ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πτυχιακή Εργασία

Στατιστική μηχανική μετάφραση από τα Αρχαία Ελληνικά στα Αγγλικά

Βασίλειος Καλογηράς

Επιβλέπων: Ίων Ανδρουτσόπουλος
Βοηθός επιβλέπων: Δημήτριος Μαυροειδής

Περιεχόμενα

Κεφάλαιο 1: Εισαγωγή

1.1 Αντικείμενο της εργασίας

Σκοπός της εργασίας ήταν η δημιουργία ενός συστήματος στατιστικής μηχανικής μετάφρασης (statistical machine translation) [1] που να μεταφράζει κείμενα από τα Αρχαία Ελληνικά στα Αγγλικά. Στα συστήματα αυτού του είδους, οι μεταφράσεις παράγονται χρησιμοποιώντας στατιστικά μοντέλα, των οποίων οι παράμετροι εκτιμώνται από παράλληλα σώματα κειμένων (parallel corpora), δηλαδή συλλογές κειμένων στις οποίες κάθε κείμενο διατίθεται σε πολλές γλώσσες, όπως στα πρακτικά του Ευρωπαϊκού Κοινοβουλίου [2] ή του Καναδικού Κοινοβουλίου [3]. Για τους σκοπούς της εργασίας, χρησιμοποιήσαμε αρχαία ελληνικά κείμενα και τις αγγλικές τους μεταφράσεις, όπως παρέχονται από την ψηφιακή βιβλιοθήκη Perseus του Πανεπιστημίου Tufts [4].

1.2 Διάρθρωση

Στο κεφάλαιο 2 παρουσιάζονται συνοπτικά η θεωρία και οι βασικές έννοιες πάνω στις οποίες στηρίζεται η εργασία. Στο κεφάλαιο 3 περιγράφονται τα δεδομένα μας, τα εργαλεία που χρησιμοποιήθηκαν και η αρχιτεκτονική του συστήματος που αναπτύχθηκε. Στο κεφάλαιο 4 παρουσιάζονται τα πειράματα της εργασίας, καθώς και τα αποτελέσματά τους. Στο κεφάλαιο 5 συνοψίζονται τα συμπεράσματα. Τέλος, προτείνονται ιδέες για μελλοντικές βελτιώσεις.

1.3 Ευχαριστίες

Αρχικά, ευχαριστώ πολύ τον επιβλέποντα καθηγητή μου, κ. Ίωνα Ανδρουτσόπουλο, για τις χρήσιμες συμβουλές και οδηγίες του όλη αυτή την περίοδο. Ευχαριστώ θερμότατα τον υποψήφιο διδάκτορα Δημήτριο Μαυροειδή, για τη διαρκή του βοήθεια και καθοδήγηση σε όλα τα στάδια της εργασίας. Ακόμη, θα ήθελα να ευχαριστήσω την κ. Βασιλική Κοτίνη, για το χρόνο που αφιέρωσε στην αξιολόγηση των μεταφρασμένων προτάσεων και τις συμβουλές της για τη βελτίωση του συστήματός μας. Τέλος, ευχαριστώ τα μέλη της Ομάδας Επεξεργασίας Φυσικής Γλώσσας, για τις ενδιαφέρουσες συζητήσεις και τις γνώσεις που αποκόμισα από αυτές.

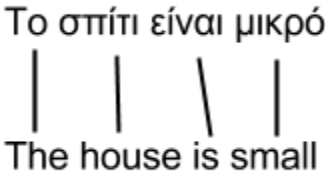
Κεφάλαιο 2: Θεωρητικό υπόβαθρο

Σε αυτό το κεφάλαιο παρουσιάζονται συνοπτικά η θεωρία και οι βασικές έννοιες πάνω στις οποίες στηρίζεται η εργασία.

2.1 Εκπαίδευση

2.1.1 Γλωσσικά και μεταφραστικά μοντέλα

Τα πρώτα στατιστικά μοντέλα μετάφρασης θεωρούσαν τις λέξεις ξεχωριστές μονάδες οι οποίες μπορούν να μεταφραστούν, παραλειφθούν, εισαχθούν και αναδιαταχθούν στην αρχική πρόταση, ώστε να προκύψει η μετάφρασή της. Έτσι, η μετάφραση ανάμεσα σε ένα ζευγάρι προτάσεων αποτελεί ουσιαστικά αντιστοίχηση λέξεων, θεωρώντας ότι λέξεις που διαγράφονται ή εισάγονται αντιστοιχούν σε ειδικές ψευδο-λέξεις της άλλης ή της αρχικής, αντίστοιχα, γλώσσας.



1. Παράδειγμα αντιστοίχησης λέξεων για μία πρόταση

Επειδή όμως, δε γνωρίζουμε ποια είναι η κατάλληλη αντιστοιχία εξ αρχής, ορίζουμε ένα μοντέλο που θα παράγει διαφορετικές μεταφράσεις μιας πρότασης, κάθε μία με διαφορετική πιθανότητα να είναι σωστή. Το πρώτο μοντέλο της IBM (IBM Model 1), για κάθε αρχική πρόταση F με λέξεις $(f_1,...,f_n)$ υπολογίζει χονδρικά την πιθανότητα η μετάφραση να είναι η πρόταση $E = (e_1,...,e_k)$ ως το γινόμενο των επιμέρους πιθανοτήτων $p(e_i|f_j)$. Τις πιθανότητες τις υπολογίζουμε βασιζόμενοι στη συχνότητα ταυτόχρονης εμφάνισης των ζευγαριών αυτών σε παράλληλα κείμενα.

Το μοντέλο αυτό δεν λαμβάνει υπόψη του το περικείμενο κάθε λέξης, συχνά όμως αυτό επηρεάζει ποια μετάφραση είναι ορθότερη. Για παράδειγμα, οι sacred και shrines αποτελούν σωστές μεταφράσεις της λέξης ἱερά. Ποια είναι όμως η καταλληλότερη; Εξαρτάται από το περιεχόμενο της πρότασης που μεταφράζουμε. Στην πρόταση, «ἡ γὰρ ἱερὰ τοῦ ἀπόλλωνος ἑβδομὰς ἀναλώσει τὴν ἡμέραν πρότερον ἢ λόγῳ τὰς δυνάμεις αὐτῆς ἁπάσας ἐπεξελθεῖν», το sacred αποτελεί καλύτερη μετάφραση. Πώς μπορεί ένα σύστημα να το γνωρίζει αυτό όμως;

Τη λύση έρχεται να δώσει το γλωσσικό μοντέλο (language model) [1] το οποίο μετρά πόσο πιθανό είναι μία ακολουθία λέξεων να είναι γλωσσικά ορθή. Συγκεκριμένα, το γλωσσικό μοντέλο επηρεάζει τόσο την επιλογή όσο και τη θέση μίας λέξης/φράσης στην υποψήφια μετάφραση. Τα πιο συνηθισμένα γλωσσικά μοντέλα είναι γλωσσικά μοντέλα ν-γραμμάτων, όπου ως ν-γραμμα εννοούμε μία ακολουθία από ν συνεχόμενες λέξεις. Για μία πρόταση `e` η πιθανότητά της `p(e)` να είναι σωστή θα ισούται με:

$$p(e) = p(e_1, e_2, ...e_v) = p(e_1)p(e_2|e_1)...p(e_v|e_1, e_2..., e_{v-1})$$

Ένα μοντέλο ν-γραμμάτων κάνει την υπόθεση πως κάθε λέξη εξαρτάται μόνο από τις ν-1 προηγούμενες. Έτσι, σε ένα μοντέλο 3-γραμμάτων ο παραπάνω υπολογισμός γίνεται:

$$p(e) \simeq p(e_1)p(e_2|e_1)...p(e_v|e_{v-2}, e_{v-1})$$

Η εκτίμηση των επιμέρους πιθανοτήτων των 3-γραμμάτων γίνεται μετρώντας τη συχνότητα εμφάνισής τους σε κείμενα εκπαίδευσης. Στην περίπτωση μας, αφού η μετάφραση γίνεται από τα Αρχαία Ελληνικά στα Αγγλικά, γίνεται καταμέτρηση των συχνοτήτων σε όλα τα αγγλικά κείμενα που έχουμε διαθέσιμα.

Η αξιολόγηση ενός γλωσσικού μοντέλου `LM` μιας γλώσσας `L` γίνεται υπολογίζοντας τη διασταυρωμένη εντροπία του `LM` και της `L`, που μπορεί να εκτιμηθεί ως εξής [6]:

$$H(L, LM) = -\frac{1}{n} log\, p_{LM}(W_1^n)$$

όπου W_1^n είναι μια μεγάλη ακολουθία n λέξεων της `L`, στην πράξη ένα μεγάλο σώμα κειμένων της `L`. Όσο χαμηλότερη η διασταυρωμένη εντροπία, τόσο καλύτερο το γλωσσικό μοντέλο. Στην πράξη, συχνότερα χρησιμοποιείται η περιπλοκή (perplexity), που ορίζεται ως εξής:

$$perplexity(L, LM) = 2^{H(L, LM)}$$

Τα μοντέλα 3-γραμμάτων είναι συνήθως καλύτερα από τα μοντέλα 2-γραμμάτων ή 1-γραμμάτων, αλλά για ν > 3 η βελτίωση αρχίζει και γίνεται μικρή [7]. Συχνό πρόβλημα των γλωσσικών μοντέλων, ιδιαίτερα όσο μεγαλώνει το ν, είναι ότι πολλά ν-γράμματα δεν εμφανίζονται στο σώμα εκπαίδευσης. Για να αποφύγουμε να δώσουμε μηδενική πιθανότητα σε κάποιο ν-γραμμα (οπότε μηδενίζεται και το γινόμενο πιθανοτήτων που υπολογίζει το γλωσσικό μοντέλο) έχουν αναπτυχθεί διάφορες τεχνικές εξομάλυνσης. Έτσι, υπάρχουν τεχνικές που δίνουν μία μικρή πιθανότητα σε κάθε ν-γραμμα που δεν έχουμε συναντήσει (add-one smoothing, Good-Turing smoothing) και μέθοδοι χρησιμοποίησης ή μίξης διαφορετικής τάξης μοντέλων (interpolation, back-off). Τέλος, οι πιο διαδεδομένες μέθοδοι εξομάλυνσης αυτή τη στιγμή (Kneser-Ney smoothing, modified Kneser-Ney smoothing) λαμβάνουν υπόψη και τον αριθμό των διαφορετικών εμφανίσεων μιας λέξης σε συνδυασμό με την προηγούμενη από αυτήν λέξη [1].

