ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΠΙΣΤΗΜΗ ΤΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Διπλωματική Εργασία
Μεταπτυχιακού Διπλώματος Ειδίκευσης

«*Αυτόματη Εξαγωγή Δίγλωσσων Λεξικών από Παράλληλα Σώματα Κειμένων*»

Ιωάννης Χάρλας

Επιβλέπων: **Ι. Ανδρουτσόπουλος**
Επόπτης: **Σ. Πιπερίδης (Ινστιτούτο Επεξεργασίας του Λόγου)**

ΑΘΗΝΑ, ΙΟΥΝΙΟΣ 2005

Πίνακας Περιεχομένων

Περίληψη 5

1. Εισαγωγή 7

1.1. Περιγραφή Προβλήματος 7
1.1.1. Συντακτικές Διαφορές 7
1.1.2. Μορφολογικές Διαφορές 8
1.1.3. Γλωσσική Οικογένεια 8
1.2. Αντικειμενικός Σκοπός του Συστήματος 9
1.3. Δομή Εγγράφου 9

2. Επισκόπηση (*υπάρχουσες τεχνολογίες*) 10

2.1. Στοίχιση Κειμένων 10
2.1.1. Αλγόριθμος Gale & Church 11
2.1.2. Άλλοι Αλγόριθμοι στοίχισης προτάσεων 11
2.2. Ομαδοποίηση Λέξεων 12
2.3. Αντιστοίχηση Λέξεων 13
2.3.1. Εντοπισμός Υποψήφιων Μεταφράσεων 14
2.3.2. Βαθμολογία και Επιλογή Αντιστοιχίας Λέξεων 15
2.3.2.1. Αντιστοίχηση Λέξεων με «Συσχέτιση» 15
2.3.2.1.1. Μέτρηση Συνεμφανίσεων 15
2.3.2.1.2. Ταίριασμα Συμβολοσειρών (String Matching) 16
2.3.2.2. Αντιστοίχηση Λέξεων με «Εκτίμηση». 17
2.3.2.2.1. Μοντέλα Πιθανοτήτων 17
2.3.2.2.2. Μοντέλα Γράφων 17
2.3.2.3. Σύνθετες Μέθοδοι Αντιστοίχησης 18
2.4. Εντοπισμός και Αντιστοίχηση Πολυλεκτικών Όρων 18

3. Αρχιτεκτονική Συστήματος 20

3.1. Προ-επεξεργασία 20

3.1.1. Εντοπισμός Προτάσεων 20
3.1.2. Στοίχιση Προτάσεων 21
3.2. Φάση 1η 21
3.2.1. Εντοπισμός Λεκτικών Μονάδων 21
3.2.2. Εντοπισμός Υποψήφιων Μεταφράσεων 22
3.2.2.1. Μαθηματικός Τύπος Αξιολόγησης 22
3.2.2.2. Όριο Αποκοπής 23
3.2.2.3. Σύγκριση Μαθ. τύπου M(x,y) με Dice coefficient 25
3.2.3. Ομαδοποίηση Λέξεων (*Word Conflation*) 25
3.2.3.1. Ομαδοποίηση ***χωρίς*** Γλωσσική πληροφορία 26
3.2.3.2. Ομαδοποίηση ***με*** Γλωσσική πληροφορία 27
3.3. Φάση 2η : Αντιστοίχηση (Μετάφραση) Λέξεων 28
3.3.1. Βαθμολογία κατά Λέξη 28
3.3.2. Βαθμολογία κατά Αντιπρόσωπο 28
3.3.3. Εντοπισμός Αντίστοιχης Λέξεως 29
3.4. Φάση 3η : Αντιστοίχηση Πολυλεκτικών Όρων 31
3.4.1. Εντοπισμός Διλεκτικών όρων 31
3.4.2. Εντοπισμός Πολυλεκτικών όρων 32
3.4.3. Μετάφραση Πολυλεκτικών όρων 33

4. Αξιολόγηση 34

4.1. Δεδομένα Αξιολόγησης. 34
4.1.1. «Παράλληλα» Κείμενα 34
4.1.2. Μορφοποίηση πριν την επεξεργασία 35
4.2. Μέθοδος Αξιολόγησης. 35
4.3. Αξιολόγηση Στοίχισης Προτάσεων. 36
4.4. Αξιολόγηση Ομαδοποίησης Λέξεων. 36
4.4.1. Ομαδοποίηση χωρίς Γλωσσική πληροφορία 36
4.4.2. Ομαδοποίηση ΜΕ Γλωσσική πληροφορία 37
4.5. Αξιολόγηση Αντιστοίχησης Λέξεων 38
4.6. Αξιολόγηση Αντιστοίχησης Πολύ-λεκτικών Όρων 39

5. ***Συμπεράσματα – Μελλοντικό έργο*** 40

5.1. Συμπεράσματα 40

5.2. Μελλοντικές Κατευθύνσεις 40

Αναφορές 41

Παραρτήματα

Α. Εντοπισμένοι Πολυλεκτικοί Όροι 43

Περίληψη

Η εργασία που παρουσιάζουμε αποτελεί μια προσπάθεια υλοποίησης ενός συστήματος αυτοματοποιημένης εξαγωγής δίγλωσσων λεξικών, του οποίου την αποτελεσματικότητα δοκιμάζουμε χρησιμοποιώντας παράλληλα σώματα κειμένων σε Ελληνικά – Τουρκικά. Η δυσκολία στο επιλεγμένο ζεύγος γλωσσών έγκειται στο ότι δεν υπάρχουν κατάλληλα γλωσσικά εργαλεία για να υποστηρίξουν μια τέτοια προσπάθεια, αλλά και κυρίως στο ότι οι δύο αυτές γλώσσες έχουν ριζικές διαφορές μεταξύ τους, τόσο από άποψης σύνταξης όσο και μορφολογίας. Η προσπάθειά μας επικεντρώθηκε στην χρήση στατιστικών μεθόδων, με την ελάχιστη δυνατή γλωσσική πληροφορία, για τη δημιουργία ενός δίγλωσσου λεξικού μονολεκτικών και πολυλεκτικών όρων. Για την αντιστοίχηση των πολυλεκτικών όρων, το σύστημά μας καταφέρνει κατ' αρχάς να τους εντοπίσει, ανεξάρτητα από το πλήθος των λέξεων που τους αποτελούν, ακόμα και αν δεν αποτελούνται από αυστηρά συνεχόμενες λέξεις. Το παρουσιαζόμενο σύστημα δοκιμάστηκε και μεταξύ Ελληνικών – Αγγλικών. Τα αποτελέσματα από την προσπάθειά μας αξιολογήθηκαν ως αρκετά ικανοποιητικά, χωρίς να αποκλείουν την δυνατότητα περαιτέρω βελτίωσης του συστήματος.

Abstract

The work we present is an attempt to create a system for Automatic Extraction of Bilingual Lexicons, whose effectiveness we test on Greek and Turkish parallel corpora. The difficulty in the selected pair of languages is the lack of appropriate linguistic tools that could support such an effort, and also the fact that the two languages have radical differences in both their syntax and morphology. Our efforts were focused on using statistical methods, with minimal language-specific information, to construct a bilingual lexicon of single and multi-word terms. In order to match the multi-word terms, our system initially manages to locate them, regardless of how many words they consist of, even when their words are not strictly adjacent. The system we present has also been tested on Greek and English corpora. The results of our work have been rated as quite satisfactory, without eliminating the possibility of further improvements.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή κ. Kemal Oflazer, Sabanci University, Faculty of Engineering and Natural Sciences για την άμεση βοήθεια που μου παρείχε στην ολοκλήρωση της εργασίας μου με τη λημματοποίηση όλων των τουρκικών λέξεων των κειμένων που χρησιμοποίησα.

Επίσης θα ήθελα να ευχαριστήσω τον καθηγητή μου, κ. Ιωάννη Τοπσόγλου, καθηγητή της τουρκικής γλώσσας του Κέντρου Σπουδών Νοτιο-Ανατολικής Ευρώπης, για τον χρόνο τον οποίο αφιέρωσε στην αξιολόγηση των αποτελεσμάτων της εργασίας μου, χωρίς τη βοήθεια του οποίου θα ήταν πολύ δύσκολη η ολοκλήρωσή της.

Κεφάλαιο 1ο

Εισαγωγή

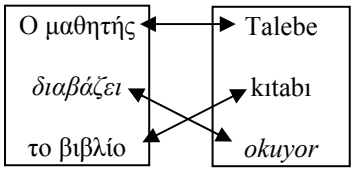
Στο τομέα της Μηχανικής Μετάφρασης (ΜΜ) είναι συνηθισμένη η χρήση εργαλείων που μπορούν να υποβοηθήσουν σημαντικά τη διαδικασία μετάφρασης. Πολύ σημαντικά μεταφραστικά εργαλεία είναι τα δίγλωσσα λεξικά με αντιστοιχίες λέξεων μεταξύ δύο γλωσσών: της Γλώσσας – Πηγής (Source Text) και της Γλώσσας – Στόχου (Target Text = μεταφρασμένο κείμενο). Η έλλειψη όμως λεξικών σε ηλεκτρονική μορφή για ζεύγη γλωσσών όπως Ελληνικά – Τουρκικά καθώς και η δυσκολία κατασκευής τους με τη διαδικασία πληκτρολόγησης της κάθε λέξεως καθιστούν ιδιαίτερα σημαντική τη δυνατότητα υλοποίησής τους με αυτοματοποιημένες μεθόδους. Από πειράματα που έχουν γίνει στο χώρο της ΜΜ, οι καθαρά Στατιστικές Μέθοδοι έχουν να επιδείξουν πολύ καλά αποτελέσματα. Συνήθως η χρήση τους συνδυάζεται και με τεχνικές Μηχανικής Μάθησης και όταν τα δεδομένα εκμάθησης αφορούν κάποιο συγκεκριμένο Γνωστικό Τομέα (Domain), τότε τα αποτελέσματα των Στατιστικών Μεθόδων είναι ανώτερα ακόμα και από τεχνικές που χρησιμοποιούν χειρωνακτικά κατασκευασμένους γλωσσικούς πόρους. Σημαντικότατο πλεονέκτημα των στατιστικών αλγορίθμων είναι ότι μπορούν να εφαρμοστούν για πολύ μεγαλύτερο πλήθος γλωσσών χωρίς καμία μετατροπή.

Ένα τέτοιο δίγλωσσο λεξικό επιχειρούμε να κατασκευάσουμε από την επεξεργασία δύο παράλληλων κειμένων. Χρησιμοποιούμε μόνο Στατιστικές Μεθόδους, με την ελάχιστη δυνατή γλωσσική πληροφορία. Οι δοκιμές μας έγιναν σε δύο γλώσσες με μεγάλες διαφορές στη σύνταξη και μορφολογία, την Ελληνική και την Τουρκική.

1.1. Περιγραφή του Προβλήματος

Η αυτόματη αντιστοίχηση των λέξεων δύο παράλληλων σωμάτων κειμένων χωρίς τη χρήση γλωσσικής πληροφορίας αποτελεί αναμφίβολα μια δύσκολη υπόθεση. Η δυσκολία του εγχειρήματος θα γίνει περισσότερο κατανοητή αν αναλογιστούμε τα προβλήματα που θα αντιμετωπίσουμε και που οφείλονται στις διαφορές που έχουν οι γλώσσες της εργασίας.

