DeNeefe, Kevin Knight, and Heiko Vogler. A decoder for probabilistic synchronous tree insertion grammars. In *Proceedings of the 2010 Workshop on Applications of Tree Automata in Natural Language Processing*, pages 10–18, 2010.

Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

Philippe Dreuw, Carol Neidle, Vassilis Athitsos, Stan Sclaroff, and Hermann Ney. Benchmark databases for video-based automatic sign language recognition. In *LREC*, 2008.

Eleni Efthimiou, Stavroula-Evita Fotinea, and Galini Sapountzaki. Feature-based natural language processing for gsl synthesis. *Sign Language & Linguistics*, 10(1):3–23, 2007.

Eleni Efthimiou, SE Fontinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Goudenove. Dicta-sign–sign language recognition, generation and modelling: a research effort with applications in deaf communication. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 80–83, 2010.

Eleni Efthimiou, Stavroula-Evita Fotinea, Athanasia-Lida Dimou, Theodore Goulas, and Dimitris Kouremenos. From grammar-based mt to post-processed sl representations. *Universal Access in the Information Society*, 15(4):499–511, 2016.

Ralph Elliott, JRW Glauert, Vince Jennings, and JR Kennaway. An overview of the sigml notation and sigml signing software system. In *Fourth International Conference on Language Resources and Evaluation, LREC*, pages 98–104, 2004.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144, 2011.

Stavroula-Evita Fotinea, Eleni Efthimiou, and Dimitris Kouremenos. Generating linguistic content for greek to gsl conversion. In *Proceedings of the HERCMA-2005 Conference (The 7th Hellenic European Conference on Computer Mathematics and its Applications), Athens*. Citeseer, 2005.

Stavroula-Evita Fotinea, Eleni Efthimiou, George Caridakis, and Kostas Karpouzis. A knowledge-based sign synthesis architecture. *Universal Access in the Information Society*, 6(4):405–418, 2008.

Michel Galley and Christopher D Manning. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 848–856. Association for Computational Linguistics, 2008.

Laurie Gerber and Jin Yang. Systran mt dictionary development. In *Machine Translation: Past, Present and Future. In: Proceedings of Machine Translation Summit VI. October*, pages 211–218, 1997.

IJ Good. The serial test for sampling numbers and other tests for randomness. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 49, pages 276–284. Cambridge University Press, 1953.

Angus B Grieve-Smith. English to american sign language machine translation of weather reports. In *Proceedings of the Second High Desert Student Conference in Linguistics (HDSL2), Albuquerque, NM*, pages 23–30, 1999.

Thomas Hanke. ilex-a tool for sign language lexicography and corpus analysis. In *LREC*, 2002.

Thomas Hanke. Hamnosys-representing sign language data in language resources and language processing contexts. In *LREC*, volume 4, pages 1–6, 2004.

Birgit Hellwig and Dieter Van Uytvanck. Eudico linguistic annotator (elan) version 2.4 manual. Technical report, Technical report, MaxPlanck Institute for Psycholinguistics, Nijmegen, The …, 2005.

Gregory Hickok, Ursula Bellugi, and Edward S Klima. The neural organization of language: evidence from sign language aphasia. *Trends in cognitive sciences*, 2(4):129–136, 1998.

Nini Hoiting and Dan I Slobin. Transcription as a tool for understanding. *Directions in sign language acquisition*, 2:55, 2002.

M Huenerfauth. Misconceptions, technical challenges, and new technologies for generating american sign language animation. *Universal Access in the Information Society*, 6:419–434, 2007.

Matt Huenerfauth, Mitch Marcus, and Martha Palmer. *Generating American Sign Language classifier predicates for English-to-ASL machine translation*. PhD thesis, University of Pennsylvania, 2006.

Matthew P Huenerfauth. American sign language natural language generation and machine translation systems. Technical report, Technical Report, computer and information sciences, University of Pennsylvania, 2003.

John Hutchins. Machine translation: A concise history. *Computer aided translation: Theory and practice*, 13:29–70, 2007.

William John Hutchins and Harold L Somers. *An introduction to machine translation*, volume 362. Academic Press London, 1992.

Nancy Ide and Chris Brew. Requirements, tools, and architectures for annotated corpora. In *Proceedings of data architectures and software support for large corpora*, pages 1–5. Citeseer, 2000.