2.1.2 Μοντέλο μετάφρασης κατά φράσεις

Τελικά, για την εύρεση της πιθανότητας μετάφρασης μίας πρότασης, στην περίπτωσή μας από τα Αρχαία Ελληνικά (a) στα Αγγλικά (e), εκμεταλλευόμαστε το νόμο του Bayes από όπου προκύπτει:

$$argmax_e p(e|a) = argmax_e p(a|e)p(e)$$

Αυτό μας δίνει την δυνατότητα ορισμού ξεχωριστού γλωσσικού $p(e)$ και μεταφραστικού $p(a|e)$ μοντέλου. Ο "ανάποδος" αυτός τρόπος παρουσίασης είναι γνωστός ως μοντέλο του θορυβώδους καναλιού και είναι δανεισμένος από τη θεωρία πληροφορίας. Η ιδέα είναι ότι αρχικά είχαμε μία πρόταση στα Αγγλικά αλλά της συνέβη κάποια παραμόρφωση και πριν φτάσει σε εμάς μετατράπηκε σε Αρχαία Ελληνικά. Τώρα, σκοπός μας είναι να ανακατασκευάσουμε την αρχική αγγλική πρόταση.

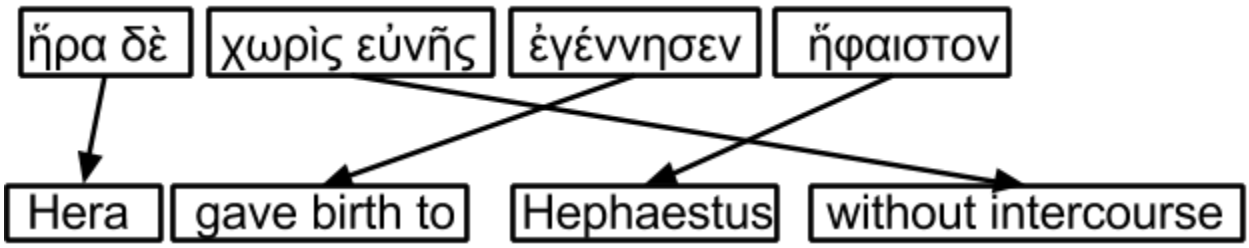
Μέχρι τώρα έχουμε βασίσει τη μετάφρασή μας στο πρώτο μοντέλο IBM, το οποίο όμως έχει αρκετές αδυναμίες. Για παράδειγμα, για την αρχική μας πρόταση «Το σπίτι είναι μικρό» οι μεταφράσεις "The house is small" και "The small house is" θα έχουν ακριβώς την ίδια πιθανότητα να είναι σωστές. Αυτό και άλλα προβλήματα, όπως ποιος είναι ο πιθανότερος αριθμός λέξεων που χρειάζονται για να μεταφραστεί η κάθε λέξη ή πόσο πιθανό είναι δύο λέξεις να βρίσκονται κοντά σε μία μεταφρασμένη πρόταση, ήρθαν να λύσουν τα μοντέλα IBM 2-5 [8]. Πιο συγκεκριμένα, το δεύτερο μοντέλο IBM λύνει το πρόβλημα στη διάταξη. Έτσι, για κάθε λέξη μίας πρότασης τώρα υπολογίζεται και η πιθανότητα από μία θέση `i` που βρίσκεται αρχικά να μετακινηθεί σε μία άλλη θέση `j` αφού μεταφραστεί. Το τρίτο μοντέλο IBM εισάγει την έννοια της παραγωγής (fertility). Δηλαδή, πόσες λέξεις είναι πιθανότερο να παραχθούν κατά τη μετάφραση μίας συγκεκριμένης λέξης. Στο τέταρτο μοντέλο IBM, η θέση όπου τοποθετούνται οι μεταγενέστερες αγγλικές λέξεις (στην περίπτωσή μας) που παράχθηκαν από μια αρχαιοελληνική λέξη εξαρτάται από τη θέση των προηγούμενων λέξεων που δημιούργησε η ίδια λέξη (relative alignment model). Αν είχαμε ορίσει μαθηματικά τα προηγούμενα μοντέλα IBM, θα παρατηρούσαμε ότι το τρίτο και τέταρτο κάνουν μία παράλειψη. Δίνουν θετική πιθανότητα σε προτάσεις που τοποθετούν περισσότερες από μία λέξεις σε μία θέση, πράγμα αδύνατο στην πραγματικότητα. Για το λόγο αυτό το πέμπτο μοντέλο IBM αποθηκεύει ποιες θέσεις παραμένουν κενές και επιτρέπει να τοποθετηθούν καινούργιες λέξεις μόνο σε αυτές.

Τα μοντέλα αυτά χρησιμοποιούνταν για αρκετό διάστημα, αλλά κατόπιν άρχισαν να χρησιμοποιούνται μοντέλα μετάφρασης κατά φράσεις [5]. Ο στόχος των τελευταίων είναι να μειώσουν τους περιορισμούς της μετάφρασης κατά λέξεις χρησιμοποιώντας φράσεις (όχι μεμονωμένες λέξεις) ως μονάδες μετάφρασης. Ως φράση, ορίζεται μία οποιαδήποτε συνεχόμενη ακολουθία λέξεων.

Ένα από τα πλενεκτήματα της μετάφρασης κατά φράσεις είναι η δυνατότητα μετάφρασης ιδιωματισμών. Έτσι, για την πρόταση "It's rainning cats and dogs", αντί να δημιουργήσουμε τη μετάφραση «Βρέχει γάτες και σκύλους» μπορούμε να θεωρήσουμε και τις τέσσερις λέξεις σαν ένα ενιαίο σύνολο που μεταφράζεται ως «ρίχνει καρεκλοπόδαρα».

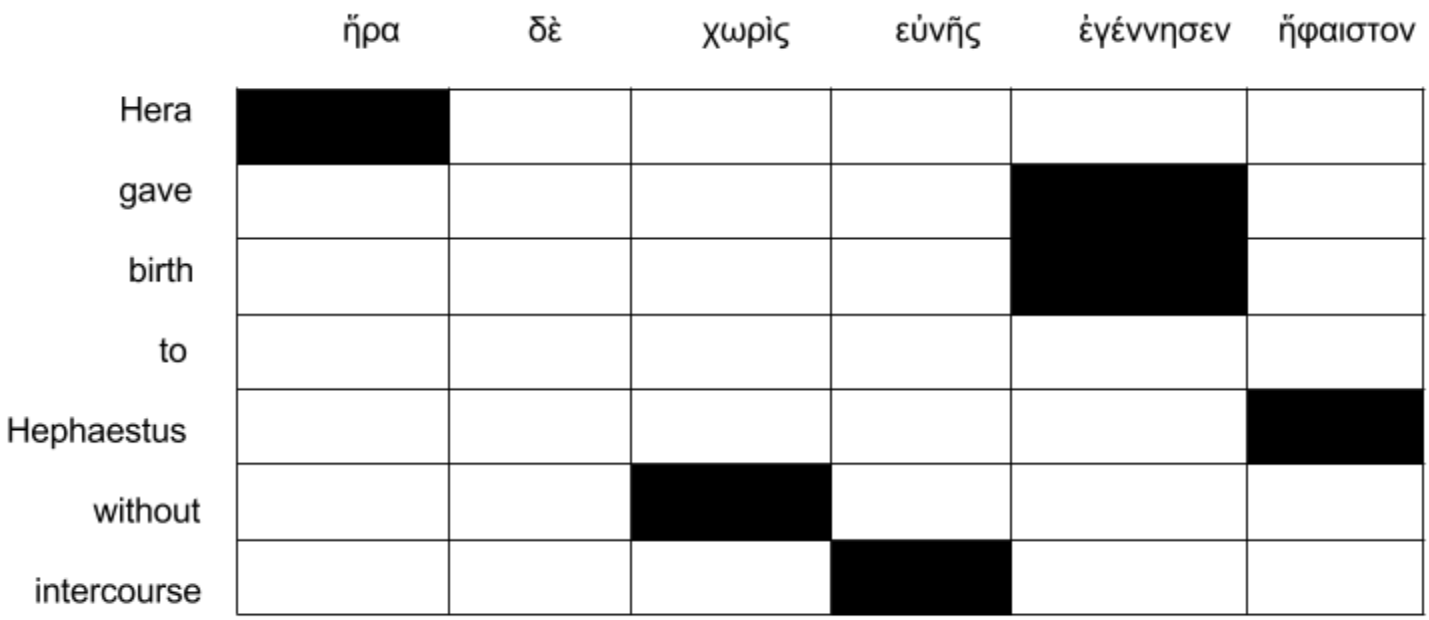
Λόγω της υπεροχής της, τα τελευταία χρόνια, η μετάφραση κατά φράσεις αποτελεί το καθιερωμένο μοντέλο μετάφρασης και αυτό που χρησιμοποιούμε και εμείς στην εργασία. Τα μοντέλα της IBM, όμως, εξακολουθούν να χρησιμοποιούνται προκειμένου να γίνει αντιστοίχηση λέξεων (word alignment) σε παράλληλα σώματα κειμένων.