1.1.1. Συντακτικές διαφορές



Όλες οι γλώσσες μπορούν να χωριστούν σε 3 μεγάλες κατηγορίες ανάλογα με τη σύνταξή τους. Η κατηγοριοποίηση αυτή αφορά την σχετική θέση που έχουν το Υποκείμενο (Υ), το Αντικείμενο (Α) και το Ρήμα (Ρ) μεταξύ τους. Έτσι έχουμε τις εξής κατηγορίες:

- 1η κατηγορία : «Υποκείμενο – Ρήμα – Αντικείμενο» (Υ-Ρ-Α). Στην κατηγορία αυτή ανήκει η Ελληνική γλώσσα. (π.χ. Ο μαθητής *διαβάζει* το βιβλίο) καθώς και πολλές άλλες γλώσσες, όπως τα Αγγλικά.
- 2η κατηγόρια : «Υποκείμενο – Αντικείμενο – Ρήμα» (Υ-Α-Ρ), όπου το ρήμα είναι στο τέλος της πρότασης. Στην κατηγόρια αυτή ανήκουν τα Τουρκικά (π.χ. Talebe kıtabı *okuyor*).
- Η 3η κατηγορία είναι η πιο σπάνια και είναι της μορφής Αντικείμενο – Ρήμα – Υποκείμενο (Α-Ρ-Υ).

Η αντιστοίχηση λέξεων σε ζεύγη γλωσσών που ανήκουν σε διαφορετικές συντακτικές κατηγορίες είναι πιο πολύπλοκη από ό,τι σε ζεύγη γλωσσών ίδιας συντακτικής κατηγορίας και πολλά συστήματα Μηχανικής Μετάφρασης που δεν έχουν σχεδιαστεί για να αντιμετωπίσουν τέτοιες περιπτώσεις, όπως το IBM Model 5, δεν μπορούν να αποδώσουν ικανοποιητικά [9]. Το σύστημά μας μπορεί να εφαρμοστεί σε γλώσσες όλων των συντακτικών κατηγοριών, αφού όπως θα δούμε οι διαφορές στη σύνταξη δεν επηρεάζουν το αποτέλεσμα του μηχανισμού αντιστοίχησης λέξεων.

1.1.2. Μορφολογικές Διαφορές

Άλλη μια κατηγοριοποίηση των γλωσσών είναι σε «Συγκολλητικές» (Agglutinative) ή «Μη Συγκολλητικές». Ο διαχωρισμός αυτός γίνεται με βάση τη μορφολογία τους και συγκεκριμένα ως προς την τάση τους να προσκολλούν σε μία μεγάλη λέξη άλλες μικρότερες. Οι προστιθέμενες λέξεις συνήθως προσκολλώνται στο τέλος των λέξεων και μπορεί να είναι προθέσεις, αντωνυμίες ή ακόμα και βοηθητικά ρήματα. Η Τουρκική γλώσσα όπως και η Ουγγρική και Φιλανδική είναι συγκολλητικές, σε αντίθεση με τις Ελληνική και Αγγλική οι οποίες δεν είναι.

Εκτός όμως από τις διαφορές μεταξύ των δύο γλωσσών, το γεγονός ότι τα ουσιαστικά και τα ρήματα κλίνονται τόσο στην Ελληνική όσο και στην Τουρκική είναι ένα κοινό στοιχείο το οποίο τις διαφοροποιεί από άλλες γλώσσες όπως τα Αγγλικά. Όμως αν και πρόκειται για ομοιότητα, το κοινό αυτό στοιχείο αποτελεί πρόβλημα στην αντιστοίχηση των λέξεων με στατιστική μέθοδο, διότι η ίδια λέξη μπορεί να εμφανίζεται στο κείμενο με πολλές διαφορετικές μορφές. Ακόμα και μετά από ομαδοποίηση των λέξεων των προερχόμενων από το ίδιο λήμμα, η ακριβής αντιστοίχηση της κάθε λέξης δεν είναι εύκολη.

π.χ. :	ΕΛ	TU
	φίλος	= arkadaş
δεν είμαι στους φίλους μου	=	arkadaşlarımda değilim

1.1.3. Γλωσσική Οικογένεια

Είναι γεγονός ότι οι γλώσσες μπορούν να χωριστούν σε μεγαλύτερες οικογένειες που έχουν κοινή προέλευση. Έτσι, για παράδειγμα, έχουμε τις «Δυτικοευρωπαϊκές» γλώσσες, που έχουν «καταγωγή» από τα Λατινικά, αλλά και άλλες όπως τις Σλαβικές, Αραβικές κ.λ.π. γλώσσες. Είναι συχνό φαινόμενο λέξεις δύο γλωσσών της ίδιας «γλωσσικής οικογένειας» που περιγράφουν την ίδια έννοια να έχουν πολλά από τα γράμματα τους κοινά (cognates). Στην περίπτωση όμως Ελληνικών - Τουρκικών οι δύο γλώσσες ανήκουν σε διαφορετικές οικογένειες και μόνο όταν πρόκειται για λέξεις που έχουν «δανειστεί» η μία γλώσσα από την άλλη, παρατηρείται κάποια μικρή αντιστοιχία των γραμμάτων των αντίστοιχων λέξεων.

Αν λάβουμε υπόψη μας ότι υπάρχουν μέθοδοι μετάφρασης που συγκρίνουν τα γράμματα που αποτελούν μια λέξη με τα γράμματα μιας λέξης της άλλης γλώσσας και προσπαθούν να εκτιμήσουν την πιθανότητα να αποτελεί η μία μετάφραση της άλλης, μπορούμε να καταλάβουμε τις σημαντικές δυσκολίες που θα αντιμετώπιζαν αυτές οι μέθοδοι στην αντιστοίχηση λέξεων δύο γλωσσών που ανήκουν σε διαφορετικές οικογένειες γλωσσών, με εντελώς διαφορετικό αλφάβητο και ελάχιστες κοινές ρίζες.

Ένα άλλο πρόβλημα στην μετάφραση μεταξύ γλωσσών που δεν ανήκουν στην ίδια γλωσσική οικογένεια είναι οι εμφανίσεις ανωμαλιών στην μετάφραση, που οφείλονται σε περιπτώσεις λέξεων οι οποίες δεν έχουν αντίστοιχη μονολεκτική μετάφραση στην άλλη γλώσσα. Για παράδειγμα, στα Τουρκικά η λέξη “dayı” σημαίνει «ο θείος από την οικογένεια της μητέρας», μια έννοια για την οποία δεν υπάρχει αντίστοιχη Ελληνική λέξη.

1.2. Αντικειμενικός Σκοπός του Συστήματος

Το σύστημα που αναπτύξαμε έχει τη δυνατότητα, με μόνο δεδομένο τα παράλληλα σώματα κειμένων δύο γλωσσών, με οποιεσδήποτε διαφορές στη δομή των γλωσσών (όπως Ελληνικά και Τουρκικά), με ελάχιστη γλωσσική πληροφορία και με υποβοήθηση από ένα εργαλείο στοίχισης κειμένων (**Tr•AID Align** [2a]), να κατασκευάζει εύκολα και γρήγορα ένα δίγλωσσο λεξικό, στο οποίο θα συμπεριλαμβάνονται τόσο μονολεκτικοί όσο και πολυ-λεκτικοί όροι. Το λεξικό αυτό θα είναι εστιασμένο σε συγκεκριμένο γνωστικό τομέα (domain), ανάλογα με τα κείμενα που θα χρησιμοποιηθούν ως δεδομένα εκπαίδευσης.

Μπορούμε να αναλογιστούμε τη σημασία ενός τέτοιου συστήματος αν λάβουμε υπόψη:

- Πόσο δύσκολο και χρονοβόρο θα ήταν να κατασκευάσουμε το λεξικό χειρωνακτικά, ιδιαίτερα αν τα κείμενα περιέχουν τεχνικούς όρους που δεν περιλαμβάνονται σε γενικής χρήσεως υπάρχοντα λεξικά.

- Τις πολλαπλές χρήσεις που μπορεί να έχει ένα τέτοιο εργαλείο, τόσο σε διαφορετικές εφαρμογές (π.χ. ενσωμάτωση σε συστήματα ΜΜ) όσο και σε διαφορετικά ζεύγη γλωσσών.

- Μια από τις δυνατότητες χρήσης ενός τέτοιου συστήματος θα μπορούσε να είναι και η αποκωδικοποίηση / αποκρυπτογράφηση κειμένων. Σε περίπτωση, δηλαδή, που διαθέτουμε ένα κρυπτογραφημένο κείμενο μαζί με την αποκρυπτογραφημένη του μορφή, θα ήταν δυνατό να κατασκευάσουμε αυτόματα ένα λεξικό με τις λέξεις του κειμένου και βασιζόμενοι σ' αυτές να βοηθηθούμε στην αποκρυπτογράφηση άλλων κειμένων.

Η επιλογή του συγκεκριμένου ζεύγους γλωσσών (Ελληνικών – Τουρκικών) έγινε όχι μόνο γιατί δεν υπάρχουν έτοιμα ανάλογα προγράμματα, αλλά και γιατί, λόγω των μεγάλων διαφορών που υπάρχουν μεταξύ αυτών των γλωσσών, δοκιμάζεται έτσι με τα αυστηρότερα κριτήρια η αποτελεσματικότητα του συστήματός μας.

Τα αποτελέσματα των δοκιμών μας είναι αρκετά ενθαρρυντικά και σε αρκετά ενδιάμεσα στάδια του συστήματος το ποσοστό επιτυχίας είναι μεγαλύτερο από 90 %.

Τέλος αξίζει να σημειωθεί ότι παράλληλα με τη διαδικασία αντιστοίχησης λέξεων χωρίς γλωσσική πληροφορία, δοκιμάσαμε και την αποτελεσματικότητα της χρήσης περιορισμένης γλωσσικής πληροφορίας με τη μορφή επιπλέον εισαγόμενων γλωσσικών μετα-δεδομένων. Τα δεδομένα που αφορούσαν την Τουρκική γλώσσα αποκτήθηκαν μετά από συνεργασία με τον καθηγητή κ. Kemal Oflazer[1].

Αξιολογώντας τα αποτελέσματα του συστήματός μας εκτιμούμε ότι θα μπορούσαμε να είχαμε ακόμα καλύτερα αποτελέσματα αν είχαν υπάρξει περισσότερα δεδομένα εκπαίδευσης αλλά και γλωσσικά εργαλεία που να υποστηρίζουν και τις δύο γλώσσες (π.χ. λημματοποιητές – lemmatizers).