Hitoshi Isahara and Yuriko Uchida. Analysis, generation and semantic representation in contrast—a context-based machine translation system. *Systems and computers in Japan*, 26(14):37–53, 1995.

Gurpreet Singh Josan and Gurpreet Singh Lehal. A punjabi to hindi machine translation system. In *22nd International Conference on on Computational Linguistics: Demonstration Papers*, pages 157–160. Association for Computational Linguistics, 2008.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition, 2000. ISBN 0130950696.

Hiroyuki Kaji. An efficient execution method for rule-based machine translation. In *Proceedings of the 12th conference on Computational linguistics-Volume 2*, pages 824–829. Association for Computational Linguistics, 1988.

George Karypis. Evaluation of item-based top-n recommendation algorithms. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 247–254. ACM, 2001.

Slava Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE transactions on acoustics, speech, and signal processing*, 35(3):400–401, 1987.

Martin Kay. Functional unification grammar: A formalism for machine translation. In *Proceedings of the 10th International Conference on Computational Linguistics*, pages 75–78. Association for Computational Linguistics, 1984.

Michael Kipp. Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European Conference on Speech Communication and Technology*, 2001.

Poul Søren Kjærsgaard. Reftex: a context-based translation aid. In *Proceedings of the third conference on European chapter of the Association for Computational Linguistics*, pages 109–112. Association for Computational Linguistics, 1987.

Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling. In *icassp*, volume 1, page 181e4, 1995.

Philipp Koehn. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Conference of the Association for Machine Translation in the Americas*, pages 115–124. Springer, 2004.

Philipp Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

Philipp Koehn and Hieu Hoang. Factored translation models. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 868–876, 2007.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics, 2003.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. Edinburgh system description for the 2005 iwslt speech translation evaluation. In *International Workshop on Spoken Language Translation (IWSLT) 2005*, 2005.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics, 2007.

E Koleli. *A new Greek part-of-speech tagger, based on a maximum entropy classifier*. PhD thesis, Thesis, Athens University of Economics, 2011.

Stefan Kombrink, Tomáš Mikolov, Martin Karafiát, and Lukáš Burget. Recurrent neural network based language modeling in meeting recognition. In *Twelfth annual conference of the international speech communication association*, 2011.

Dimitris Kouremenos, Stavroula-Evita Fotinea, Eleni Efthimiou, and Klimis Ntalianis. A prototype greek text to greek sign language conversion system. *Behaviour & Information Technology*, 29(5):467–481, 2010.

Zdeněk Krňoul, Jakub Kanis, Miloš Železnỳ, and Luděk Müller. Czech text-to-sign speech synthesizer. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 180–191. Springer, 2007.

Penny Labropoulou, Elena Mantzari, and Maria Gavrilidou. Lexicon-morphosyntactic specifications: Language specific instantiation. *PP-PAROLE, MLAP*, pages 63–386, 1996.

Sung-Chien Lin, Chi-Lung Tsai, Lee-Feng Chien, Ker-Jiann Chen, and Lin-Shan Lee. Chinese language model adaptation based on document classification and multiple domain-specific language models. In *Fifth European Conference on Speech Communication and Technology*, 1997.

Wolfgang Macherey, Franz Josef Och, Ignacio Thayer, and Jakob Uszkoreit. Lattice-based minimum error rate training for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 725–734. Association for Computational Linguistics, 2008.

James H Martin and Daniel Jurafsky. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall Upper Saddle River, 2009.

Guillem Massó and Toni Badia. Dealing with sign language morphemes in statistical machine translation. In *4th workshop on the representation and processing of sign languages: corpora and sign language technologies, Valletta, Malta*, pages 154–157, 2010.

Evgeny Matusov, Richard Zens, David Vilar, Arne Mauser, Maja Popovic, Saša Hasan, and Hermann Ney. The rwth machine translation system. In *TC-STAR Workshop on Speech-to-Speech Translation*, pages 31–36, 2006.

Rachel I Mayberry. First-language acquisition after childhood differs from second-language acquisition: The case of american sign language. *Journal of Speech, Language, and Hearing Research*, 36(6):1258–1270, 1993.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

Robert C Moore and William Lewis. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 conference short papers*, pages 220–224. Association for Computational Linguistics, 2010.

Sara Morrissey. *Data-driven machine translation for sign languages*. PhD thesis, Dublin City University, 2008.