Πιο συγκεκριμένα, ας θεωρήσουμε το παράδειγμα:



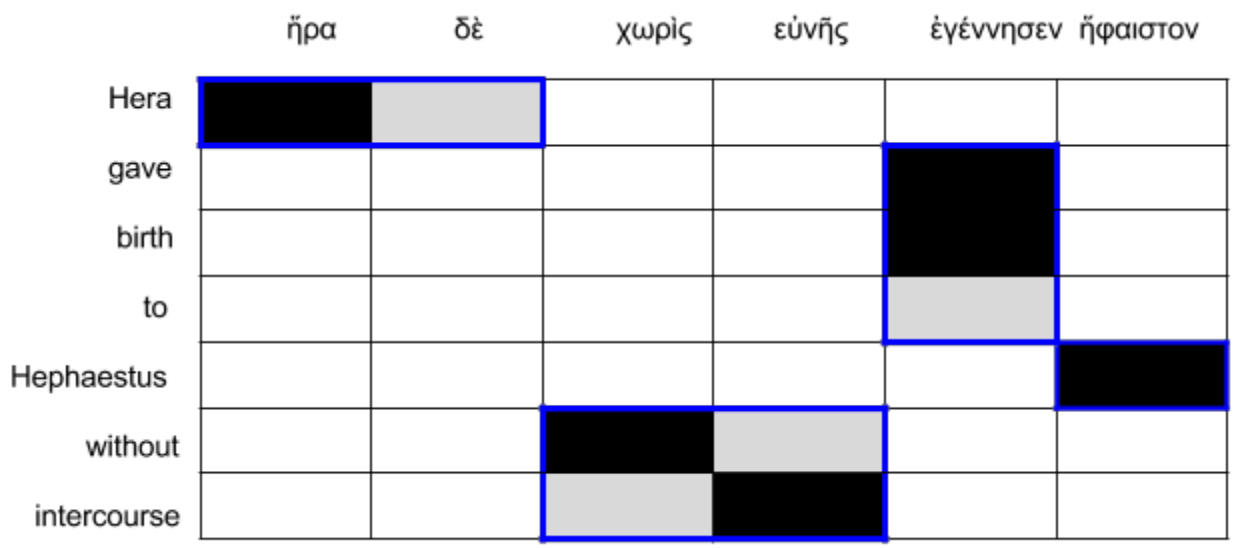
2.Αντιστοίχηση με το μοντέλο φράσεων

Όπως παρατηρούμε, η αρχαιοελληνική πρόταση χωρίζεται σε φράσεις, οι οποίες στη συνέχεια μεταφράζονται και τελικά αναδιατάσσονται. Αρχικά χρησιμοποιούμε τα μοντέλα IBM που περιγράψαμε για να εξαγάγουμε τις πιθανότερες αντιστοιχίες λέξεων.



3. Πίνακας αντιστοίχησης λέξεων

Έπειτα, εξάγουμε όλα τα δυνατά ζευγάρια φράσεων χωρίς να παραβιάζουμε την αντιστοίχηση `A` των λέξεών μας. Ένα ζευγάρι `(a,e)` φράσεων είναι συνεπές ως προς μία αντιστοίχηση λέξεων `A` αν όλες οι αντιστοιχίσεις των λέξεών του `a` γίνονται με λέξεις που υπάρχουν στο σύνολο `e`. Έτσι οι ευθυγραμμίσεις `ἥρα δὲ - Hera, ἐγέννησεν ἥφαιστον - gave birth to Hephaestus, χωρὶς εὐνῆς ἐγέννησεν ἥφαιστον - gave birth to Hephaestus without intercourse` θα είναι σωστές, ενώ οι `ἥρα δὲ - Hera gave intercourse, εὐνῆς ἐγέννησεν ἥφαιστον - Hephaestus without intercourse` όχι.



4. Πίνακας αντιστοίχησης φράσεων

Τελικά με βάση τα κείμενα που έχουμε στη διάθεσή μας υπολογίζουμε τις πιθανότητες του κάθε ζεύγους φράσεων (βλ. [1] για περισσότερες λεπτομέρειες).

2.2 Ρύθμιση παραμέτρων (tuning)

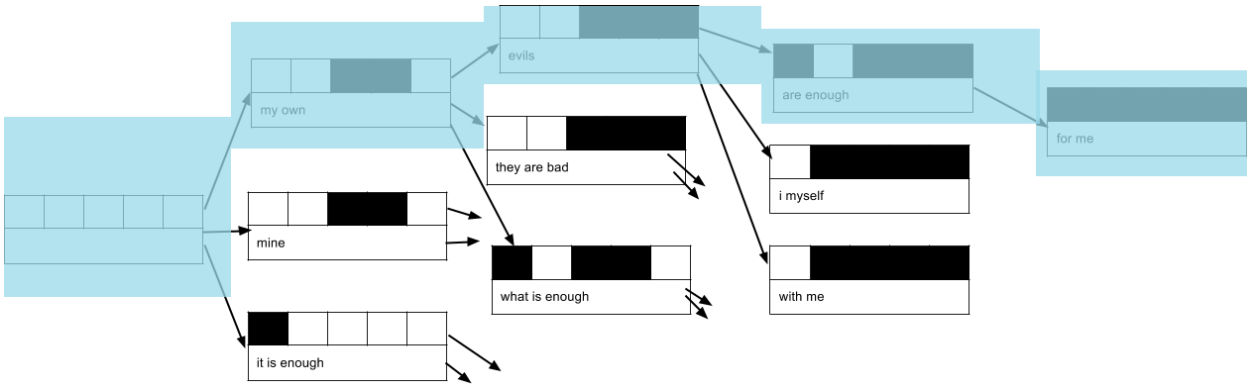
Όσο καλύτερο και πληρέστερο πίνακα μετάφρασης φράσεων διαθέτουμε, τόσο καλύτερες μπορούμε να αναμένουμε ότι θα είναι και οι παραγόμενες μεταφράσεις του συστήματός μας. Ωστόσο, οι μετάφραση είναι συνάρτηση και άλλων παραμέτρων. Για το λόγο αυτό το βήμα της «ρύθμισης» τους (tuning) είναι ένα από τα σημαντικότερα στη διαδικασία εκπαίδευσης. Συγκεκριμένα, τέσσερις είναι οι παράμετροι που επηρεάζουν την πιθανότητα μιας μετάφρασης να είναι η επικρατέστερη. Το μεταφραστικό μοντέλο, εκφράζει την πιθανότητα μία αρχαιοελληνική και μία αγγλική φράση να αποτελούν καλή μετάφραση η μία της άλλης. Το γλωσσικό μοντέλο, δηλώνει τη γραμματική ποιότητα μιας πρότασης. Το μοντέλο στρέβλωσης, αντικατοπτρίζει την αναδιάταξη των λέξεων μέσα σε μία πρόταση αφού έγινε η μετάφρασή της. Συνήθως, όσο περισσότερες οι αναδιατάξεις τόσο μεγαλύτερο και το κόστος της μετάφρασης. Τέλος, υπάρχει το μοντέλο ποινής λέξης που «τιμωρεί» παραγόμενες μεταφράσεις πολύ μεγαλύτερες ή μικρότερες σε σχέση με την αρχική πρόταση.

Η διαδικασία της ρύθμισης των βαρών έχει ως εξής. Αρχικά, κατασκευάζουμε ένα σύνολο με παράλληλες προτάσεις που δεν υπήρχαν στα κείμενα εκπαίδευσης (development/ tuning set) και τις τροφοδοτούμε στο σύστημά μας προς μετάφραση. Εν συνεχεία, επαναληπτικά και με βάση τις δοσμένες μεταφράσεις αξιολογούμε αυτές που βγάλαμε εμείς μέσω κάποιας μετρικής (συνήθως BLEU [9]). Ρυθμίζουμε, τα βάρη κάθε παραμέτρου κατάλληλα και επαναξιολογούμε. Τελικά, όταν η περαιτέρω ρύθμιση των βαρών δεν προσφέρει σημαντική βελτίωση στην μετρική μας σταματάμε.

2.3 Αποκωδικοποίηση

Παρουσιάσαμε δύο μοντέλα παραγωγής μεταφράσεων, ένα που βασίζεται σε λέξεις και ένα σε φράσεις. Στόχος της αποκωδικοποίησης στη στατιστική μηχανική μετάφραση είναι να βρει την πιθανότερη μετάφραση με βάση αυτά τα μοντέλα. Συγκεκριμένα, η μετάφραση γίνεται λέξη προς λέξη ή φράση προς φράση, αντίστοιχα. Με αυτό τον τρόπο κάθε στιγμή έχουμε κατασκευασμένα υποψήφια σύνολα (ενδεχομένως ημιτελών) μεταφράσεων που ονομάζονται υποθέσεις. Η κάθε υπόθεση μπορεί να παρασταθεί με μία δομή δεδομένων που περιέχει πληροφορίες όπως: ποιες λέξεις της μετάφρασης έχει δημιουργήσει μέχρι τώρα, ποιες λέξεις του αρχικού κειμένου έχει καλύψει, τι πιθανότητα δίνει η ως τώρα μετάφραση κλπ.

Έστω ότι θέλουμε να μεταφράσουμε την πρόταση «ἀρκεῖ ἐμοὶ τὰ ἐμὰ κακά». Αρχικά, ξεκινάμε με μία κενή υπόθεση και έστω ότι αποφασίζουμε να μεταφράσουμε την τρίτη και τέταρτη λέξη της αρχαιοελληνικής πρότασης ως `my own` ενώ ταυτόχρονα τοποθετούμε τις παραγόμενες αγγλικές λέξεις στην αρχή της μετάφρασης που αναπτύσσουμε. Επίσης, υπολογίζουμε όλες τις πιθανότητες που περιγράψαμε παραπάνω (γλωσσικό μοντέλο, μεταφραστικό κλπ.). Με αυτό τον τρόπο δημιουργούμε μία καινούργια υπόθεση που συνδέεται με την αρχική κενή μας. Ακόμη, παρατηρούμε ότι μπορούμε να ξεκινήσουμε και με τις φράσεις `mine` και `it is enough`. Ομοίως υπολογίζουμε τις πιθανότητες και για αυτές τις επιλογές. Η αποκωδικοποίηση συνεχίζεται αναδρομικά όπως φαίνεται στο σχήμα. Τελικά, αφού έχουμε αναπτύξει πλήρως όλες τις υποθέσεις βρίσκουμε ποιας το τελικό σημείο φέρει την μεγαλύτερη πιθανότητα. Από εκεί, αν ακολουθήσουμε τους δείκτες προς τα πίσω θα βρούμε την μετάφραση που έλαβε το υψηλότερο σκορ που είναι και η ζητούμενη.