1.3. Δομή του Εγγράφου

Στη συνέχεια, στο κεφάλαιο 2, θα γίνει μια σύντομη ανασκόπηση σχετικών τεχνικών που έχουν προταθεί ως σήμερα. Στο κεφάλαιο 3 θα αναλύσουμε όλα τα στάδια της διαδικασίας αντιστοίχησης των λέξεων και του εντοπισμού και αντιστοίχησης των πολυλεκτικών όρων των δύο γλωσσών στο σύστημα που αναπτύξαμε. Τα αποτελέσματα των δοκιμών μας θα τα παρουσιάσουμε στο κεφάλαιο 4 και τέλος στο κεφάλαιο 5 θα παρουσιάσουμε τα συμπεράσματα της εργασίας, και θα αναφερθούμε σε πιθανές μελλοντικές επεκτάσεις.

---

[1] Kemal Oflazer , Sabanci University , Faculty of Engineering and Natural Sciences , Orhanli, 81474 Tuzla, Istanbul, Turkey , http://people.sabanciuniv.edu/~oflazer/

Κεφάλαιο 2ο

Επισκόπηση Χρησιμοποιούμενων Τεχνικών

Η δημιουργία ενός δίγλωσσου λεξικού είναι το αποτέλεσμα μιας διαδικασίας η οποία σε γενικές γραμμές αποτελείται από τα παρακάτω αλληλοεξαρτώμενα στάδια:

- Στοίχιση Κειμένων.
- Ομαδοποίηση Λέξεων.
- Αντιστοίχηση Λέξεων.
- Εντοπισμό & Μετάφραση 2-λεκτικών & Πολυλεκτικών Όρων.

Στο κεφάλαιο αυτό θα παρουσιάσουμε μερικές από τις εφαρμοζόμενες τεχνικές ανά στάδιο, οι οποίες έχουν προταθεί ή έχουν χρησιμοποιηθεί από άλλους, ενώ στο επόμενο κεφάλαιο θα παρουσιάσουμε τις μεθόδους που χρησιμοποιήσαμε.

2.1. Στοίχιση Κειμένων (*Text Alignment*)

Το πρόβλημα της στοίχισης δύο παράλληλων κειμένων σε διαφορετικές γλώσσες είναι το πρώτο που αντιμετωπίσαμε. Σημαντικό παράγοντα στη μεθοδολογία στοίχισης παραλλήλων κειμένων αποτελεί η μονάδα κειμένου (π.χ. πρόταση, επίπεδο μικρότερο της πρότασης, φράσεις, λέξεις) [1] με βάση την οποία θα γίνει η στοίχιση. Η επιλογή της μονάδας κειμένου επηρεάζει όχι μόνο τον μηχανισμό στοίχισης κειμένων αλλά και αντιστοίχησης λέξεων. Επειδή η επιλογή του επιπέδου της πρότασης εξασφαλίζει την άρση πιθανών μεταφραστικών σημασιολογικών αμφισημιών [21], ως μονάδα κειμένου στα περισσότερα συστήματα επιλέγεται η πρόταση.

Η ορθή στοίχιση των κειμένων είναι θεμελιώδους σημασίας για την περαιτέρω συνέχεια, καθώς οποιαδήποτε λάθη σε αυτό το στάδιο θα επηρεάσουν αρνητικά όλα τα επόμενα στάδια. Το πρόβλημα που αντιμετωπίζουμε στο στάδιο αυτό είναι ότι δεν υπάρχει πάντα αντιστοίχηση 1:1 μεταξύ των προτάσεων των δύο γλωσσών, είτε λόγω ιδιομορφίας της κάθε γλώσσας είτε ακόμα και λόγω μεταφραστικών αναγκών. Στην πράξη, είναι δυνατό να υπάρχουν περιπτώσεις με αντιστοίχηση προτάσεων 2:1, 2:2, 1:2, 1:0, 0:1, 2:0, 0:2, 3:1, 1:3.

Στην συνέχεια του κειμένου θα χρησιμοποιήσουμε τον όρο «Περίοδο» για τις περιπτώσεις αντιστοίχησης οι οποίες αφορούν περισσότερες από 1 προτάσεις. Έτσι οι αντιστοιχήσεις προτάσεων 2:1, 2:2, 3:1 αφορούν «Αντιστοίχηση Περιόδων» 1:1 όπου κάθε περίοδος μπορεί να αποτελείται από διαφορετικό πλήθος προτάσεων (1, 2 ή 3).

2.1.1. Αλγόριθμος GALE & CHURCH [5]

Η πλειοψηφία των προτεινόμενων μεθοδολογιών στοίχισης προτάσεων βασίζονται σε μια πολυεπίπεδη αρχιτεκτονική με κύριο άξονα το μηχανισμό των Gale & Church [5]. Ο αλγόριθμος που πρότειναν και υλοποίησαν (μεταξύ Αγγλικών και Γαλλικών) στηρίχθηκε στην εξής απλή παρατήρηση:

Μεγάλες προτάσεις της μιας γλώσσας, μεταφράζονται σε μεγάλες προτάσεις της άλλης και μικρές προτάσεις της μιας γλώσσας μεταφράζονται σε μικρές προτάσεις της άλλης γλώσσας.

Έτσι εντόπισαν ότι υπάρχει μια σχέση στο μέγεθος των προτάσεων των δύο γλωσσών, ως προς το πλήθος των γραμμάτων που περιέχουν. Σε κάθε γράμμα της μιας γλώσσας αντιστοιχούν, κατά μέσο όρο, c γράμματα της άλλης γλώσσας με διακύμανση s^2. Η πιθανότητα μια πρόταση μήκους l_1 γραμμάτων της γλώσσας Α να μεταφράζεται σε πρόταση μήκους l_2 της γλώσσας Β υπολογίζεται με τη βοήθεια της βαθμολογίας που επιτυγχάνει το ζεύγος των προτάσεων:

$$-log_2 (1\text{-Prob}(|\delta|))$$

όπου το δ εξαρτάται από τα μήκη των 2 προτάσεων :

$$\delta = \frac{(l_2 - l_1 \times c)}{\sqrt{l_1 \times s^2}}$$

και έχει μια κανονική κατανομή με μέση τιμή = 0 και διακύμανση = 1.

Η βαθμολογία αυτή υπολογίζεται αναδρομικά για όλους τους πιθανούς συνδυασμούς των διαδοχικών ζευγών των προτάσεων της κάθε γλώσσας και με ένα αλγόριθμο δυναμικού προγραμματισμού επιλέγεται η αντιστοίχηση προτάσεων που θα έχει τη καλύτερη δυνατή βαθμολογία για το σύνολο των προτάσεων. Κατά την βαθμολογία των αντιστοιχήσεων δοκιμάζονται οι αντιστοιχήσεις ανά 4 προτάσεις (2 διαδοχικές από κάθε γλώσσα) και υπολογίζονται οι βαθμολογίες για κάθε δυνατή αντιστοίχηση (δηλ. 1:1, 1:0, 0:1, 2:1, 1:2, 2:2) και κάθε φορά επιλέγεται ο συνδυασμός εκείνος με την καλύτερη βαθμολογία.

Οι τιμές των c, s^2 υπολογίστηκαν εμπειρικά για τα ακόλουθα ζεύγη γλωσσών:

- Αγγλικά – Γαλλικά : c= 1,06 , s^2= 5,6
- Αγγλικά – Γερμανικά : c= 1,1 , s^2= 7,3 .

Η διαφορά ανάμεσα στις τιμές των c και s^2, που εμφανίζουν διαφορετικά ζεύγη γλωσσών, δεν είναι ιδιαίτερα μεγάλη ώστε να επηρεάζει σημαντικά την τελική βαθμολογία. Έτσι οι Gale & Church, για να γενικεύσουν τον αλγόριθμο τους για όλες τις γλώσσες, υιοθέτησαν τις τιμές :

$$c = 1$$
$$s^2 = 6,8$$

2.1.2. Άλλοι Αλγόριθμοι Στοίχισης Προτάσεων

Η αντιστοίχηση των προτάσεων θα μπορούσε να γίνει και με τη στοίχιση λέξεων [5], εφόσον διαθέτουμε κάποιο δίγλωσσο λεξικό σε ηλεκτρονική μορφή. Ακόμα όμως και χωρίς λεξικό θα μπορούσε να γίνει μια περιορισμένου βάθους στοίχιση, χρησιμοποιώντας μόνο κάποιες «λέξεις-άγκυρες», όπως ημερομηνίες, αριθμητικά ποσά, κύρια ονόματα, που συνήθως έχουν παρόμοια γραφή στις δύο γλώσσες. Χρησιμοποιώντας σε πρώτη φάση την αντιστοίχηση περιοχών κειμένου με βάση τις λέξεις-άγκυρες, ο Brown et al. [10] εισηγήθηκαν πριν τους Gale & Church έναν αλγόριθμο όπου ως μονάδα μέτρησης του μήκους των προτάσεων χρησιμοποιούνται οι λέξεις αντί των γραμμάτων.

2.2. Ομαδοποίηση Λέξεων (*Word Conflation*)

Οι στατιστικές μέθοδοι βασίζονται στη καταμέτρηση του πλήθους των εμφανίσεων κάθε λέξεως. Ένα πρόβλημα που θα αντιμετωπίσουμε, όμως, είναι ότι κάθε λέξη μπορεί να εμφανίζεται με πολλές μορφές σε ένα κείμενο (διαφορετικές πτώσεις, χρόνους κ.λ.π.), και έτσι αν προσπαθήσουμε να αντιστοιχήσουμε κάθε μεμονωμένη μορφή των λέξεων, η δυσκολία θα είναι μεγάλη καθώς το πλήθος τους δεν θα είναι αρκετά μεγάλο ώστε να οδηγήσει σε ασφαλή συμπεράσματα. Επιπλέον, το γεγονός ότι οι μορφές με τις οποίες μπορεί να εμφανιστούν κάποιες λέξεις σε μια γλώσσα δεν είναι σε αντιστοιχία με τις μορφές τους στις άλλες γλώσσες δυσχεραίνει ακόμα περισσότερο την αντιστοίχησή τους[2].

Τα αποτελέσματα της αντιστοίχησης λέξεων θα ήταν πιο ακριβή αν ομαδοποιούνταν όλες οι διαφορετικές μορφές της ίδιας λέξεως, ενισχύοντας έτσι το μέγεθος του δείγματος που θα χρησιμοποιηθεί για την αντιστοίχησή τους. Το πρόβλημα της ομαδοποίησης είναι ουσιαστικά πρόβλημα συσχέτισης των ποικίλων μορφών της κάθε λέξεως και η δυσκολία του είναι ανάλογη με τη μορφολογία της κάθε γλώσσας.

Οι περισσότερες μέθοδοι που έχουν χρησιμοποιηθεί μέχρι σήμερα για να συσχετίσουν και να ομαδοποιήσουν διαφορετικές μορφές της ίδιας λέξεως βασίζονται στην παρατήρηση ότι συνήθως οι διαφοροποιήσεις εντοπίζονται στην κατάληξή τους. Παρόλο που αυτή η παρατήρηση δεν είναι απόλυτη, ισχύει σε μεγάλο βαθμό για τις περισσότερες γλώσσες, ενώ οι περιπτώσεις όπου οι διαφοροποιήσεις οφείλονται σε προθέματα είναι πιο σπάνιες.