Sara Morrissey. Assessing three representation methods for sign language machine translation and evaluation. In *Proceedings of the 15th annual meeting of the European Association for Machine Translation (EAMT 2011), Leuven, Belgium*, pages 137–144, 2011.

Sara Morrissey and Andy Way. Manual labour: tackling machine translation for sign languages. *Machine Translation*, 27(1):25–64, 2013.

Sara Morrissey, Andy Way, Daniel Stein, Jan Bungeroth, and Hermann Ney. Combining data-driven mt systems for improved sign language translation. 2007.

Carol Neidle, Stan Sclaroff, and Vassilis Athitsos. Signstream: A tool for linguistic and computer vision research on visual-gestural language data. *Behavior Research Methods, Instruments, & Computers*, 33(3):311–320, 2001.

Trung Nguyen, Mary Czerwinski, and Dan Lee. Compaq quicksource: Providing the consumer with the power of artificial intelligence. In *Proceedings of the The Fifth Conference on Innovative Applications of Artificial Intelligence*, pages 142–151. AAAI Press, 1993.

Bertoldi Nicola, Tiotto Gabriele, Prinetto Paolo, Piccolo Elio, Nunnari Fabrizio, Lombardo Vincenzo, Mazzei Alessandro, Damiano Rossana, Lesmo Leonardo, and Andrea Del Principe. On the creation and the annotation of a large-scale italian-lis parallel corpus. In *Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CLST-LREC 2010)*, number 4, 2010.

Eric H Nyberg and Teruko Mitamura. The kant system: Fast, accurate, high-quality translation in practical domains. In *Proceedings of the 14th conference on Computational linguistics-Volume 3*, pages 1069–1073. Association for Computational Linguistics, 1992.

Franz Josef Och and Hermann Ney. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447. Association for Computational Linguistics, 2000.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51, 2003.

Stevan Ostrogonac, Dragiša Mišković, Milan Sečujski, Darko Pekar, and Vlado Delić. A language model for highly inflective non-agglutinative languages. In *2012 IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics*, pages 177–181. IEEE, 2012.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

Jordi Porta, Fernando López-Colino, Javier Tejedor, and José Colás. A rule-based translation from written spanish to spanish sign language glosses. *Computer Speech & Language*, 28 (3):788–811, 2014.

Siegmund Prillwitz, Regina Leven, Heiko Zienert, Thomas Hanke, and Jan Henning. Hamburg notation system for sign languages. an introductory guide, hamnosys version 2.0. *Signum, Seedorf, Germany*, 1989.

Éva Sáfár and Ian Marshall. Sign language translation via drt and hpsg. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 58–68. Springer, 2002.

Rubén San-Segundo, R Barra, R Córdoba, LF D'Haro, F Fernández, Javier Ferreiros, Juan Manuel Lucas, Javier Macías-Guarasa, Juan Manuel Montero, and José Manuel Pardo. Speech to sign language translation system for spanish. *Speech Communication*, 50(11-12):1009–1020, 2008.

Rubén San Segundo Hernández, Veronica Lopez Ludeña, Raquel Martin Maganto, David Sánchez, and Adolfo García. Language resources for spanish-spanish sign language (lse) translation. 2010.

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61: 85–117, 2015.

Holger Schwenk, Marta R Costa-Jussa, and José AR Fonollosa. Smooth bilingual *n*-gram translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

Claude E Shannon and Warren Weaver. The mathematical theory of communication. 1949. *Urbana, IL: University of Illinois Press*, 1963.

Stuart M Shieber. *An introduction to unification-based approaches to grammar*. Microtome Publishing, 2003.

John Sinclair. *Corpus, concordance, collocation*. Oxford University Press, 1991.

Dan I Slobin, Nini Hoiting, Michelle Anthony, Yael Biederman, Marlon Kuntze, Reyna Lindert, Jennie Pyers, Helen Thumann, and Amy Weinberg. Sign language transcription at the level of meaning components: The berkeley transcription system (bts). *Sign Language & Linguistics*, 4(1):63–104, 2001.

D Speers. Representation of american sign language for machine translation. 2002.

D Stein, J Bungeroth, and H Ney. Morpho-syntax based statistical methods for sign language translation. In *11th Annual conference of the European Association for Machine Translation, Oslo, Norway*, pages 169–177. Citeseer, 2006.