5. Αναζήτηση καλύτερης μετάφρασης

Όπως είναι φανερό, η αποκωδικοποίηση είναι ένα αρκετά δύσκολο πρόβλημα αφού με δεδομένη μια είσοδο οι πιθανές μεταφράσεις είναι εκθετικά περισσότερες σε αριθμό. Μάλιστα, είναι αποδεδειγμένο ότι η διαδικασία την αποκωδικοποίησης είναι NP-complete [10]. Κατά

συνέπεια, εφαρμόζονται διάφορες ευρετικές μέθοδοι αναζήτησης, που αν και δεν είναι σίγουρο ότι θα ανακαλύψουν την καλύτερη μετάφραση ελπίζουμε ότι θα βρουν μια μετάφραση πολύ κοντά στην καλύτερη. Μία τέτοια μέθοδος είναι να αναπτύσσουμε κάθε φορά τις k καλύτερες υποθέσεις (beam search). Τέλος, υπάρχουν κι άλλες τεχνικές μείωσης του χώρου υποθέσεων, όπως υπολογισμός μελλοντικού κόστους μετάφρασης, συνδυασμός υποθέσεων κλπ. [1].

Κεφάλαιο 3: Δεδομένα και εργαλεία

3.1 Παράλληλα κείμενα

Με τον όρο παράλληλο σώμα κειμένων *(parallel corpus)* περιγράφουμε μία συλλογή αρχείων (δύο ή περισσότερων γλωσσών) όπου κάθε γραμμή κειμένου (ή πρόταση ή παράγραφος) της μίας γλώσσας αντιστοιχεί στη γραμμή κειμένου με το ίδιο ακριβώς νόημα στις υπόλοιπες γλώσσες. Αυτά τροφοδοτούνται στο προς εκπαίδευση σύστημά μας. Για πολλά ζεύγη γλωσσών υπάρχουν ήδη παράλληλα σώματα, για τα Αρχαία Ελληνικά όμως δεν υπήρχε μέχρι τώρα κάτι αντίστοιχο. Ωστόσο, υπάρχουν διαθέσιμες πολλές μεταφράσεις αρχαίων συγγραφέων στα Αγγλικά που μπορούν να χρησιμεύσουν σαν πρώτη ύλη.

Τα κείμενά που χρησιμοποιήθηκαν στη εργασία προέρχονται από την ψηφιακή βιβλιοθήκη Perseus [4]. Αυτή, αποτελεί δημιουργία του Tufts University[1] της Βοστώνης και ως αποστολή της έχει την προσφορά ελεύθερης πρόσβασης σε μία, όσον το δυνατόν πληρέστερη, καταγραφή της αρχαίας Ελληνικής και Λατινικής γραμματείας και τέχνης - γλωσσικές πηγές, ιστορικά μνημεία, φυσικά αντικείμενα. Η λειτουργία της ξεκίνησε το 1995 και αυτή τη στιγμή διαθέτει μία τεράστια συλλογή κειμένων. Μέρος των κειμένων υπάρχει ελεύθερα διαθέσιμο προς μεταφόρτωση κωδικοποιημένο σε XML.

3.2 Προεπεξεργασία

Η διαδικασία που ακολουθήσαμε ώστε τα κείμενα να εξυπηρετούν την εκπαίδευση ενός αυτόματου συστήματος μετάφρασης ήταν η εξής:

Αρχικά, έπρεπε να ξεχωρίσουμε για κάθε κείμενο το κυρίως μέρος από τα μη-χρήσιμα κομμάτια (π.χ. σχόλια μεταφραστών). Για παράδειγμα, στις παρακάτω εικόνες βλέπουμε το ίδιο απόσπασμα κειμένου στις δύο γλώσσες

[1] "Tufts University." http://www.tufts.edu/

```
<teiHeader type="text" status="new">
<p><milestone ed="p" n="100" unit="line"/>By banishing the man, or by paying back
bloodshed with bloodshed, since it is this blood which brings the tempest on our
city.<milestone ed="p" n="102" unit="card"/></p></sp>
<sp>
<speaker>Oedipus</speaker>
<p>And who is the man whose fate he thus reveals?</p></sp>
<sp>
<speaker>Creon</speaker>
<p>Laius, my lord, was the leader of our land before you assumed control of this
state.</p></sp>
<sp>
<speaker>Oedipus</speaker>
<p><milestone ed="p" n="105" unit="line"/>I know it well—by hearsay, for I never saw
him.</p></sp>
<sp>
<speaker>Creon</speaker>
<p>He was slain, and the god now bids us to take vengeance on his murderers, whoever
they are.</p></sp>
<sp>
```

6. Παράδειγμα XML αρχείου με Αγγλικό κείμενο

```
<titleStmt>
<title type="work" n="Trach.">Trachiniae</title>
<author n="Soph.">Sophocles</author>
<editor role="editor" n="Jebb">Sir Richard Jebb</editor>
</titleStmt>
<l n="100">a)ndrhlatou=ntas h)\ fo/nw| fo/non pa/lin</l>
<l>lu/ontas, w(s to/d' ai(=ma xeima/zon po/lin.</l>
<milestone ed="p" n="102" unit="card"/></sp>
<sp>
<speaker>*oi)di/pous</speaker>
<l>poi/ou ga\r a)ndro\s th/nde mhnu/ei tu/xhn;</l></sp>
<sp>
<speaker>*kre/wn</speaker>
```

```
<l>h)=n h(mi/n, w)=nac, *la/i+o/s poq' h(gemw\n</l>
<l>gh=s th=sde, pri\n se\ th/nd' a)peuqu/nein po/lin.</l></sp>
<sp>
<speaker>*oi)di/pous</speaker>
<l n="105">e)/coid' a)kou/wn:  ou) ga\r ei)sei=do/n ge/ pw.</l></sp>
<sp>
<speaker>*kre/wn</speaker>
<l>tou/tou qano/ntos nu=n e)piste/llei safw=s</l>
<l>tou\s au)toe/ntas xeiri\ timwrei=n tinas.</l></sp>
```

7. Παράδειγμα XML αρχείου με Αρχαίο Ελληνικό κείμενο

Όπως παρατηρούμε, υπάρχουν σύνολα από XML tags που είτε περικλείουν την πληροφορία που θέλουμε (π.χ `<p>`) ή όχι και μας είναι άχρηστα (π.χ `<title>`). Εκμεταλλευόμενοι τα πρώτα, δημιουργήσαμε έναν XML parser που για κάθε κείμενο αντλούσε από τα χρήσιμα tags το κείμενο που περιέκλειαν. Έτσι για τα παραπάνω, θα είχαμε:

By banishing the man, or by paying back bloodshed with bloodshed, since it is this blood which brings the tempest on our city. Oedipus And who is the man whose fate he thus reveals? Creon Laius, my lord, was the leader of our land before you assumed control of this state. Oedipus I know it well—by hearsay, for I never saw him. Creon He was slain, and the god now bids us to take vengeance on his murderers, whoever they are.	a)ndrhlatou=ntas h)\ fo/nw| fo/non pa/lin lu/ontas, w(s to/d' ai(=ma xeima/zon po/lin. *oi)di/pous poi/ou ga\r a)ndro\s th/nde mhnu/ei tu/xhn; *kre/wn h)=n h(mi/n, w)=nac, *la/i+o/s poq' h(gemw\n gh=s th=sde, pri\n se\ th/nd' a)peuqu/nein po/lin. *oi)di/pous e)/coid' a)kou/wn: ou) ga\r ei)sei=do/n ge/ pw. *kre/wn tou/tou qano/ntos nu=n e)piste/llei safw=s tou\s au)toe/ntas xeiri\ timwrei=n tinas.

8. Κείμενο μετά την ανάλυση (parsing) των αρχικών αρχείων XML

Στη συνέχεια, έγινε καθαρισμός των κειμένων από περιττούς χαρακτήρες που μπορεί να είχαν παρεισφρήσει (π.χ HTML tags). Τα αρχαιοελληνικά κείμενα είναι γραμμένα σε Beta Code [11], η οποία είναι μέθοδος αποτύπωσης της στίξης και των γραμμάτων χρησιμοποιώντας μόνο χαρακτήρες ASCII. Έτσι, επόμενο βήμα ήταν η μετατροπή των χαρακτήρων του αρχαιοελληνικού κειμένου από beta code στους αντίστοιχους Unicode (UTF-8):

By banishing the man, or by paying back bloodshed with bloodshed, since it is this blood which brings the tempest on our city. Oedipus And who is the man whose fate he thus reveals? Creon Laius, my lord, was the leader of our land before you assumed control of this state. Oedipus I know it well—by hearsay, for I never saw him. Creon He was slain, and the god now bids us to take vengeance on his murderers, whoever they are.	ἀνδρηλατοῦντας ἢ φόνῳ φόνον πάλιν λύοντας, ὡς τόδ' αἷμα χειμάζον πόλιν. οἰδίπους ποίου γὰρ ἀνδρὸς τήνδε μηνύει τύχην; κρέων ἦν ἡμίν, ὦναξ, λάϊός ποθ' ἡγεμὼν γῆς τῆσδε, πρὶν σὲ τήνδ' ἀπευθύνειν πόλιν. οἰδίπους ἔξοιδ' ἀκούων· οὐ γὰρ εἰσεῖδόν γέ πω. κρέων τούτου θανόντος νῦν ἐπιστέλλει σαφῶς τοὺς αὐτοέντας χειρὶ τιμωρεῖν τινας.

9. Μετατροπή κειμένου από beta code σε UTF-8

Κατόπιν διαχωρίστηκαν οι λέξεις από τα σημεία στίξης (tokenization). Το βήμα αυτό είναι αρκετά βασικό αφού -αν παραλειφθεί- οι λέξεις "`the`" και "`the,`", για παράδειγμα, θα θεωρηθούν διαφορετικές. Αν γίνει κάτι τέτοιο η ποιότητα της μετάφρασης επηρεάζεται σημαντικά. Τελευταίο βήμα ήταν η ευθυγράμμιση (alignment) των γραμμών των κειμένων που έγινε με έτοιμο λογισμικό και περιγράφεται στη συνέχεια.