Μια λύση στο πρόβλημα της συσχέτισης των λέξεων μπορεί εύκολα να βρεθεί αν ομαδοποιήσουμε τις λέξεις που έχουν το ίδιο θέμα (Stemming) ή το ίδιο λήμμα (Lemmatization). Και οι δύο αυτές μέθοδοι όμως προϋποθέτουν αλγόριθμους εφοδιασμένους με κατάλληλη γλωσσική πληροφορία, συνήθως με μορφή κανόνων γραμματικής. Ένα τέτοιο πολύ απλό αλγόριθμο για τον εντοπισμό του θέματος μιας λέξεως, με ελάχιστους κανόνες γραμματικής, για την Αγγλική γλώσσα εισηγήθηκε ο Krovetz [12], χρησιμοποιώντας μόνο την απαλοιφή συγκεκριμένων καταλήξεων, όπως καταλήξεις πληθυντικού αριθμού για τα ουσιαστικά και διαφορετικών χρόνων για τα ρήματα. Τα αποτελέσματά του ήταν αρκετά καλά αλλά σε αυτό βοήθησε και η απλότητα της Αγγλικής γραμματικής. Από τους ευρύτερα διαδεδομένους αλγόριθμους εντοπισμού του θέματος των λέξεων είναι ο αλγόριθμος του Porter [13], ο οποίος έχει υλοποιηθεί για πολλές ευρωπαϊκές γλώσσες (όχι όμως για Ελληνικά και Τουρκικά). Η υλοποίηση του για τα Αγγλικά αποτελείται από 5 στάδια, στα οποία εφαρμόζονται στην εξεταζόμενη λέξη περίπου 1200 κανόνες για διαδοχική αφαίρεση καταλήξεων.

Δεδομένου ότι σε πολλές γλώσσες υπάρχουν περιπτώσεις όπου μια λέξη, ανάλογα με τα συμφραζόμενά της, μπορεί να έχει εντελώς διαφορετικό νόημα (π.χ. η Τουρκική λέξη "yaz" μπορεί να σημαίνει «καλοκαίρι» ή «γράψε!», με εντελώς διαφορετικό λήμμα σε κάθε περίπτωση), οι τυχόν αμφισημίες, για τον εντοπισμό του ορθού λήμματος ή θέματος, είναι δυνατό να επιλυθούν αν προηγηθεί η ανάλυση του κειμένου στα μέρη του λόγου (part-of-speech tagging).

Υπάρχουν όμως και άλλες προσεγγίσεις στο πρόβλημα ομαδοποίησης των λέξεων, που σημαντικότερο πλεονέκτημά τους είναι το γεγονός ότι δεν απαιτούν καμία απολύτως γλωσσική πληροφορία. Οι μέθοδοι αυτές βασίζονται στο ταίριασμα συμβολοσειρών (String Matching) και οι δύο πιο δημοφιλείς από αυτές είναι:

- Το Ποσοστό Μέγιστης Κοινής Υποακολουθίας (Longest Common Sub-sequence Ratio).
- Η μέθοδος των N-Grams.

Στην παράγραφο 2.3.2.1.2 θα αναλύσουμε περισσότερο την μέθοδο αξιολόγησης του Ποσοστού Μέγιστης Κοινής Υποακολουθίας. Τα αποτελέσματα όμως από την χρήση αυτής της μεθόδου μπορεί να είναι αρκετά παραπλανητικά, καθώς οι χαρακτήρες που βρίσκονται στην αρχή ή στο μέσο μιας λέξεως έχουν πολύ μεγαλύτερη βαρύτητα από τους χαρακτήρες της κατάληξης. Έτσι, για παράδειγμα, η λέξη «χρώματα» έχει εντελώς διαφορετικό νόημα από την «χρήματα», παρόλο που το ποσοστό μέγιστης κοινής υποακολουθίας είναι ιδιαίτερα υψηλό (έχουμε διαφοροποίηση μόνο στο 3^{ο} γράμμα). Για την αντιμετώπιση αυτού του προβλήματος, οι Simard κ.ά. [19] προτείνουν τον υπολογισμό

[2] Βλ. για παράδειγμα §3.2.3, πίνακα 3.

μόνο της μέγιστης *αρχικής* υποακολουθίας. Ωστόσο το πρόβλημα δεν λύνεται πλήρως, καθώς λέξεις με κοινό πρόθεμα τείνουν να ομαδοποιηθούν λάθος (π.χ. «παρακάτω» και «παρακάνω»).

Η μέθοδος των N-Grams [11] αποτελεί μια αρκετά διαδεδομένη μέθοδο ομαδοποίησης όρων χωρίς τη χρήση γλωσσικής πληροφορίας. Η κύρια ιδέα αυτής της μεθόδου είναι η ομαδοποίηση λέξεων οι οποίες έχουν πολλά κοινά υπο-τμήματα (substrings) μήκους N διαδοχικών χαρακτήρων. Η μετρική ομοιότητας δύο λέξεων υπολογίζεται με τον παρακάτω μαθηματικό τύπο:

$$\text{Ομοιότητα : } S = \frac{2 \cdot C}{A+B}$$

C = πλήθος κοινών N-Gram
A = πλήθος N-Gram της $1^{η}$ λέξεως
B = πλήθος N-Gram της $2^{η}$ λέξεως

Όσο η τιμή του S πλησιάζει το 1 τόσο πιο πιθανό είναι οι δύο λέξεις να πρέπει να ομαδοποιηθούν μαζί. Στο παράδειγμα που ακολουθεί βλέπουμε πως η τιμή του S βοηθάει να συσχετίσουμε ορθά τις λέξεις «προσάρμοσε» και «προσάρμοσαν». Επίσης, πολύ σωστά η τιμή του S είναι χαμηλότερη για τις λέξεις «προσάρμοσε» και «προσάρτησε», παρότι η μεταξύ τους διαφορά είναι μόνο 2 χαρακτήρες.

Ανάλυση σε N-Grams (N=2):		
προσάρμοσε	πρ ρο οσ σά άρ ρμ μο οσ σε	9
προσάρμοσαν	πρ ρο οσ σά άρ ρμ μο οσ σα αν	10
προσάρτησε	πρ ρο οσ σά άρ ρτ τη ησ σε	9

Συγκρινόμενες Λέξεις	Κοινά 2-Grams	C	Ομοιότητα
προσάρμοσε & προσάρμοσαν	πρ ρο οσ σά άρ ρμ μο οσ	8	2·8/(9+10) = **0.842**
προσάρμοσε & *προσάρτησε*	πρ ρο οσ σά άρ σε	6	2·6/(9+9) = **0.666**
προσάρμοσαν & *προσάρτησε*	πρ ρο οσ σά άρ	5	2·5/(10+9) = **0.526**

Όπως ισχυρίζεται ο Kosinov [11], μετά από πειράματα υπολογίστηκε ότι αν N=2, η απόδοση αυτής της μεθόδου μπορεί να ξεπεράσει ακόμα και αυτή μεθόδων stemming που βασίζονται στον αλγόριθμο του Porter.

2.3. Αντιστοίχηση Λέξεων

Στην επόμενη και κυριότερη φάση του συστήματος θα πρέπει από όλες τις αντιστοιχισμένες προτάσεις να συσχετίσουμε κάθε λέξη με ένα σύνολο από υποψήφιες λέξεις - μεταφράσεις της άλλης γλώσσας και στο τέλος να επιλέξουμε την καλύτερη από όλες.

Η «Αντιστοίχηση των λέξεων» (Word Alignment) γίνεται σε 3 στάδια:

- Εντοπισμός λεκτικών μονάδων (Tokenization).
- Εντοπισμός και περιορισμός υποψήφιων μεταφράσεων.
- Βαθμολογία και επιλογή αντιστοιχίας λέξεων.

Το κρισιμότερο στάδιο από όλα, είναι το τελευταίο. Σε αυτό γίνεται η επιλογή της καταλληλότερης μετάφρασης από το σύνολο των υποψήφιων λέξεων της άλλης γλώσσας που εντοπίσαμε στη προηγούμενη φάση. Η σωστή επιλογή μπορεί να επιτευχθεί:

- είτε με έναν αλγόριθμο Δυναμικού Προγραμματισμού που επιδιώκει να αντιστοιχήσει τις λέξεις έτσι ώστε να επιτύχει τη μέγιστη δυνατή συνολική βαθμολογία για όλα τα ζεύγη των αντιστοιχισμένων λέξεων,

- είτε με ένα «άπληστο» (greedy) αλγόριθμο, ο οποίος για κάθε λέξη της μίας γλώσσας, επιλέγει εκείνη τη λέξη της άλλης γλώσσας η οποία έχει την καλύτερη βαθμολογία από τις υπόλοιπες υποψήφιες.

Η βαθμολογία, με βάση την οποία θα γίνει η επιλογή αντίστοιχης λέξεως, είναι το αποτέλεσμα μιας αξιολόγησης η οποία βασίζεται στις μεθόδους που θα παρουσιάσουμε παρακάτω.

2.3.1. Εντοπισμός Υποψήφιων Μεταφράσεων

Οι μέθοδοι βαθμολογίας και επιλογής υποψηφίων μεταφράσεων συχνά έχουν να επιλέξουν από ένα πολύ μεγάλο πλήθος λέξεων. Για λόγους απόδοσης, αποτελεί κοινή πρακτική η μεσολάβηση ενός σταδίου περιορισμού του πλήθους των προς εξέταση λέξεων, πριν από την διαδικασία βαθμολογίας για την τελική αντιστοίχησή τους. Μέσα από αυτή τη διαδικασία δημιουργείται για κάθε λέξη της γλώσσας Πηγής, ένα υποσύνολο από τις λέξεις των αντιστοιχισμένων προτάσεων της γλώσσας Στόχου, οι οποίες έχουν περισσότερες πιθανότητες από τις υπόλοιπες να αποτελούν τη μετάφρασή της. Η διαδικασία αυτή βασίζεται συνήθως σε στατιστικές μεθόδους [1]. Σε μερικές περιπτώσεις, η φάση του περιορισμού των υποψήφιων λέξεων εφαρμόζεται μετά από την αξιολόγηση των υποψήφιων μεταφραστικών ζευγών, με σκοπό την επίλυση ισοβαθμιών.

Ο Tufiş [14] περιορίζει τη λίστα των υποψήφιων μεταφράσεων σε πολλά στάδια συνδυάζοντας περισσότερες από μία μεθόδους. Αρχικά η λίστα υποψηφίων μεταφράσεων περιορίζεται μόνο στις λέξεις που ανήκουν στην ίδια κατηγορία μερών του λόγου, πρακτική που ακολουθείται και από τους Kageura κ.ά. [16] (π.χ. τα ρήματα μεταφράζονται σε ρήματα). Η μέθοδος αυτή όμως απαιτεί τη χρήση ενός επισημειωτή των μερών του λόγου (POS-Tagger) και ενός προγράμματος εντοπισμού των λημμάτων των λέξεων. Η βασική μέθοδος που επέλεξε ο Tufiş [14] για τον περιορισμό της λίστας υποψήφιων μεταφράσεων είναι η αξιολόγηση των υποψήφιων λέξεων με βάση την Λογαριθμική Ομοιότητα (*log-likehood*) θέτοντας ως όριο απόρριψης την τιμή 9.