Daniel Stein, Christoph Schmidt, and Hermann Ney. Sign language machine translation overkill. In *Federico M, Lane I, Paul M, Yvon F (eds) International workshop on spoken language translation*, pages 337–344, Paris, France, 2010. URL http://mt-archive.info/IWSLT-2010-Stein.pdf.

William C Stokoe, Dorothy C Casterline, and Carl G Croneberg. *A dictionary of American Sign Language on linguistic principles*. Linstok Press, 1976.

William C Stokoe Jr. Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of deaf studies and deaf education*, 10(1):3–37, 2005.

Andreas Stolcke. Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*, 2002.

Nicolas Stroppa and Andy Way. Matrex: Dcu machine translation system for iwslt 2006. In *International Workshop on Spoken Language Translation (IWSLT) 2006*, 2006.

V Sutton. Sutton movement shorthand: a quick visual easy-to-learn method of recording dance movement-book one: The classical ballet key. *The Movement Shorthand Society, Irvine, Ca*, 1973.

Valerie Sutton. *Lessons in Sign Writing: Textbook*. SignWriting, 1995.

Rachel Sutton-Spence and Bencie Woll. *The linguistics of British Sign Language: an introduction*. Cambridge University Press, 1999.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218, 2012.

Peter Toma. Systran as a multilingual machine translation system. In *Proceedings of the Third European Congress on Information Systems and Networks, Overcoming the language barrier*, pages 569–581, 1977.

Sneha Tripathi and Juran Krishna Sarkhel. Approaches to machine translation. 2010.

Arturo Trujillo. Transfer machine translation. In *Translation Engines: Techniques for Machine Translation*, pages 121–166. Springer, 1999.

Lynette Van Zijl and Andries Combrink. The south african sign language machine translation project: issues on non-manual sign generation. In *Proceedings of the 2006 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries*, pages 127–134. South African Institute for Computer Scientists and Information Technologists, 2006.

Tony Veale and Alan Conway. Cross modal comprehension in zardoz an english to sign-language translation system. In *Proceedings of the Seventh International Workshop on Natural Language Generation*, pages 249–252. Association for Computational Linguistics, 1994.

Ian H Witten and Timothy C Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *Ieee transactions on information theory*, 37 (4):1085–1094, 1991.

Chung-Hsien Wu, Hung-Yu Su, Yu-Hsien Chiu, and Chia-Hung Lin. Transfer-based statistical translation of taiwanese sign language using pcfg. *ACM transactions on Asian language information processing (TALIP)*, 6(1):1, 2007.

Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In *Proceedings of the 39th annual meeting on association for computational linguistics*, pages 523–530. Association for Computational Linguistics, 2001.

Richard Zens, Franz Josef Och, and Hermann Ney. Phrase-based statistical machine translation. In *Annual Conference on Artificial Intelligence*, pages 18–32. Springer, 2002.

Liwei Zhao, Karin Kipper, William Schuler, Christian Vogler, Norman Badler, and Martha Palmer. A machine translation system from english to american sign language. In *Conference of the Association for Machine Translation in the Americas*, pages 54–67. Springer, 2000.

Ελένη Ευθυμίου, Γρηγόρης Σταϊνχάουερ, Βασίλης Κουρμπέτης, and Μαριάννα Κατσογιάννου. Το έργο ΝΟΗΜΑ για την Λεξικογράφηση της Ελληνικής Νοηματικής Γλώσσας (ΕΝΓ). *Λογοπλοήγηση: Ενημερωτικό δελτίο ανθρωπίνων δικτύων γλωσσικής τεχνολογίας*, 7:14–17, 2000.

Ελένη Ευθυμίου and Ευίτα Φωτεινέα. *Μαθαίνω τα Νοήματα*. Ινστιτούτο Επεξεργασίας του Λόγου / Ε.Κ. "Αθηνά", 2006.

Μαριάννα Κατσογιάννου. Η Ελληνική Νοηματική Γλώσσα, ένα άλλο μέσο επικοινωνίας. *Translatum Journal*, (2), 2002.

Θεόδωρος Μαγγανάρης. *Εγχειρίδιο νοηματικής γλώσσας*. Υπουργείο Υγείας και Πρόνοιας, 2002.