Με την ολοκλήρωση των βημάτων το μέγεθος της συλλογής μας ήταν:

Αριθμός παράλληλων γραμμών	140.415
Αριθμός Αγγλικών λέξεων	2.833.165
Αριθμός Αρχαίων Ελληνικών λέξεων	2.289.520

3.3 Λογισμικό

3.3.1 Ευθυγράμμιση κειμένων

Η ιδανική μορφή που θέλουμε να έχουν τα δύο αρχεία με τα κείμενά μας πριν το ξεκίνημα της εκπαίδευσης είναι οι προτάσεις που βρίσκονται στην ίδια γραμμή να αποτελούν η μία την ακριβή μετάφραση της άλλης. Έτσι, στο παρακάτω απόσπασμα:

νόσῳ δὲ τὸν βίον ὁ τέταρτος ὑπεξῆλθεν .
ὁ δὲ πέμπτος ὑπὸ ποιμένων ἐσφάγη , καὶ ὁ ἕκτος ὁμοίως σφαγῇ κατέστρεψε τὸν βίον .
ὁ δὲ ἕβδομος καὶ τῆς πόλεως καὶ τῆς βασιλείας παρανομῶν ἐξηλάθη ·
ἐξ οὗ τῆς βασιλείας καταλυθείσης εἰς τοὺς ὑπάτους τὰ τῆς ἀρχῆς μετετέθη .

the fourth died of a disease .
the fifth was murdered by some shepherds .
the sixth lost his life in a similar manner .
the seventh was expelled from the city and kingdom for violating the laws .
from that time kingly rule came to an end , and the administration of government was transferred to consuls .

10. Κείμενο χωρίς ευθυγράμμιση

μετά από κάποιου είδους επεξεργασία θα θέλαμε να καταλήξουμε να έχουμε:

νόσῳ δὲ τὸν βίον ὁ τέταρτος ὑπεξῆλθεν .
ὁ δὲ πέμπτος ὑπὸ ποιμένων ἐσφάγη , καὶ ὁ ἕκτος ὁμοίως σφαγῇ κατέστρεψε τὸν βίον .
ὁ δὲ ἕβδομος καὶ τῆς πόλεως καὶ τῆς βασιλείας παρανομῶν ἐξηλάθη ·
ἐξ οὗ τῆς βασιλείας καταλυθείσης εἰς τοὺς ὑπάτους τὰ τῆς ἀρχῆς μετετέθη .

the fourth died of a disease .
the fifth was murdered by some shepherds . the sixth lost his life in a similar manner .
the seventh was expelled from the city and kingdom for violating the laws .
from that time kingly rule came to an end , and the administration of government was transferred to consuls .

11. Κείμενο μετά την ευθυγράμμιση

Στο παραπάνω παράδειγμα βλέπουμε ότι όχι μόνο χρειάζεται να ενώσουμε τις δύο αγγλικές προτάσεις σε μία, αλλά πρέπει να μετακινήσουμε και τις επόμενες προτάσεις μία θέση πάνω.

Το πρόβλημα της ορθής ευθυγράμμισης κειμένων συναντάται πολύ συχνά στην κατασκευή παράλληλων σωμάτων και για το λόγο αυτό έχουν προταθεί διάφοροι αλγόριθμοι. Γενικά, υπάρχουν δύο προσεγγίσεις για την ευθυγράμμιση σε επίπεδο πρότασης, η μία βασίζεται στο μέγεθος των προτάσεων και η άλλη στις αντιστοιχίες μεταξύ λέξεων. Ειδικότερα, η πρώτη στηρίζεται στο γεγονός ότι οι μεγάλες προτάσεις τείνουν να παραμένουν μεγάλες, ή αντίστοιχα μικρές, και όταν μεταφραστούν [12]. Αυτού του είδους οι υλοποιήσεις μας δίνουν το πλεονέκτημα ότι είναι αρκετά γρήγορες αλλά όχι πάντα ακριβείς. Από την άλλη, οι μέθοδοι που βασίζονται στην αντιστοιχία λέξεων είναι τις περισσότερες φορές πιο ακριβείς, ωστόσο έχουν και μεγαλύτερο κόστος σε χρόνο. Επίσης, για να λειτουργήσουν πρέπει να υπάρχει διαθέσιμο κάποιο λεξικό ή σύνολο συγγενών λέξεων. Τέλος, υπάρχουν υλοποιήσεις που εκμεταλλεύονται και τις δύο μεθόδους για να επιτύχουν παραλληλοποίηση.

Αρχικά, χρησιμοποιήσαμε το hunalign [13] το οποίο αποτελεί υλοποίηση της πρώτης τεχνικής ευθυγράμμισης. Στη συνέχεια δοκιμάσαμε τον bilingual sentence aligner της Microsoft [14] που επωφελείται και των δύο μεθόδων. Ειδικότερα, στο πρώτο του βήμα ο αλγόριθμος λειτουργεί όπως οι οι μέθοδοι που βασίζονται στο μέγεθος της πρότασης. Στη συνέχεια, από τις προτάσεις που θεωρεί ότι είναι σωστά ευθυγραμμισμένες με μεγάλη πιθανότητα εξάγει κάποιες αντιστοιχίες λέξεων (λεξικό) με τη βοήθεια του πρώτου μοντέλου της IBM. Τέλος, αυτές τις ανατροφοδοτεί στο πρώτο βήμα του αλγορίθμου με σκοπό να πετύχει καλύτερη αντιστοίχηση. Τελικά, επειδή η δεύτερη υλοποίηση φάνηκε να μας δίνει ελαφρώς καλύτερα αποτελέσματα την προτιμήσαμε έναντι της πρώτης.

3.3.2 Εκπαίδευση συστήματος

Για να μπορέσουμε να μεταφράσουμε επιτυχώς κείμενα από μία γλώσσα σε μία άλλη χρησιμοποιώντας στατιστική μηχανική μετάφραση δύο είναι τα κύρια στοιχεία που πρέπει να υλοποιηθούν απαραίτητα. Το γλωσσικό και το μεταφραστικό μοντέλο που περιγράψαμε παραπάνω. Το πρώτο μας βοηθά να αναπτύσσουμε γραμματικά σωστές προτάσεις. Το μεταφραστικό μοντέλο από την άλλη, έχει ως σκοπό να μεταφράσει επιτυχημένα στη γλώσσα στόχο μας μία φράση/λέξη και το πετυχαίνει χρησιμοποιώντας έναν οι περισσότερους πίνακες φράσεων και αναδιατάξεων. Ο πίνακας φράσεων έχει αποθηκευμένη τη συχνότητα με την οποία κάποιο ζευγάρι ν-γραμμάτων της γλώσσας πηγής συνυπήρξε στην ίδια πρόταση με ένα ζευγάρι ν-γραμμάτων της γλώσσας στόχου. Η συχνότητα αυτή μας δίνει πληροφορία για το πόσο πιθανό είναι αυτό το ζευγάρι μετάφρασης να το ξανασυναντήσουμε σε κάποιο άλλο παρόμοιο κείμενο. Ο πίνακας αναδιάταξης περιέχει συχνότητες που περιγράφουν τις αλλαγές που συμβαίνουν στη σειρά των λέξεων από τη γλώσσα πηγής στη γλώσσα στόχο. Για παράδειγμα, πόσο πιθανότερο είναι μία φράση να έχει μεταφραστεί σε "nice view" αντί "view nice". Στη συνέχεια, συνδυάζοντας τα παραπάνω μπορούμε να ψάξουμε μέσα στο διαθέσιμο χώρο

μεταφράσεων για την καλύτερη πιθανοτικά μετάφραση.

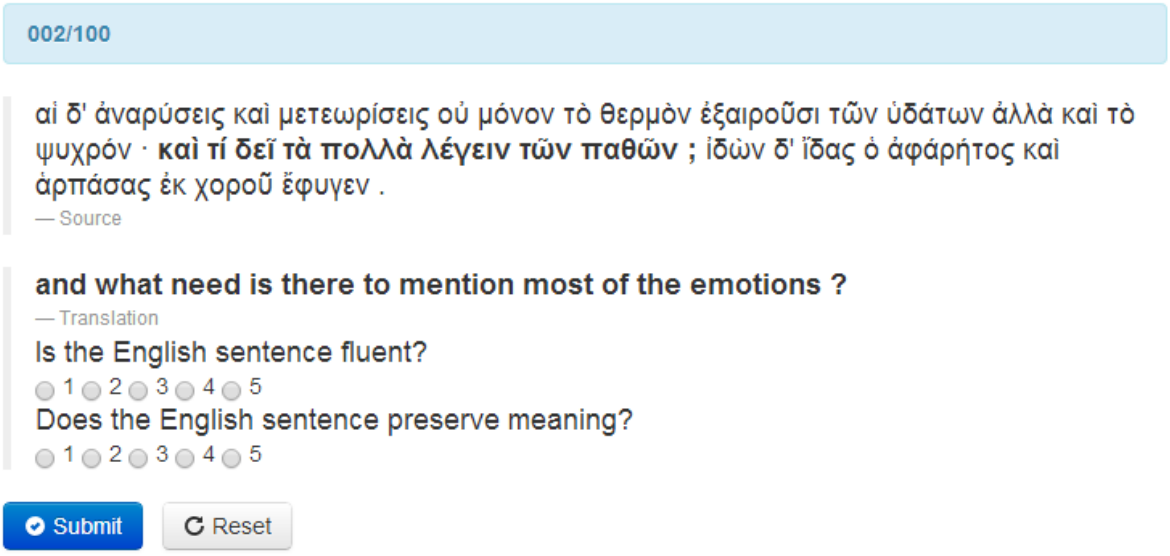
Η διαδικασία που περιγράφηκε μπορεί να αυτοματοποιηθεί με τη βοήθεια του Moses [15]. Το Moses, είναι ένα ολοκληρωμένο σύστημα στατιστικής μηχανικής μετάφρασης που παίρνει ως είσοδο ένα ζευγάρι παράλληλων σωμάτων, γραμμένα σε οποιαδήποτε γλώσσα θέλουμε, και υλοποιεί όλα τα βήματα που είναι απαραίτητα για την εκπαίδευση ενός στατιστικού μεταφραστικού συστήματος.