Ένας άλλος τρόπος με τον οποίο είναι δυνατό να επιτευχθεί το φιλτράρισμα, σύμφωνα με τους Tufiş [14] και Brown κ.ά. [15], είναι αποκλείοντας τις υποψήφιες λέξεις των αντιστοιχισμένων περιόδων της γλώσσας Στόχου, που έχουν πολύ διαφορετική θέση (στη πρόταση που εμφανίζονται) από τη θέση της λέξεως που εξετάζεται. Το όριο που έθεσαν και οι δύο είναι μια ακτίνα ±2 θέσεων από τη θέση της εξεταζόμενης λέξεως. Έτσι π.χ. η $4^{η}$ λέξη μιας περιόδου της γλώσσας Πηγής, είναι πιθανότερο να συσχετίζεται με κάποια από τις {$2^{η}$, $3^{η}$, $4^{η}$, $5^{η}$ και $6^{η}$} λέξεις της αντίστοιχης περιόδου της γλώσσας Στόχου. Προφανώς προϋπόθεση για να λειτουργήσει ορθά αυτή η μέθοδος είναι η σύνταξη των 2 γλωσσών να είναι παρόμοια. Επιπλέον των παραπάνω, ο Tufiş [14] επιλύει τις ισοβαθμίες των υποψήφιων μεταφράσεων συγκρίνοντας και αξιολογώντας τα υποψήφια μεταφραστικά ζεύγη με ταίριασμα συμβολοσειρών (string matching)[3].

Οι Brown κ.ά. [15] βασίζουν την διαδικασία φιλτραρίσματος των υποψήφιων μεταφράσεων στην Αμοιβαία Δεσμευμένη Πιθανότητα (Mutual Conditional Probability) μεταξύ ενός ζεύγους λέξεων των εξεταζόμενων γλωσσών. Αν δηλαδή δεδομένης μιας λέξεως από τη γλώσσα πηγή (x) η πιθανότητα να εμφανιστεί στο ίδιο ζεύγος περιόδων μια συγκεκριμένη λέξη της γλώσσας στόχου (y) είναι αρκετά μεγάλη, αλλά και δεδομένης της ίδιας λέξεως από τη γλώσσα στόχο (y) η πιθανότητα να εμφανιστεί στο ίδιο ζεύγος περιόδων η συγκεκριμένη λέξη της γλώσσας πηγής (x) είναι και αυτή αρκετά μεγάλη, τότε το ζεύγος (x, y) είναι πιθανό να αποτελεί μεταφραστικό ζεύγος.

Στην υλοποίηση αυτής της ιδέας, οι Brown κ.ά. απορρίπτουν από τις λίστες υποψήφιων μεταφράσεων όλα εκείνα τα ζεύγη λέξεων που δεν επιτυγχάνουν βαθμολογία καλύτερη τουλάχιστον από ένα εκ των 3 ορίων αποκοπής που έθεσαν οι ίδιοι. Αν C είναι το πλήθος των κοινών συνεμφανίσεων των (x, y), αποφαινόμαστε ότι οι δύο λέξεις είναι πιθανό να αποτελούν μετάφραση η μίας της άλλης αν:

$$(P(x|y) \geq thr[C] \text{ AND } P(y|x) \geq thr[C]) \text{ OR}$$
$$(P(x|y) \geq thr_1[C] \text{ AND } P(y|x) \geq thr_1[C]) \text{ OR}$$

[3] βλ. § 2.3.2.1.2. και § 2.2.

$$(P(x|y) \geq thr_2[C] \ \ AND \ \ P(y|x) \geq thr_2[C])$$

όπου οι συναρτήσεις thr[C], $thr_1[C]$, $thr_2[C]$ είναι τα 3 διαφορετικά όρια αποκοπής. Οι συναρτήσεις αυτές ορίζονται έτσι ώστε για λέξεις που εμφανίζονται ελάχιστες φορές (C = μικρό) το όριο αποκοπής υποψήφιων (thr[C]) να είναι μεγάλο, ενώ για συχνά συνεμφανιζόμενες λέξεις (C = μεγάλο) το όριο αποκοπής να είναι μικρό.

x = λέξη της γλώσσας Πηγής
y = λέξη της γλώσσας Στόχου
f(x) = πλήθος εμφανίσεων της x

f(x,y) = πλήθος κοινών εμφανίσεων της x και της y σε αντιστοιχισμένες περιόδους

Ο Smadja [6] επιτυγχάνει το φιλτράρισμα των υποψήφιων λέξεων[4] με την αντικατάσταση του f(y) από το f(x,y) στον συντελεστή Dice (βλ. § 2.3.2.1.1.Α). Εξετάζοντας δηλαδή τις εμφανίσεις της λέξεως x, υπολογίζεται άμεσα η τιμή του f(x). Στη συνέχεια εξετάζοντας μόνο τις αντιστοιχισμένες περιόδους που περιέχουν την x, υπολογίζεται εύκολα το πλήθος των κοινών συνεμφανίσεων της x με την λέξη y δηλαδή η f(x,y). Η τεχνική αυτή έχει το πλεονέκτημα ότι φιλτράρει τις υποψήφιες λέξεις αποφεύγοντας τον υπολογισμό του f(y) που θα απαιτούσε την αναζήτηση όλων των εμφανίσεων της y. Για τη τιμή της f(x,y) ισχύει πάντοτε ότι

$f(x,y) \leq f(y)$ και κατά συνέπεια :

$$\text{Dice}(x,y) = \frac{2 \cdot f(x,y)}{f(x)+f(y)} \leq \frac{2 \cdot f(x,y)}{f(x)+f(x,y)}$$

Όσες υποψήφιες μεταφράσεις (y), αξιολογούμενες με τον ελαστικότερο, τροποποιημένο συντελεστή Dice, δεν υπερβούν μια προκαθορισμένη τιμή εκτιμάται ότι δεν έχουν σχέση με την εξεταζόμενη λέξη (x) και αγνοούνται. Οι υπόλοιπες συμπεριλαμβάνονται στις πιθανές μεταφράσεις της x και σε δεύτερο χρόνο θα αναζητηθεί η ακριβής τιμή του f(y), ώστε να υπολογιστεί η κανονική τιμή του συντελεστή Dice.

2.3.2. Βαθμολογία και Επιλογή Αντιστοιχίας Λέξεων

Μέχρι τώρα υπάρχουν αρκετές προτάσεις για το πώς θα πρέπει να υπολογιστεί μια βαθμολογία για κάθε πιθανή μετάφραση. Οι περισσότερες προτάσεις έχουν δοκιμαστεί και εφαρμοστεί μεταξύ ευρωπαϊκών γλωσσών, οι οποίες ανήκουν στην ίδια οικογένεια γλωσσών και έχουν πολλά κοινά (συντακτικά και μορφολογικά) χαρακτηριστικά μεταξύ τους. Πολλές από τις μεθόδους αυτές όμως δεν θα μπορούσαμε να τις εφαρμόσουμε με ικανοποιητικά αποτελέσματα στο ζεύγος Ελληνικών - Τουρκικών. Παρακάτω θα αναφερθούμε σε αυτές τις προτάσεις, χωρισμένες σε δύο μεγάλες κατηγορίες [17] και μία επιπλέον κατηγορία που συνδυάζει τις άλλες δύο, και παράλληλα θα εξηγήσουμε γιατί κάποιες από αυτές θα παρουσίαζαν προβλήματα σε ένα ζεύγος γλωσσών τόσο διαφορετικών μεταξύ τους.

Α. Μέθοδοι *Συσχέτισης* [17]:
Β. Μέθοδοι *Εκτίμησης* [17]:
Γ. Σύνθετες μέθοδοι

2.3.2.1. Μέθοδοι βασισμένες στη Συσχέτιση

Οι μέθοδοι αυτής της κατηγορίας χρησιμοποιούν ευρεστικές μεθόδους, οι οποίες λαμβάνουν υπόψη τους είτε το πλήθος των συνεμφανίσεων των δύο λέξεων είτε την ομοιότητα συμβολοσειρών (String Matching).

2.3.2.1.1. Μέτρηση Συνεμφανίσεων

[4] Χρησιμοποιεί την περιγραφόμενη μέθοδο για αντιστοίχηση πολυλεκτικών όρων αλλά μπορεί να εφαρμοστεί και στην αντιστοίχηση λέξεων.

Μετρώντας τις συνεμφανίσεις δύο λέξεων διαφορετικών γλωσσών στα παράλληλα κείμενα, είναι δυνατό να αποφανθούμε με αρκετή ακρίβεια για το αν η μία είναι μετάφραση της άλλης. Έχουν προταθεί αρκετές μέθοδοι αξιολόγησης αυτού του είδους.

A. Ευρύτατα διαδεδομένος είναι ο Συντελεστής $Dice_{(x,y)}$ (Dice Coefficient):

$$Dice(x,y) = \frac{2 \cdot f(x,y)}{f(x) + f(y)}$$

$f(x)$, $f(y)$: οι συχνότητες εμφάνισης, σε όλο το κείμενο, των λέξεων x, y
$f(x,y)$: το πλήθος των κοινών συνεμφανίσεων των x, y σε στοιχισμένες περιόδους.

Ο Συντελεστής Dice είναι ισοδύναμος με τη μέση αρμονική των δύο εξαρτημένων πιθανοτήτων p(x|y) και p(y|x). Οι τιμές που παίρνει κυμαίνονται στο [0, 1], με τη τιμή 1 να αποδίδεται σε ζεύγη λέξεων που είναι πολύ πιθανό η μία να είναι μετάφραση της άλλης.

B. Άλλη συχνά χρησιμοποιούμενη μετρική βασίζεται στο «Αμοιβαίο Πληροφοριακό Περιεχόμενο»:

$$I(x,y) = log \frac{p(x,y)}{p(x) \cdot p(y)}$$

Η παραπάνω σχέση βασίζεται στη παρατήρηση ότι όταν η παρουσία μιας λέξης x στο κείμενο της μίας γλώσσας παρέχει αρκετή πληροφορία για την παρουσία της λέξεως y σε αντίστοιχη περίοδο στο κείμενο της άλλης γλώσσας και αντίστροφα, τότε πιθανότατα η μία λέξη είναι μετάφραση της άλλης.

2.3.2.1.2. Ταίριασμα Συμβολοσειρών (*String Matching*)

Οι μέθοδοι αξιολόγησης αυτού του είδους συσχετίζουν ζεύγη λέξεων εξετάζοντας τους κοινούς χαρακτήρες των δύο λέξεων. Δύο αλγόριθμοι που χρησιμοποιούνται συνήθως για αυτό το σκοπό, η μέθοδος των N-Grams και η μέθοδος υπολογισμού του ποσοστού της μεγαλύτερης κοινής υποακολουθίας χαρακτήρων (Longest Common String Sub-sequence), προαναφέρθηκαν στην § 2.2 για την «ομαδοποίηση» των λέξεων της κάθε γλώσσας. Η βασική διαφορά κατά τη χρήση τους στην φάση της αντιστοίχησης λέξεων είναι ότι εδώ οι δύο αλγόριθμοι εφαρμόζονται σε λέξεις διαφορετικών γλωσσών. Παρουσιάζουμε αναλυτικά τη δεύτερη μέθοδο, λόγω των ιδιαιτεροτήτων που παρουσιάζει η εφαρμογή της σε δύο τόσο διαφορετικές γλώσσες. Η μέθοδος των N-Grams χρησιμοποιείται όπως στην ενότητα § 2.2.