Συντομογραφίες

Short Term	English	Greek
3D	Three Dimensional	Τρισδιάστατο
3D avatar	A three(3) Dimensional digital persona	Τρισδιάστατο ανθρωποειδές
API	Application Program Interfaces	Διεπαφές προγραμματισμού εφαρμογής
ASL	American Sign Language	Αμερικανική Νοηματική Γλώσσα
BLUE	bilingual evaluation understudy	Δίγλωσση μελέτη αξιολόγησης
BSL	British Sign Language	Βρετανική Νοηματική Γλώσσα
BTS	Berkeley Transcription System	Σύστημα μεταγραφής του Μπέρκλεϋ (Berkeley University)
CAD	Computer Aided Design	Σχεδιασμός με υπολογιστή
CBR	Case Based Reasoning	Συλλογιστική Βάσει Υποθέσεων
DDBMTs	Data-Driven Based MT Systems	Συστήματα Μηχανικής Μετάφρασης με δεδομένα
DGS	Deutsche/German Sign Language	Γερμανική Νοηματική Γλώσσα
DNN	Deep Neural Network	Βαθύ Νευρωνικό Δίκτυο
EBMT	Example-Based Machine Translation	Μηχανική Μετάφραση Βάσει Παραδείγματος
ELAN	EUDICO Linguistic Annotator	EUDICO Επισημειωτής Γλωσσικών Χαρακτηριστικών
GIZA++	Toolkit for training of statistical translation models	Εργαλειοθήξη για την εκπαίδευση στατιστικών μοντέλων μετάφρασης
gloss	A brief notation, the meaning of a word	Σύντομο σχόλιο για την απόδοση της σημασίας ενός λήμματος
glossa	Glossa	Γλώσσα
GSL	Greek Sign Language	Ελληνική Νοηματική Γλώσσα

HamNoSys	Hamburg Notation System	Σύστημα επισημείωσης του Αμβούργου
HMT	Hybrid Machine Translation	Υβριδικό Σύστημα Μηχανικής Μετάφρασης
HTML	Hypertext Markup Language	Γλώσσα Σήμανσης Υπερκειμένου
ICT	Information and Communications Technology	Τεχνολογία Πληροφοριών και Επικοινωνιών
IDGS	Institute for German Sign Language and Communication for the Deaf at the University of Hamburg	Ινστιτούτο Γερμανικής Νοηματικής Γλώσσας και Επικοινωνίας για τους Κωφούς στο Πανεπιστήμιο του Αμβούργου
iLEX	integrated Lexicon	ολοκληρωμένο Λεξικό
IPA	International Phonetic Alphabet	Διεθνές Φωνητικό Αλφάβητο
KD	Discrimination Coefficient	Συντελεστής Διάκρισης
LM	Language Model	Γλωσσικό Μοντέλο
MM	Machine Traslation	Μηχανική Μετάφραση
MOSES	Moses is a free software, statistical machine translation	Ο Moses είναι ένα ελεύθερο λογισμικό, στατιστικής μηχανικής μετάφρασης
NLP	Natural Language Processing	Φυσική Επεξεργασία Γλώσσας
NLTK	Natural Language Toolkit	Εργαλειοθήκη Φυσικής Γλώσσας
OpenGL	Open Graphics Library	Ανοιχτή Βιβλιοθήκη Γραφικών
PCFG	Probabilistic Context-Free Grammar	Πιθανοτική Γραμματική Χωρίς Πλαίσιο
POS	Part Of Speech	Μέρος Του Λόγου (Γλωσσολογικός Χαρακτηρισμός Λήμματος)
ppl	perplexity	περιπλοκή
RBMT	Rule Based Machine Translation	Μηχανική Μετάφραση Βασισμένη σε Κανόνες
SiGML	The HamNoSys into XML form	Το HamNoSys σε μορφή XML
SL	Sign Language	Νοηματική Γλώσσα
SMT	Statistical Machine Translation	Στατιστική Μηχανική Μετάφραση
SRILM	SRI Language Modeling (SRILM) toolkit	Εργαλειοθήκη SRI για τη Μοντελοποίηση Γλωσσών (SRILM)
STEP	Scripting Technology for Embodied Persona	Τεχνολογία σεναρίων για ενσωματωμένο ανθρωποειδές