3.3.3 Αξιολόγηση συστήματος

Μέχρι τώρα έχουμε περιγράψει συνοπτικά πώς μπορούμε να δημιουργήσουμε το δικό μας αυτόματο μεταφραστικό σύστημα αλλά, δεν έχουμε αναφέρει πώς θα διακρίνουμε αν οι μεταφράσεις μας είναι ικανοποιητικές. Οι τρόποι είναι δύο, είτε με κάποιο αλγόριθμο αυτόματης αξιολόγησης είτε χρησιμοποιώντας ανθρώπους αξιολογητές. Ο δεύτερος τρόπος είναι πιο χρονοβόρος αλλά μας δίνει καλύτερη εικόνα της ποιότητας των μεταφράσεων

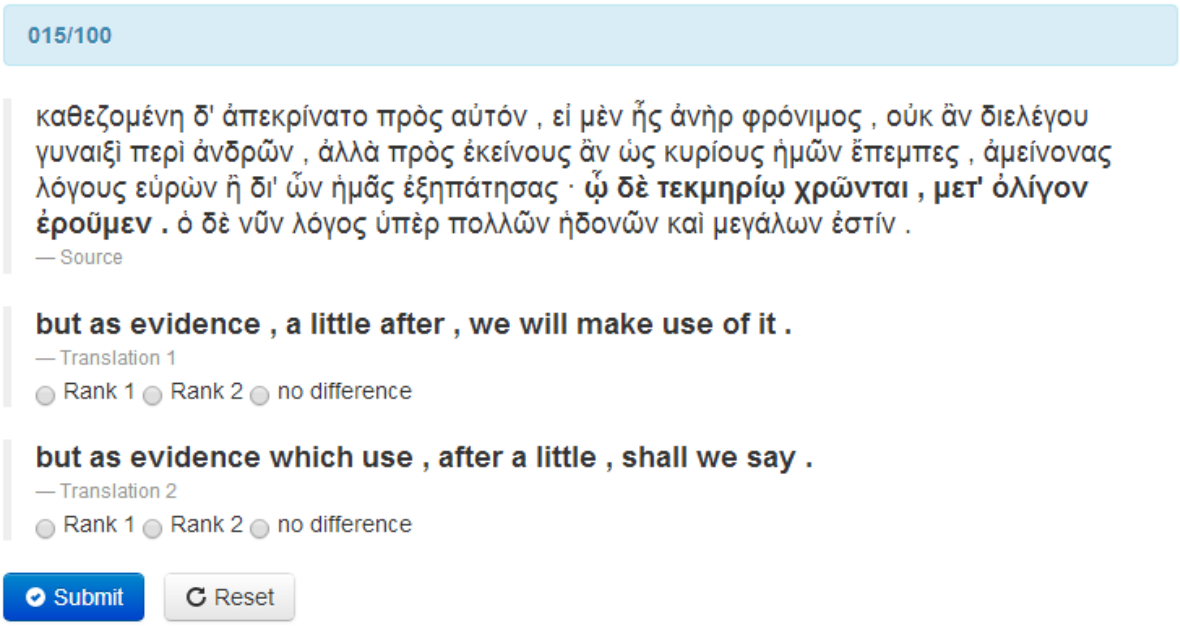
Ως αλγόριθμος αξιολόγησης χρησιμοποιήθηκε η μετρική BLEU [9]. Η κεντρική ιδέα πίσω από το BLEU είναι ότι όσο πιο πολλά κοινά έχει μία μετάφραση που έχει δημιουργηθεί από ένα αυτόματο σύστημα με αυτή που έχει γίνει από επαγγελματία άνθρωπο μεταφραστή τόσο καλύτερη θα είναι. Γι' αυτό το λόγο όταν αξιολογούμε κάποιες προτάσεις πρέπει να τις έχουμε μεταφρασμένες σωστά από πριν. Τελικά το αποτέλεσμα που παίρνουμε είναι ένα αριθμός από 0 έως 1. Η τιμή αυτή δείχνει πόσο μοιάζουν το μεταφρασμένο από το σύστημα με το κείμενο αναφοράς (του ανθρώπου μεταφραστή), με τις τιμές πιο κοντά στο 1 να υποδηλώνουν μεγαλύτερη ομοιότητα.

Η αξιολόγηση από ανθρώπους έγινε με τη βοήθεια του Appraise [16] το οποίο είναι ένα εργαλείο ανοιχτού λογισμικού για χειροκίνητη αξιολόγηση μεταφράσεων μηχανικής μετάφρασης. Συγκεκριμένα, οι εργασίες αξιολόγησης χωρίστηκαν σε εργασίες εκτίμησης ποιότητας και εκτίμησης κατάταξης. Στην πρώτη περίπτωση δινόταν μία μεταφρασμένη πρόταση στα Αγγλικά και στόχος ήταν να βαθμολογηθεί στην κλίμακα 1(κακή) έως 5(άριστη) ως προς τη διατήρηση του νοήματος και τη σωστή χρήση του λόγου, λαμβάνοντας υπόψη και την αρχική Αρχαία Ελληνική πρόταση.



12. Αξιολόγηση πρότασης με το Appraise

Κατά την εκτίμηση κατάταξης παρείχαμε δύο μεταφράσεις της ίδιας πρότασης από διαφορετικά συστήματα. Σκοπός ήταν να υποδειχθεί ποια είναι η καλύτερη από τις δύο.



13. Σύγκριση προτάσεων στο Appraise

Κεφάλαιο 4: Πειράματα

Αφού τελειώσαμε με την κατασκευή των παράλληλων κειμένων προχωρήσαμε στο επόμενο βήμα, που ήταν και ο αρχικός μας στόχος, την υλοποίηση ενός πειραματικού μεταφραστικού συστήματος. Στο κεφάλαιο αυτό παρουσιάζουμε τα πειράματα που έγιναν μαζί με τα αποτελέσματά τους.

4.1 Περιγραφή πειραμάτων

Όπως αναφέραμε και στο προηγούμενο κεφάλαιο τα κείμενά μας προέρχονται από μια πληθώρα συγγραφέων. Αυτό έχει σαν αποτέλεσμα το εύρος των περιεχομένων τους να εκτείνεται από κομμάτια της Ευκλείδειας γεωμετρίας μέχρι Ομηρικά αποσπάσματα. Επομένως, μπορεί να γίνει αντιληπτό ότι η επίδοση του συστήματός μας θα μειωθεί αν έχουμε πολλά δεδομένα εκπαίδευσης που προέρχονται από διαφορετική θεματική περιοχή από τα δεδομένα αξιολόγησης [17]. Ιδανικά, θα μπορούσαμε να δημιουργήσουμε ένα σύστημα για κάθε συγγραφέα ή θεματική περιοχή. Ωστόσο, κάτι τέτοιο θα ήταν άσκοπο αφού τα δεδομένα που έχουμε στη διάθεση μας για κάθε διαφορετικό πεδίο είναι πολύ λίγα και δεν είναι ικανά να εκπαιδεύσουν ικανοποιητικά ένα σύστημα στατιστικής μηχανικής μετάφρασης.

Ο περιορισμένος αριθμός παράλληλων κειμένων για διάφορες θεματικές περιοχές αποτελεί πρόβλημα στην στατιστική μηχανική μετάφραση. Αυτή η έλλειψη δεδομένων (data sparseness), μπορεί να αντιμετωπιστεί, εν μέρει, αν προσθέσουμε κείμενα άλλων περιοχών στη συλλογή μας. Όμως, η μετάφραση είναι εκ φύσεως διφορούμενη εξαιτίας, κυρίως, της χρησιμοποίησης των ίδιων λέξεων σε διαφορετικές θεματικές περιοχές. Επομένως, με την προσθήκη «ξένων» κειμένων μπορεί να επηρεάσουμε αρνητικά το σύστημά μας. Για το λόγο αυτό, έχουν αναπτυχθεί μέθοδοι συνδυασμού διαφορετικών γλωσσικών ή και μεταφραστικών μοντέλων που αντιμετωπίζουν το πρόβλημα. Ένας τέτοιος τρόπος είναι η γραμμική παρεμβολή μοντέλων [18]. Δηλαδή, δημιουργούμε ένα μοντέλο για κάθε διαφορετικό είδος από κείμενα και στο τέλος παίρνουμε το σταθμισμένο μέσο τους. Ο συνδυασμός των βαρών που θα χρησιμοποιηθεί θα πρέπει να είναι τέτοιος ώστε να ελαχιστοποιείται η περιπλοκή του μοντέλου μας με βάση κάποιο σύνολο εναρμόνισης.

Τα πειράματά μας μπορούν να χωριστούν σε δύο κατηγορίες. Στόχος της πρώτης ήταν να δημιουργηθεί ένα γενικό σύστημα μετάφρασης. Της δεύτερης, να αναπτυχθεί ένα σύστημα που θα ειδικεύεται στην μετάφραση κειμένων ενός συγκεκριμένου συγγραφέα.

Για την πρώτη κατηγορία χρησιμοποιήθηκαν όλα τα κείμενα για την εκπαίδευση. Εξαίρεση αποτελούν οι γραμμές που αφαιρέθηκαν για τα σύνολα εναρμόνισης (tuning) και δοκιμών (testing). Αυτές, πάρθηκαν τυχαία από το σύνολο των κειμένων.

Για να κατασκευαστεί ένα σύστημα που θα ειδικεύεται σε κάποιον συγκεκριμένο τομέα πρέπει

να το μάθουμε να δίνει περισσότερη βαρύτητα στα δεδομένα εκπαίδευσης που προέρχονται από αυτόν. Επομένως, σύμφωνα και με όσα περιγράψαμε νωρίτερα, για τα μεταφραστικά μοντέλα η διαδικασία που ακολουθήσαμε έχει ως εξής. Αρχικά, εκπαίδευση δύο διαφορετικών μεταφραστικών μοντέλων. Το πρώτο με όλα τα κείμενα εκπαίδευσης και το δεύτερο μόνο με κείμενα γραμμένα από έναν συγκεκριμένο συγγραφέα. Στη συνέχεια, παρεμβολή των δύο μοντέλων με βάρη τα οποία ελαχιστοποίησαν την περιπλοκή σε ένα σύνολο προτάσεων αναφοράς, στην περίπτωσή μας στις προτάσεις του συνόλου εναρμόνισης (tuning set). Την παραπάνω διαδικασία την υλοποιήσαμε με εργαλεία του Moses ενώ παρόμοια είναι τα βήματα και στη περίπτωση της παρεμβολής γλωσσικών μοντέλων.