Ο εντοπισμός της μέγιστης κοινής υποακολουθίας χαρακτήρων μεταξύ 2 λέξεων είναι δυνατό να γίνει με τη χρήση δυναμικού προγραμματισμού [18]. Ειδικότερα όμως στις περιπτώσεις που εξετάζουμε, όπου οι δύο λέξεις δεν ανήκουν στην ίδια γλώσσα αλλά σε γλώσσες οι οποίες μπορεί να έχουν διαφορετικά αλφάβητα, εφαρμόζεται παράλληλα κάποιος αλγόριθμος αντιστοίχησης (mapping) των γραμμάτων – χαρακτήρων της γλώσσας Πηγής με αυτά της γλώσσας Στόχου. Έτσι π.χ. εφαρμόζοντας τον αλγόριθμο αυτό μεταξύ Ελληνικών - Τουρκικών, θα μπορούσαμε να αντιστοιχήσουμε Τουρκικά με Ελληνικά γράμματα όπως στον πίνακα που ακολουθεί:

Τουρκικά	Ελληνικά
a	α
v	β
v	(α ή ε)+υ
d	δ
d	ντ
p	π
r	ρ

Παραδείγματα Αντιστοίχησης Τουρκικών – Ελληνικών Γραμμάτων.

Η αντιστοίχηση των γραμμάτων δύο γλωσσών δεν είναι πάντα 1:1. Υπάρχουν περιπτώσεις που η αναλογία είναι διαφορετική (π.χ. 2:6 για τα Τουρκικά "ı" και "i" που αντιστοιχίζονται με τα «ι», «υ», «η», «οι», «ει», «υι»). Αν και η δημιουργία λίστας αντιστοίχησης γραμμάτων των δύο γλωσσών αποτελεί γλωσσική πληροφορία, είναι τόσο μικρή που μπορούμε να την θεωρήσουμε αμελητέα.

Το Ποσοστό της Μέγιστης Κοινής Υποακολουθίας (LCSR) είναι το πηλίκο του μήκους της Μέγιστης Κοινής Υποακολουθίας προς το μήκος της μεγαλύτερης (σε χαρακτήρες) από τις δύο λέξεις. Για παράδειγμα:

Ε	**υ**	**ρ**	ω	**π**	η
Α	**v**	**r**	u	**p**	a

$$\text{LCSR (Ευρώπη, Avrupa)} = \frac{\text{όμοιοι χαρακτήρες}}{\text{χαρακτήρες μεγαλύτερης λέξης}} = \frac{3}{6} = 0.5$$

Όσο πλησιέστερα στο 1 είναι το LCSR, τόσο μεγαλύτερη είναι η πιθανότητα οι δύο λέξεις να συσχετίζονται.

Η αξιολόγηση με αυτό το κριτήριο έχει δοκιμαστεί, σε συνδυασμό με άλλες, ως συμπληρωματική και όχι από μόνη της [17][14]. Οι περισσότερες δοκιμές όμως με τη χρήση αυτής της μετρικής έγιναν σε γλώσσες με κοινή καταγωγή (Αγγλοσαξονικές), οι οποίες έχουν παρόμοιο αλφάβητο, ενώ πολλές λέξεις τους έχουν κοινή ρίζα. Η χρήση αυτής της μεθόδου αξιολόγησης θα είχε ικανοποιητικά αποτελέσματα π.χ. στην αντιστοίχηση Ελληνικών και Αγγλικών ιατρικών όρων, κύριων ονομάτων (τοπωνύμια κ.λ.π.), αλλά θα είχε μάλλον φτωχά αποτελέσματα σε τόσο διαφορετικές γλώσσες όπως η Ελληνική με τη Τουρκική. Στο παραπάνω παράδειγμα βλέπουμε πως ο αλγόριθμος αυτός δεν καταφέρνει να εντοπίσει ομοιότητα μεταξύ δύο κυρίων ονομάτων προφανέστατα ίδιων μεταξύ τους.

2.3.2.2. Μέθοδοι βασισμένοι στην «Εκτίμηση»

2.3.2.2.1. Μοντέλα Πιθανοτήτων.

Ένας μεγάλος αριθμός μεθόδων αξιολόγησης ζευγών υποψήφιων μεταφράσεων βασίζεται σε εκτιμήσεις με μοντέλα πιθανοτήτων. Έτσι δεδομένης της εμφάνισης μιας λέξης x σε πρόταση της μιας γλώσσας, υπολογίζουμε την πιθανότητα η λέξη y της άλλης γλώσσας να εμφανιστεί στην αντίστοιχη πρόταση. Η πιθανότητα εμφάνισης της y δεδομένης της x είναι:

$$P(y|x) = \frac{P(x|y)\cdot P(y)}{P(x)}$$

Όταν επιχειρούμε να εντοπίσουμε την μετάφραση της x συγκρίνοντάς την με κάθε λέξη y της άλλης γλώσσας, η P(x) στον παραπάνω τύπο θα είναι πάντα η ίδια, οπότε αρκεί να βρούμε εκείνη τη λέξη y που μεγιστοποιεί το γινόμενο :

$$P(x|y)\cdot P(y)$$

2.3.2.2.2. Μοντέλα Γράφων.

Ο Gaussier [20] και οι Kageura κ.ά. [16] προτείνουν την κατασκευή ενός γράφου του οποίου οι κορυφές είναι οι λέξεις των δύο γλωσσών. Δύο κορυφές που σχηματίζονται από λέξεις διαφορετικής γλώσσας συνδέονται με ακμές μόνο αν σε προηγούμενη φάση έχουν χαρακτηριστεί ως υποψήφια μεταφραστικά ζεύγη [16].

Οι Kageura κ. ά. [16] θέτουν ως βάρη των ακμών του γράφου το πλήθος των κοινών εμφανίσεων των λέξεων που αποτελούν τις κορυφές τους, και σχηματίζουν αρχικά ένα γράφο που περιέχει όλες τις λέξεις των κειμένων (όχι απαραίτητα συνεκτικό γράφο). Σε επόμενο στάδιο αφαιρούνται οι ακμές που οφείλονται σε λάθος συσχετίσεις και οι οποίες θα οδηγήσουν σε «κόψιμο» (cut) του γράφου και σε δημιουργία μικρότερων υπογράφων. Ως λάθος ακμές επιλέγονται εκείνες για τις οποίες ισχύει τουλάχιστον ένα από τα παρακάτω:

- Έχουν βάρος < 3 (δηλαδή $f(x,y)<3$).
- Καμία από τις λέξεις – κορυφές δεν έχει συχνότητα > 3.
- Και οι δύο λέξεις – κορυφές έχουν περισσότερες από 1 ακμές.
- Το «κόψιμο» (cut) δεν θα δημιουργήσει υπογράφους χωρίς τουλάχιστον 1 λέξη με συχνότητα > 3.

Σημαντικό πλεονέκτημα του παραπάνω αλγόριθμου είναι το γεγονός ότι εκτός από τις αντιστοιχήσεις λέξεων 1:1, μπορεί να εντοπίζει και αντιστοιχίες 1:n ή n:m. Στις περιπτώσεις όπου μια λέξη μπορεί να αντιστοιχισθεί με πολλές λέξεις της άλλης γλώσσας, τα βάρη των ακμών καταδεικνύουν την καταλληλότερη αντιστοίχηση για το γνωστικό πεδίο κειμένου των εξεταζόμενων κειμένων (η ακμή που έχει με το μεγαλύτερο βάρος).

2.3.2.3. Σύνθετες Μέθοδοι Αντιστοίχησης

Οι σύνθετες μέθοδοι αξιολόγησης [17] αποτελούν ένα μίγμα των παραπάνω μεθόδων, όπου η βαθμολογία από κάθε μέθοδο έχει τη δική της βαρύτητα. Η τελική αξιολόγηση ενός ζεύγους λέξεων υπολογίζεται ως το άθροισμα όλων των διαφορετικών αξιολογήσεων πολλαπλασιασμένων με την αντίστοιχη βαρύτητα τους. Η βαρύτητα της κάθε μεθόδου αξιολόγησης δεν είναι προκαθορισμένη και μπορεί να ρυθμιστεί για συγκεκριμένα ζεύγη γλωσσών μέσω μεθόδων μηχανικής μάθησης. Από δοκιμές που έγιναν σε Σουηδικά - Αγγλικά και Σουηδικά - Γερμανικά [17] υπολογίστηκε ότι π.χ. η αξιολόγηση με βάση τη μεγαλύτερη κοινή υποακολουθία γραμμάτων πρέπει να έχει το μικρότερο συντελεστή βαρύτητας, αφού σπάνια επαρκεί για να εντοπίσει τη μετάφραση μιας λέξης.

2.4 Εντοπισμός και Αντιστοίχηση Πολυλεκτικών Όρων

Οι πολυλεκτικοί όροι είναι ακολουθίες από 2 έως 4 (πιθανώς μη συνεχόμενες) λέξεις ή και σπανιότερα μεγαλύτερου μεγέθους, οι οποίες αποτελούν ένα τόσο συμπαγές σύνολο ώστε συχνά μεταφράζονται σε άλλες γλώσσες ως μία λέξη ή ως ένα άλλο συμπαγές σύνολο λέξεων όπου η λέξη προς λέξη αντιστοίχηση των λέξεων των δύο συνόλων δεν ενδείκνυται. Το πρόβλημα μετάφρασης αυτών των όρων είναι διπλό, αφού πρέπει πρώτα να εντοπιστούν και σε δεύτερη φάση να μεταφραστούν.

Παράδειγμα από τα δεδομένα αξιολόγησης:
Ηνωμένες **Π**ολιτείες της **Α**μερικής ⇔ **ABD**
ABD = Amerşka Birleşik Devletleri
Στο Ελληνικό κείμενο η αναφορά στις Η.Π.Α. έγινε ολογράφως ενώ στο Τουρκικό με συντομογραφία.

Ο Kupiec [7] εισηγήθηκε ένα σύστημα το οποίο όμως απαιτεί πρώτα την επεξεργασία και επισημείωση των μερών του λόγου στο κείμενο (POS Tagging), ώστε να εντοπιστούν μετά τα ουσιαστικά που περιέχονται σ' αυτό. Κατόπιν τα ουσιαστικά αντιστοιχίζονται και αυτό αποτελεί τη βάση για τον εντοπισμό των πολυλεκτικών όρων.

Παρόμοια μέθοδο ακολούθησε και ο Van der Eijk [8], ο οποίος, ομοίως, θεωρεί ότι οι πολυλεκτικοί όροι αποτελούνται *μόνο* από ουσιαστικά

Ένα πρόβλημα που έχουν και οι δύο παραπάνω μέθοδοι είναι ότι αναζητούν πολυλεκτικούς όρους που αποτελούνται αποκλειστικά από ουσιαστικά [6], ενώ πολλές φορές οι πολυλεκτικοί όροι συμπεριλαμβάνουν επίθετα, αντωνυμίες και προθέσεις. Επιπλέον η επιλογή του συντελεστή βαρύτητας [8] ανάλογα με τη θέση των λέξεων στη πρόταση δεν είναι δόκιμη για γλώσσες με διαφορετική σύνταξη.