VRML	Virtual Reality Modeling Language	Γλώσσα μοντελοποίησης εικονικής πραγματικότητας
XCES	Corpus Encoding Standard for XML	Πρότυπο κωδικοποίησης Corpus για XML
XML	eXtensible Markup Language	επεκτάσιμη Γλώσσα Σήμανσης
ΓΠ	Language Source	Γλώσσα Πηγή
ΓΣ	Language Target	Γλώσσα Στόχος
ΕΝΓ	Greek Sign Language	Ελληνική Νοηματική Γλώσσα
ΙΝΓ	Spain Sign Language	Ισπανική Νοηματική Γλώσσα
ΝΓ	Sign Language	Νοηματική Γλώσσα
ΣΓΜ	Statistical Language Models	Στατιστικά Γλωσσικά Μοντέλα
ΣΜΕΝΓ	Short Transcription of the Greek Sign Language	Σύντομης Μεταγραφής της Ελληνικής Νοηματικής Γλώσσας
ΣΜΜ	Statistical Machine Traslation	Σύστημα Μηχανικής Μετάφρασης

Βιογραφικό Σημείωμα

Προσωπικά Στοιχεία

Ονομα:	**Δημήτριος**
Επίθετο:	**Κουρεμένος**
Πατρώνυμο:	Αναστάσιος
Τόπος | Ημ. γέννησης:	Αθήνα | 22 Απριλίου 1974

Πληροφορίες επικοινωνίας

Διεύθυνση:	Αβύδου 111 , Άνω Ιλίσια, Αθήνα
email:	dkourem@gmail.com
LinkedIn:	https://www.linkedin.com/in/dkourem/

Εκπαίδευση

2001-2020 **Υποψήφιος Διδάκτορας στη Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών ΕΜΠ**
Πεδίο: *"Μοντελοποίηση της Ελληνικής Νοηματικής Γλώσσας για εφαρμογή σε συστήματα στατιστικής μηχανικής μετάφρασης"*
Επιβλέπων: Καθ. Ανδρέας - Γεώργιος Σταφυλοπάτης

1999-2000 **Διεπιστημονικό Διαπανεπιστημιακό Πρόγραμμα Μεταπτυχιακών Σπουδών στην "Γλωσσική Τεχνολογία" στη σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών ΕΜΠ και σχολή Φιλοσοφικής Μαθηματικής Γλωσσολογίας του Καποδιστριακού Πανεπιστημίου Αθηνών**
Πεδίο: *"Γλωσσική Τεχνολογία"*

1992-1998 **Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών ΕΜΠ**
Διπλωματική εργασία: *"Πολυμέσα και δικτυακά πολυμέσα με εφαρμογή στα Άτομα με ειδικές ανάγκες ακοής"*

Κατάλογος Δημοσιεύσεων

Άρθρα σε Επιστημονικά Περιοδικά

1. Kouremenos, D., Ntalianis, K., and Kollias, S. (2018). **A novel rule based machine translation scheme from Greek to Greek Sign Language: Production of different types of large corpora and Language Models evaluation.** *Computer Speech & Language, 51, 110-135.*

2. Efthimiou, E., Fotinea, S. E., Dimou, A. L., Goulas, T., and Kouremenos, D. (2016). **From grammar-based MT to post-processed SL representations.** *Universal Access in the Information Society, 15(4), 499-511.*

3. Kouremenos, D., Fotinea, S. E., Efthimiou, E., and Ntalianis, K. (2010). **A prototype Greek text to Greek Sign Language conversion system.** *Behaviour & Information Technology, 29(5), 467-481.*

Ανακοινώσεις σε Συνέδρια

1. Kouremenos, D., Ntalianis, K. S., Siolas, G., and Stafylopatis, A. (2018). **Statistical Machine Translation for Greek to Greek Sign Language Using Parallel Corpora Produced via Rule-Based Machine Translation.** *In CIMA@ ICTAI (pp. 28-42).*

2. D. KOUREMENOS, K. NTALIANIS, and A. STAFYLOPATIS (2018). **A Rule Based Machine Translation Scheme from Greek to Greek Sign Language: Production of Different Types of Large Corpora.** *WSEAS TRANSACTIONS on COMPUTERS, vol. 17.*

3. Fotinea, Stavroula-Evita, Eleni Efthimiou, and Dimitris Kouremenos (2005). **Generating linguistic content for Greek to GSL conversion.** *Proceedings of the HERCMA-2005 Conference (The 7th Hellenic European Conference on Computer Mathematics and its Applications), Athens.*