Τέλικα, για τα πειράματα της δεύτερη κατηγορίας ακολουθήσαμε τα παρακάτω επιμέρους βήματα. Αρχικά, χρησιμοποιήσαμε με την ίδια βαρύτητα όλα τα δεδομένα εκπαίδευσης. Με αυτό τον τρόπο δεν εκμεταλλευόμασταν καθόλου τη γνώση που είχαμε για το συγγραφέα των δεδομένων αξιολόγησης. Στη συνέχεια, παρεμβάλλαμε γραμμικά τα γλωσσικά και μεταφραστικά μοντέλα με τα αντίστοιχα που είχαν δημιουργηθεί μόνο από κείμενα του συγγραφέα των δεδομένων αξιολόγησής μας. Έτσι, πετύχαμε το εκάστοτε σύστημα να δίνει περισσότερη βαρύτητα στην πληροφορία που λαμβάνει από αυτά και άρα, πιθανότατα καλύτερα αποτελέσματα στη μετρική μας [19].

Όσον αφορά τα εξειδικευμένα συστήματά μας, το ένα θέλαμε να ειδικεύεται στην μετάφραση του Ξενοφώντα. Γι' αυτό, εκτός της αρχικής δημιουργήσαμε και μια μικρότερη συλλογή κειμένων που περιείχε μόνο τα κείμενα του Ξενοφώντα. Για το άλλο, επιλέξαμε τον Πλούταρχο. Επίσης, στα δεδομένα αξιολόγησης είχαμε διαθέσιμες δύο διαφορετικές μεταφράσεις για κάθε πρόταση αρχαίου κειμένου. Έτσι, περιμέναμε ότι θα βελτιωνόταν η αξιολόγηση της ποιότητας των μεταφράσεων (BLEU score).

4.2 Αξιολόγηση

4.2.1 Αξιολόγηση BLEU

Από την εκτέλεση των παραπάνω πειραμάτων λάβαμε τα εξής:

Γενικό σύστημα μετάφρασης:

Translation Model	Language Model	BLEU
genres	genres	
all	all	0.1464

17. Γενικό σύστημα μετάφρασης

Σύστημα εξειδικευμένο σε κείμενα του Ξενοφώντα:

Translation Model		Language Model		BLEU
genres	Interpolated TM	genres	Interpolated LM	
all	-	all	-	0.1079
all	-	all	yes	0.1087
Xenofontas	-	Xenofontas	-	0.0881
all	yes	all	yes	0.1137

18. Συστήματα αξιολογημένα σε κείμενα του Ξενοφώντα

Σύστημα εξειδικευμένο σε κείμενα του Πλουτάρχου:

Translation Model		Language Model		BLEU
genres	Interpolated TM	genres	Interpolated LM	
all	-	all	-	0.1256
all	-	all	yes	0.1263
Ploutarxos	-	Ploutarxos	-	0.0925
all	yes	all	yes	0.1239

19. Συστήματα αξιολογημένα σε κείμενα του Πλουτάρχου

Παρατηρούμε ότι η χρησιμοποίηση παρεμβολής στα μοντέλα μας όντως βοήθησε στη βελτίωση του BLEU, περισσότερο στην περίπτωση των κειμένων του Ξενοφώντα, αλλά και στην περίπτωση του Πλουτάρχου (όπου την καλύτερη επίδοση είχε το σύστημα που χρησιμοποιούσε

παρεμβολή στο γλωσσικό μοντέλο -- δεύτερη γραμμή του πίνακα 19)

Είναι γεγονός, όμως, ότι όλα τα αποτελέσματα είναι ιδιαίτερα χαμηλά. Αυτό μπορεί να δικαιολογηθεί, εν μέρει, από το μικρό μέγεθος των δεδομένων εκπαίδευσης και την πληθώρα διαφορετικών συγγραφέων που το απάρτιζαν. Ως αποτέλεσμα, στις μεταφράσεις που παράχθηκαν υπήρχε σημαντικό πλήθος φράσεων που έμεναν αμετάφραστες. Αυτό μπορεί να συνέβη είτε γιατί δεν είχαν εμφανιστεί στα δεδομένα εκπαίδευσης είτε επειδή υπήρχαν ελάχιστες φορές και δεν κατέστη δυνατό να δημιουργηθεί κάποια αντιστοιχία με σχετικά υψηλή πιθανότητα. Επίσης, παρατηρήσαμε κατά την κατασκευή των συνόλων εναρμόνισης και αξιολόγησης, ότι η ευθυγράμμιση των κειμένων είχε αρκετά λάθη. Πολλές προτάσεις είτε είχαν εντελώς λάθος αντιστοίχηση είτε έλειπε σημαντικό μέρος της μετάφρασης και γι' αυτό τις διαγράφαμε. Κάτι τέτοιο όμως ήταν αδύνατο να γίνει και για το πολύ μεγαλύτερο σύνολο εκπαίδευσης. Αυτό ήταν και το σημαντικότερο πρόβλημα αφού αν η αρχική ευθυγράμμιση είναι λάθος επηρεάζεται αρνητικά η εκπαίδευση και άρα το τελικό σύστημα.

4.2.2 Σύγκριση με άλλα συστήματα

Από τα αποτελέσματα των πειραμάτων μας και μόνο, δεν ήταν δυνατό να είμαστε σίγουροι αν η ποιότητα των μεταφράσεων θα μπορούσε να βελτιωθεί αισθητά ακόμη, ή αν με δεδομένο το μέγεθος και το περιεχόμενο της συλλογής μας είχαμε αγγίξει κάποιο άνω φράγμα της μεθόδου μας. Επίσης, δεν είχε κατασκευαστεί μέχρι τώρα κάποιο άλλο, αντίστοιχο σώμα κειμένων, ούτε αυτόματο σύστημα μετάφρασης για να συγκρίνουμε τα αποτελέσματά μας.

Γι' αυτό, σκεφτήκαμε να υλοποιήσουμε ένα σύστημα αυτόματης μετάφρασης από τα Νέα Ελληνικά στα Αγγλικά, χρησιμοποιώντας σύνολα δεδομένων περίπου ίσου μεγέθους με τα σύνολα δεδομένων των προηγούμενων πειραμάτων μας, και να συγκρίνουμε τα αποτελέσματα BLEU του νέου συστήματος με αυτά των προηγούμενων πειραμάτων μας.

Δημιουργήσαμε ένα καινούργιο σώμα παράλληλων κειμένων, το οποίο περιείχε κείμενα από τις ακόλουθες συλλογές:

SETIMES [20]: παράλληλη συλλογή κειμένων αποτελούμενη από ειδησεογραφικά άρθρα της ομώνυμης σελίδας,

OpenSubtitles [21]: συλλογή από υπότιτλους ταινιών του οργανισμού opensubtitles,

Europarl [2]: σύνολο παράλληλων κειμένων από τα πρακτικά του Ευρωπαϊκού Κοινοβουλίου.

με μέγεθος:

Αριθμός παράλληλων γραμμών	134.176
Αριθμός Αγγλικών λέξεων	2.540.981
Αριθμός Ελληνικών λέξεων	2.564.442

Ακολουθώντας τον ίδιο συλλογισμό με τη δική μας συλλογή, διενεργήσαμε δύο πειράματα. Στο ένα τα δεδομένα αξιολόγησης είχαν παρθεί από το σύνολο των κειμένων, στο άλλο, μόνο από τη συλλογή του Europarl.

Τα αποτελέσματά μας ήταν τα εξής:

Γενικό μεταφραστικό σύστημα:

Translation Model	Language Model	BLEU
genres	genres	
all	all	0.3046

20. Γενικό μεταφραστικό σύστημα από Νέα Ελληνικά στα Αγγλικά

Εξειδικευμένο μεταφραστικό σύστημα:

Translation Model		Language Model		BLEU
genres	Interpolated TM	genres	Interpolated LM	
all	-	all	-	0.2455
all	-	all	yes	0.2498
Europarl	-	Europarl	-	0.2438
all	yes	all	yes	0.2534

21. Συστήματα μετάφρασης από τα Νέα Ελληνικά στα Αγγλικά αξιολογημένα σε κείμενα του Europarl

Αν και για τα δύο συστήματα αυτόματης μετάφρασης (Αρχαία Ελληνικά προς Αγγλικά, Νέα Ελληνικά προς Αγγλικά) χρησιμοποιήθηκε περίπου το ίδιο μέγεθος δεδομένων εκπαίδευσης, η υπεροχή (των παραλλαγών) του συστήματος μετάφρασης Νέων Ελληνικών είναι ξεκάθαρη, αν συγκρίνουμε με τα αντίστοιχα αποτελέσματα του συστήματος μετάφρασης Αρχαίων Ελληνικών. Φυσικά, δεν μπορούμε να κάνουμε άμεση σύγκριση των αποτελεσμάτων, αφού τα είδη των κειμένων που δόθηκαν στα δύο συστήματα είναι πολύ διαφορετικά, αλλά θα μπορούσαμε να ισχυριστούμε ότι το κύριο πρόβλημα του συστήματος μετάφρασης των Αρχαίων Ελληνικών δεν είναι τόσο το μέγεθος του σώματος εκπαίδευσης, όσο το είδος των κειμένων που περιέχει (π.χ. φιλοσοφία, μαθηματικά, ιστορία), ενδεχομένως και η ποιότητα των δεδομένων εκπαίδευσης (π.χ. λάθη ευθυγράμμισης προτάσεων).