Μία ακόμα προσέγγιση στην αντιστοίχηση πολυλεκτικών όρων, με εμπορική χρήση, έγινε από τους Smadja, McKeown, Hatzivassiloglou [6]. Το σύστημά τους αποτελείται από 2 επιμέρους προγράμματα. Το πρώτο είναι το XTRACT, που εντοπίζει αρχικά τους πολυλεκτικούς όρους, χωρίς να περιορίζεται μόνο σε ουσιαστικά. Το δεύτερο είναι το *Champolion* που στην συνέχεια τους μεταφράζει.

Το κυριότερο μειονέκτημα αυτής της προσέγγισης είναι ότι το XTRACT για να λειτουργήσει καλύτερα βασίζεται σε άλλα εργαλεία, όπως επισημειωτή μερών του λόγου και συντακτικό αναλυτή, τα οποία δεν είναι διαθέσιμα για όλες τις γλώσσες. Τα σημαντικότερα από τα κριτήρια που έθεσαν [6] για τον εντοπισμό των πολυ-λεκτικών όρων είναι:

- Μέχρι 4 λέξεις να παρεμβάλλονται μεταξύ κάθε δύο λέξεων που σχηματίζουν τον πολυλεκτικό όρο.
- Οι λέξεις που αποτελούν τον πολυλεκτικό όρο να συνεμφανίζονται τουλάχιστον 5 φορές.

Η μέθοδος μετάφρασης των πολυλεκτικών όρων που χρησιμοποίησαν σε πρώτη φάση εντοπίζει τις προτάσεις που περιέχουν τον προς μετάφραση πολυλεκτικό όρο και μετά όλες τις λέξεις της γλώσσας Στόχου οι οποίες περιέχονται στις αντιστοιχισμένες προτάσεις και είναι πιθανό να συσχετίζονται με τον όρο αυτό. Η επόμενη φάση αποτελείται από πολλαπλά στάδια αξιολόγησης αρχικά των πιθανών 2-λεκτικών όρων, στην συνέχεια των 3-λεκτικών όρων κοκ, προσθέτοντας σε κάθε στάδιο μία επιπλέον λέξη. Οι νέοι όροι προκρίνονται στο επόμενο στάδιο αξιολόγησης μόνο αν επιτύχουν βαθμολογία μεγαλύτερη από κάποιο κατώφλι (αξιολόγηση με το συντελεστή Dice). Η διαδικασία τελειώνει όταν κανένας νέος πολυλεκτικός όρος δεν περάσει το κατώφλι[5]. Τότε επιλέγεται ως καλύτερη μετάφραση ο όρος εκείνος ο οποίος σε κάποιο από τα προηγούμενα στάδια πέτυχε την καλύτερη βαθμολογία σύμφωνα με το συντελεστή Dice. Το αποτέλεσμα μπορεί να είναι μονολεκτικός ή n-λεκτικός όρος.

Το XTRACT είχε αποτελεσματικότητα 80 % στον εντοπισμό των πολυλεκτικών όρων γεγονός που οδήγησε και το *Champollion* σε κάποια λάθη, δίνοντας κατά μέσο όρο αποτελεσματικότητα στη μετάφραση 73 % (78 % αν δεν ληφθούν υπόψη τα λάθη του XTRACT)[6].

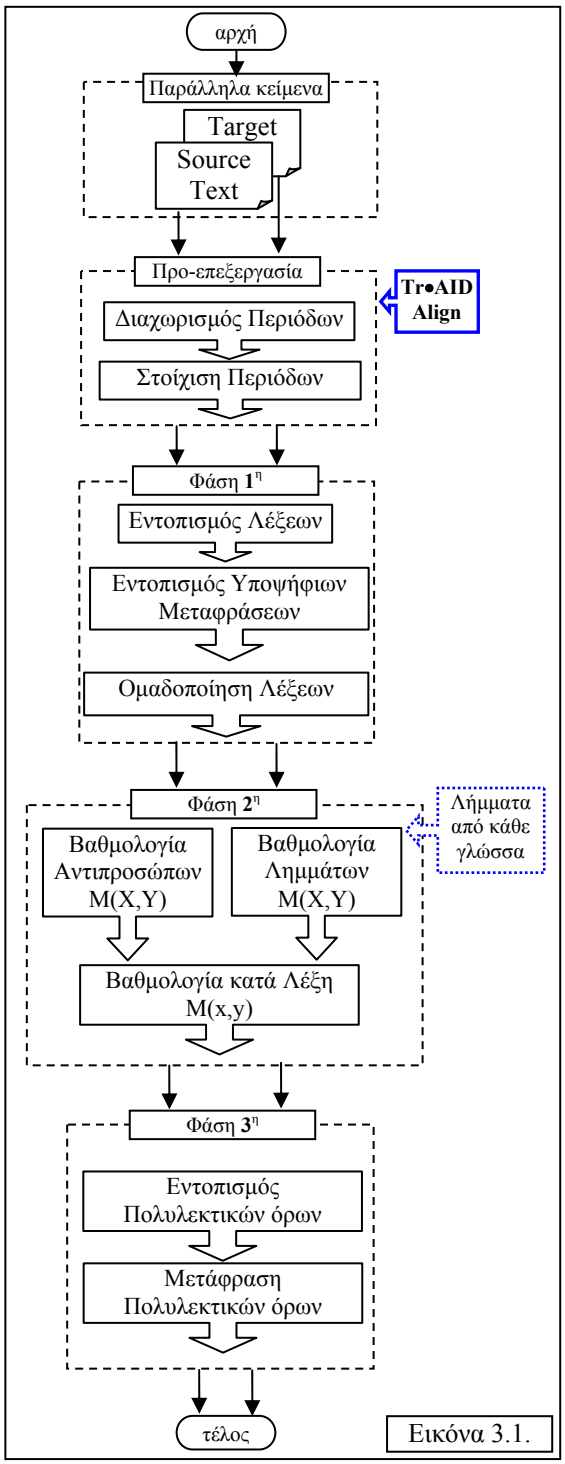
Κοινό χαρακτηριστικό των 3 μεθόδων που αναφέραμε είναι η απαραίτητη ενσωμάτωση γλωσσικής πληροφορίας σε αυτές με τη μορφή εργαλείων επισημείωσης μερών του λόγου, στοιχείο που προσπαθούμε να αποφύγουμε στο σύστημά μας.

[5] Η τιμή του κατωφλίου είναι εμπειρικά επιλεγμένη (0.1), χωρίς αυτό να σημαίνει ότι η τιμή αυτή είναι δεσμευτική (μπορεί να ποικίλει ανάλογα με τις εξεταζόμενες γλώσσες).

Κεφάλαιο 3ο

Αρχιτεκτονική Συστήματος

Η λειτουργία του συστήματός μας χωρίζεται σε τρεις κύριες φάσεις και υποβοηθείται από άλλη μια φάση προ-επεξεργασίας με την χρήση του **Tr•AID Align** [2a]. Η υλοποίηση του συστήματος έγινε σε Java.



Εικόνα 3.1.

3.1. Προ-Επεξεργασία.

3.1.1. Διαχωρισμός Προτάσεων

Το πρώτο πρόβλημα που πρέπει να αντιμετωπιστεί είναι ο διαχωρισμός των προτάσεων. Η λύση στο πρόβλημα αυτό δεν είναι τόσο απλή διότι το σύμβολο της τελείας (.) δεν αποτελεί πάντα και αλλαγή πρότασης (π.χ. « κ. Α. Ιωάννου, Α.Σ.Ο.Ε.Ε. , κ.λ.π.…»). Με τη βοήθεια κειμένων εκπαίδευσης που έδειχναν για κάθε μία τελεία αν αποτελεί αλλαγή προτάσεως ή όχι, παριστάνοντας κάθε τελεία ως διάνυσμα ιδιοτήτων και χρησιμοποιώντας τους παρακάτω αλγόριθμους μηχανικής μάθησης από τη βιβλιοθήκη WEKA [22]:

- τον αλγόριθμο `J4.8` [6], ο οποίος είναι μια υλοποίηση σε Java του C4.5,
- τη μέθοδο του αφελούς ταξινομητή Bayes (Naive Bayes) [7],
- τον αλγόριθμο των k κοντινότερων γειτόνων (`k-NN`) [8] με k=5 και βάρη αντίστροφα της απόστασης,

κατασκευάσαμε μία επιτροπή ταξινομητών (με μέλη τους ταξινομητές που προκύπτουν με τους τρεις αλγορίθμους μάθησης) που αποφασίζει για κάθε μία τελεία αν αποτελεί τέλος πρότασης ή όχι. Η απόφαση της επιτροπής λαμβάνεται πλειοψηφικά.[9]

[6] weka.classifiers.trees.J48
[7] weka.classifiers.bayes.NaiveBayesSimple
[8] weka.classifiers.lazy.IBk
[9] Η μονάδα διαχωρισμού προτάσεων βασίστηκε σε αντίστοιχη άσκηση του μαθήματος «Λογική και Τεχνητή Νοημοσύνη» του ΜΠΣ «Πληροφοριακά Συστήματα».

3.1.2. Στοίχιση Προτάσεων.

Στο επόμενο στάδιο γίνεται στοίχιση των δύο κειμένων βασιζόμενη στη στοίχιση των περιόδων τους. Για την στοίχιση των περιόδων των δύο κειμένων επιλέξαμε τον αλγόριθμο των Gale & Church (ενότητα 2.1.1). Για την εφαρμογή του αλγόριθμου στο εξεταζόμενο ζεύγος γλωσσών υπολογίσαμε τις τιμές των παραμέτρων c και s^2, οι οποίες για τα Ελληνικά και Τουρκικά είναι:

- c = 0,86 (δηλ. κάθε γράμμα της Ελληνικής αντιστοιχεί κατά μέσο όρο σε 0.86 γράμματα της Τουρκικής)
- $s^2 = 0,63704$

Όπως βλέπουμε, η τιμή του c για το ζεύγος Ελληνικών – Τουρκικών προσεγγίζει την τιμή που χρησιμοποίησαν στον αλγόριθμό τους οι Gale & Church.

Ένα πρόβλημα όμως στην εφαρμογή του παραπάνω αλγόριθμου είναι το ότι δεν προβλέπει αντιστοιχήσεις 2:0, 0:2, 3:1, 1:3, αφού δεν είχαν εντοπιστεί τέτοιες στα κείμενα Αγγλικών-Γαλλικών όπου εφαρμόστηκε ο αλγόριθμος. Αν και οι περιπτώσεις όπου δύο προτάσεις της μιας γλώσσας δεν αντιστοιχούν σε καμία της άλλης (2:0 και 0:2) οφείλονται μάλλον σε κακή μετάφραση, οι αναλογίες 1:3 και 3:1 ή ακόμα 3:2 και 2:3 δεν αποκλείονται στην περίπτωση Ελληνικών – Τουρκικών. Αυτό το πρόβλημα οδήγησε σε κάποια επιπλέον λάθη κατά την αξιολόγηση του αλγόριθμου Gale & Church που εφαρμόσαμε μέσω του προγράμματος **Tr•AID Align** [2a]. Τα αποτελέσματα ήταν ενθαρρυντικά για την εφαρμογή του αλγόριθμου σε αυτό το ζεύγος γλωσσών, αφού στοίχισε τις προτάσεις με πολύ καλά ποσοστά επιτυχίας:[10]

- 87 % σε κείμενα όπου το ένα είναι «χαλαρή» μετάφραση του άλλου[11].
- 100 % σε κείμενα όπου το ένα είναι πιστή μετάφραση του άλλου[12].