4.2.3 Ανθρώπινη αξιολόγηση

Από τα προηγούμενα αποτελέσματα έγινε φανερό ότι το σύστημά μας δε ήταν ακόμη σε θέση να μεταφράζει με μεγάλο ποσοστό επιτυχίας, έτσι η ανθρώπινη αξιολόγηση δε ήταν απαραίτητο να πραγματοποιηθεί. Ωστόσο, επειδή είχαμε ήδη υλοποιήσει την πλατφόρμα ανθρώπινης αξιολόγησης και μας ενδιέφερε να γίνει έλεγχος στην ποιότητα και την ευχρηστία της, ζητήσαμε από έναν ειδικό να την χρησιμοποιήσει για να αξιολογήσει παραγόμενες αγγλικές μεταφράσεις αρχαιοελληνικών κειμένων. Η αξιολόγηση έγινε στα κείμενα του Πλουτάρχου και μόνον για δύο από τις παραλλαγές του συστήματος μετάφρασης από τα Αρχαία Ελληνικά στα Αγγλικά. Πιο συγκεκριμένα, για την εκτίμηση ποιότητας ξεχωρίσαμε τυχαία πενήντα προτάσεις από το σύνολο μεταφράσεων που παρήγαγε το καλύτερο σύστημά μας. Για την εκτίμηση κατάταξης πήραμε άλλες 50, διαφορετικές από τις προηγούμενες προτάσεις, από τα δύο συστήματα που είχαν παρεμβολή στα μοντέλα τους.

Από την εκτίμηση ποιότητας των μεταφράσεων είχαμε την εξής αξιολόγηση:

Quality Ranking			
number of sentences	percentage	fluency	meaning preservation
38	76%	1	1
1	2%	1	2
3	3%	2	1
4	8%	2	2
1	2%	2	3
2	4%	3	3
1	2%	5	5

22. Αποτελέσματα ανθρώπινης αξιολόγησης

Από την εκτίμηση κατάταξης των δύο μεταφραστικών μοντέλων, 26 προτάσεις (52%) είχαν σχεδόν πανομοιότυπη μετάφραση από τα δύο συστήματα. Σε 16 προτάσεις (32%), το σύστημα στο οποίο είχαμε παρεμβάλει τα δύο γλωσσικά μας μοντέλα ήταν καλύτερο, ενώ σε 8 (16%) ήταν καλύτερο το σύστημα που είχε παρεμβολή και στα μεταφραστικά μοντέλα. Τα αποτελέσματα της αξιολόγησης δεν ήταν ιδιαίτερα ενθαρρυντικά για το σύστημά μας, κάτι το οποίο περιμέναμε βάσει και των αποτελεσμάτων BLEU.

Εκτός από την αξιολόγηση, όμως, λάβαμε σημαντική ανατροφοδότηση σχετικά με τις αδυναμίες του συστήματός μας και κάποιους από τους λόγους που οδήγησαν στη χαμηλή βαθμολογία στις αξιολογήσεις. Αρχικά, μας επισημάνθηκε ότι τυχαίες προτάσεις χωρίς τα συμφραζόμενά τους, όπως αυτές που χρησιμοποιήσαμε για την εκπαίδευση και αξιολόγηση των συστημάτων, είναι δύσκολο να μεταφραστούν (τουλάχιστον από τους ανθρώπους) αλλά και να αξιολογηθούν, αφού τα συμφραζόμενα παίζουν σημαντικό ρόλο στην κατανόησή τους.

Ακόμη, η αγγλική μετάφραση συχνά δεν λάμβανε υπόψη βασικούς συντακτικούς κανόνες της Αρχαίας Ελληνικής, με αποτέλεσμα να επηρεάζεται σε μεγάλο βαθμό η μεταφραστική συνοχή. Αυτό το πρόβλημα μπορούσε ενδεχομένως να αντιμετωπιστεί αν είχαμε περισσότερα δεδομένα εκπαίδευσης αλλά και χρησιμοποιώντας ένα σύστημα μετάφρασης βασισμένο και στη σύνταξη [22, 23], αντί για σύστημα βασισμένο μόνο σε φράσεις.

Κεφάλαιο 5: Συμπεράσματα

5.1 Ανασκόπηση

Στόχος της εργασίας ήταν η κατασκευή ενός αυτόματου συστήματος μετάφρασης από τα Αρχαία Ελληνικά στα Αγγλικά. Έτσι, αρχικά δημιουργήσαμε ένα παράλληλο σώμα κειμένων χρησιμοποιώντας της ψηφιακή βιβλιοθήκη Perseus [4]. Στη συνέχεια, με τη βοήθεια του Moses [15], ενός ολοκληρωμένου συστήματος στατιστικής μηχανικής μετάφρασης, εκπαιδεύσαμε και αξιολογήσαμε τρεις παραλλαγές συστημάτων μετάφρασης βασισμένες στα κείμενα που είχαμε αντλήσει νωρίτερα. Αυτή ήταν η πρώτη μας απόπειρα δημιουργίας συστήματος μετάφρασης μεταξύ των δύο γλωσσών αλλά, από όσο γνωρίζουμε, και η πρώτη γενικώς αφού στη βιβλιογραφία δεν υπάρχει κάποια αντίστοιχη εργασία. Οι χαμηλές επιδόσεις του συστήματος οφείλονται μάλλον στο μικρό μέγεθος του σώματος εκπαίδευσης, στα σφάλματα ευθυγράμμισης των προτάσεων και στην πληθώρα και δυσκολία των θεμάτων των κειμένων εκπαίδευσης.

5.2 Μελλοντικές βελτιώσεις

Υπάρχουν αρκετά περιθώρια βελτίωσης του συστήματος μετάφρασης των Αρχαίων Ελληνικών. Αρχικά, θα πρέπει να προστεθούν περισσότερα κείμενα εκπαίδευσης. Θα ήταν ενδιαφέρον επίσης να πειραματιστεί κανείς με τη μετάφραση από τα Αρχαία Ελληνικά στα Νέα Ελληνικά, αν εξασφαλίσει κατάλληλο παράλληλο σώμα κειμένων, αξιοποιώντας π.χ. κοινές λέξεις ή ρίζες λέξεων κατά την ευθυγράμμιση προτάσεων. Θα μπορούσε επίσης να χρησιμοποιηθεί ένα ιεραρχικό μοντέλο στατιστικής μηχανικής μετάφρασης, που να ενσωματώνει περισσότερες πληροφορίες για το συντακτικό των γλωσσών. Ακόμη, θα ήταν ενδιαφέρον να γίνει χωρισμός των κειμένων βάσει θεματικών ενοτήτων (δράμα, κωμωδία, κλπ.) ή χρονικών περιόδων και να μελετηθεί το πώς επηρεάζεται η ποιότητα των μεταφράσεων. Τέλος, παρουσιάσαμε μία απλή μέθοδο ανθρώπινης αξιολόγησης που θα μπορούσε όμως να βελτιωθεί αρκετά. Θα μπορούσε να προστεθεί λειτουργικότητα όπως π.χ. να μπορεί να γίνει κατηγοριοποίηση των λαθών από τον αξιολογητή (συντακτικά λάθη, νοηματικά,κλπ.) και να δίνεται η δυνατότητα χειροκίνητης διόρθωσης των λανθασμένων μεταφράσεων. Επίσης, θα πρέπει να αξιολογηθούν περισσότερες προτάσεις και από περισσότερους αξιολογητές για να ελεγχθεί και το ποσοστό της μεταξύ τους συμφωνίας.

Αναφορές

[1] Koehn, Philipp. (2010). *Statistical Machine Translation* (1st ed.). New York, NY, USA: Cambridge University Press.

[2] Koehn, Philipp. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *Conference Proceedings: the tenth Machine Translation Summit* (p./pp. 79--86), Phuket, Thailand: AAMT.

[3] Roukos, Salim, David Graff, and Dan Melamed. Hansard French/English LDC95T20. Web Download. *Philadelphia: Linguistic Data Consortium*, 1995.

[4] Perseus Digital Library. Ed. Gregory R. Crane. Tufts University. http://www.perseus.tufts.edu (accessed March, 2013).

[5] Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. 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* (NAACL '03), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 48-54.

[6] Jurafsky, Dan, and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Pearson Prentice Hall, 2009. Print.

[7] Chen, Stanley F. and Goodman, Joshua. *An empirical study of smoothing techniques for language modeling*. Harvard University, Center for Research in Computing Technology, TR-10-98:1–63, 1998

[8] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.* 19, 2 (June 1993), 263-311.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (ACL '02). Association for Computational Linguistics, Stroudsburg, PA, USA, 311-318

[10] Kevin Knight. 1999. Decoding complexity in word-replacement translation models. *Comput. Linguist.* 25, 4 (December 1999), 607-615.

[11] Wikipedia contributors. "Beta Code." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 10 Feb. 2014. Web. 9 Jul. 2014.

[12] William A. Gale and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics* (ACL '91). Association for Computational Linguistics, Stroudsburg, PA, USA, 177-184.

[13] Daniel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, Viktor Nagy (2005).Parallel corpora for medium density languages *In Proceedings of the RANLP 2005*, pages 590-596.

[14] Robert C. Moore. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users* (AMTA '02), Stephen D. Richardson (Ed.). Springer-Verlag, London, UK, UK, 135-144.

[15] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (ACL '07). Association for Computational Linguistics, Stroudsburg, PA, USA, 177-180.

[16] Christian Federmann Appraise: An Open-Source Toolkit for Manual Evaluation of Machine Translation Output In *The Prague Bulletin of Mathematical Linguistics volume 98*, Prague, Czech Republic, 9/2012

[17] Barry Haddow and Philipp Koehn. 2012. Analysing the effect of out-of-domain data on SMT systems. In *Proceedings of the Seventh Workshop on Statistical Machine Translation* (WMT '12). Association for Computational Linguistics, Stroudsburg, PA, USA, 422-432.

[18] Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (EACL '12). Association for Computational Linguistics, Stroudsburg, PA, USA, 539-549.

[19] Holger Schwenk and Philipp Koehn. (2008). Large and Diverse Language Models for Statistical Machine Translation. In *IJCNLP*'08 (pp. 661–666).

[20] Tiedemann, Jörg. (2009). News from {OPUS} - {A} Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, & R. Mitkov (Eds.), *Recent Advances in Natural Language Processing* (Vol. V, pp. 237–248). Borovets, Bulgaria: John Benjamins, Amsterdam/Philadelphia.

[21] Jörg Tiedemann, 2012, Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC 2012)

[22] David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (ACL '05). Association for Computational Linguistics, Stroudsburg, PA, USA, 263-270.

[23] Wikipedia contributors. "Context-free grammar." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 8 May. 2014. Web. 9 Jul. 2014.