3.2. Φάση 1η

Στην 1η Φάση κατασκευάζουμε ένα ανεστραμμένο ευρετήριο όπου για κάθε διαφορετική λέξη σχηματίζουμε δύο λίστες: μια λίστα με τις προτάσεις στις οποίες εμφανίζεται η λέξη και μία λίστα με τις συν-εμφανιζόμενες λέξεις της άλλης γλώσσας. Για κάθε συνεμφανιζόμενη λέξη, η δεύτερη λίστα περιέχει και έναν μετρητή που δείχνει πόσες φορές έχουν συνεμφανιστεί οι δύο λέξεις.

3.2.1. Εντοπισμός Λέξεων

Στην 1η φάση εντοπίζονται αρχικά οι λέξεις των κειμένων, εξαιρώντας αριθμούς, ημερομηνίες, σημεία στίξης κλπ. Η διαδικασία αυτή είναι πολύ απλή και δεν θα την αναλύσουμε περισσότερο. Αγνοήθηκαν κατά τη διαδικασία του εντοπισμού λέξεων οι πολύ συχνές λέξεις και των δύο γλωσσών (άρθρα, σύνδεσμοι κλπ.), που δεν παρουσιάζουν ενδιαφέρον.

---

[10] Τα αποτελέσματα περιγράφονται αναλυτικότερα στην § 5.3.
[11] Η πρόκληση του 21ου Αιώνα (NATO) βλ. § 5.1.1.
[12] Σχέδιο ANAN (Comprehensive Settlement of the Cyprus Problem) βλ. § 5.1.1.

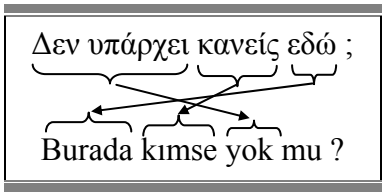
3.2.2. Εντοπισμός Υποψήφιων Μεταφράσεων

Στη συνέχεια κατασκευάζουμε ένα ανεστραμμένο ευρετήριο για όλες τις λέξεις της κάθε γλώσσας και παράλληλα δημιουργούμε μια λίστα από πιθανές μεταφράσεις της κάθε λέξεως. Υποψήφιες μεταφράσεις μιας λέξεως της γλώσσας πηγής είναι όλες οι λέξεις της γλώσσας στόχου οι οποίες εμφανίζονται στις αντιστοιχισμένες περιόδους των περιόδων όπου εμφανίζεται η προς μετάφραση λέξη. Έστω, για παράδειγμα, ότι η λέξη x της γλώσσας πηγής εμφανίζεται στην $3^η$ και $7^η$ πρόταση του κειμένου εκείνης της γλώσσας. Έστω, επίσης, ότι η $3^η$ πρόταση του κειμένου της γλώσσας πηγής αντιστοιχίζεται με την $4^η$ πρόταση του κειμένου της γλώσσας Στόχου, ενώ η $7^η$ πρόταση αντιστοιχίζεται με την $8^η$ και $9^η$ (περίπτωση 1:2). Τότε η x θα έχει ως *υποψήφιες* μεταφράσεις τις λέξεις y_1, y_2,...y_n της γλώσσας Στόχου που εμφανίζονται στην $4^η$, $8^η$ και $9^η$ πρόταση του κειμένου εκείνης της γλώσσας.

Το πρόβλημα όμως με τις λίστες αυτές, σε ένα μεγάλο κείμενο, είναι ότι γίνονται υπερβολικά μεγάλες και καταλαμβάνουν πολύ μεγάλο μέρος από τη διαθέσιμη μνήμη. Επίσης είναι δυνατό να συμπεριληφθούν σε αυτή τη λίστα και λέξεις που δεν είναι μεταφράσεις αλλά τυχαίνει να συνεμφανίζονται σε πολλές προτάσεις. Στη συνέχεια παρουσιάζουμε μία μέθοδο περιορισμού των υποψήφιων μεταφράσεων, αποκλείοντας εκείνες που είναι λιγότερο πιθανό να συσχετίζονται με την εξεταζόμενη λέξη.

Σημειώνουμε πως οι μέθοδοι περιορισμού υποψηφίων μεταφράσεων που βασίζονται στη σχετική θέση των λέξεων μέσα στις αντιστοιχισμένες περιόδους (ενότητα 2.3.1) δυστυχώς δεν μπορούν να εφαρμοστούν στην περίπτωση Ελληνικών – Τουρκικών, διότι οι δύο γλώσσες έχουν εντελώς διαφορετική σύνταξη, με αποτέλεσμα οι αντίστοιχες λέξεις να έχουν πολύ διαφορετικές θέσεις σε κάθε ζεύγος προτάσεων.

Στο παράδειγμα που ακολουθεί η σχετική θέση των Τουρκικών λέξεων είναι εντελώς ανεστραμμένη από αυτή των Ελληνικών λέξεων:



Σημειώνουμε, επίσης, ότι οι μέθοδοι που χρησιμοποίησαν οι Brown κ.ά. και ο Smadja (ενότητα 2.3.1) χρησιμοποιούν τις συχνότητες εμφάνισης των λέξεων σε όλο το κείμενο, ενώ εμείς (όπως περιγράφουμε παρακάτω) χρησιμοποιούμε τις συχνότητες των λέξεων στο τμήμα του κειμένου που έχουμε επεξεργαστεί μέχρι στιγμής. Το πλεονέκτημα της μεθόδου μας είναι η ταχύτερη δημιουργία της λίστας υποψήφιων μεταφράσεων και η εξοικονόμηση μνήμης.

3.2.2.1. Μαθηματικός Τύπος Αξιολόγησης

Για τον περιορισμό του προβλήματος της παρείσφρησης άσχετων λέξεων στη λίστα με τις υποψήφιες μεταφράσεις, εκτιμούμε την πιθανότητα να είναι η λέξη της γλώσσας στόχου μετάφραση μιας λέξεως της γλώσσας πηγής με τη χρήση του παρακάτω τύπου [3], [4]:

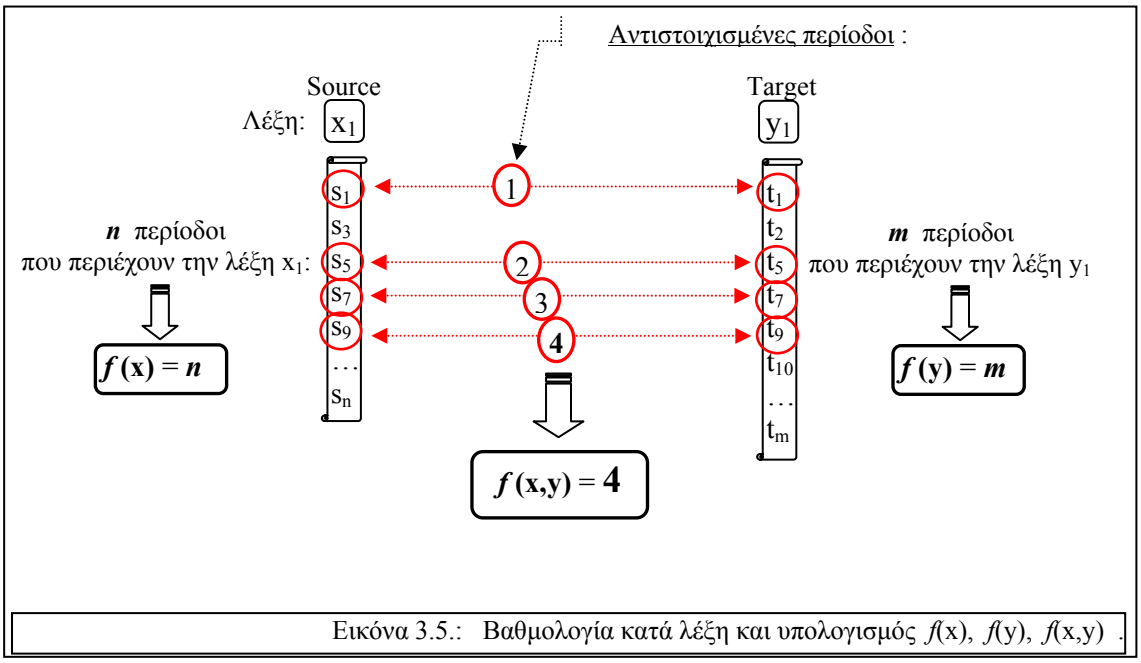
$$M(x,y) = \frac{\sqrt{(f(x)-m)^2 + (f(y)-m)^2 + (f(x,y)-m)^2}}{m} \qquad \textbf{(Εξίσωση 3.2)}$$

$$m = \frac{f(x) + f(y) + f(x,y)}{3}$$

f(x) = το πλήθος εμφανίσεων (η συχνότητα) της λέξης x στο κείμενο της γλώσσας πηγής.

f(y) = το πλήθος εμφανίσεων (η συχνότητα) της λέξης y στο κείμενο της γλώσσας στόχου.

f(x,y) = το πλήθος των κοινών εμφανίσεων της λέξης x με τη λέξη y σε αντιστοιχισμένες περιόδους.



Εικόνα 3.5.: Βαθμολογία κατά λέξη και υπολογισμός f(x), f(y), f(x,y) .

Αν η λέξη y είναι μετάφραση της x (και μάλιστα η μόνη μετάφραση), τότε θα πρέπει :

$$f(x) = f(y) = f(x,y)$$

δηλαδή θα πρέπει όχι μόνο να εμφανίζονται συνολικά τις ίδιες φορές στα αντίστοιχα κείμενα αλλά και να εμφανίζονται πάντα σε αντιστοιχισμένες περιόδους. Η σχέση (3.2) είναι συμμετρική, αφού πάντα M(x,y) = M(y,x), και μπορεί να χρησιμοποιηθεί για να εκτιμήσει αν η μετάφραση της y είναι η x.

Ένδειξη ότι
x μεταφράζεται σε **y**
και αντίστροφα, όταν:
M(x,y) → **0**

Όμως η πλήρης ταύτιση των συχνοτήτων δύο λέξεων θα συμβεί μόνο σε ιδανικές περιπτώσεις, καθώς ελεύθερες μεταφράσεις και εμφάνιση συνώνυμων λέξεων τείνουν να διαφοροποιούν την συχνότητα εμφάνισης λέξεων που η μία είναι μετάφραση της άλλης. Στην ιδανική περίπτωση όπου η μετάφραση της λέξεως x είναι μόνο η y, η τιμή του M(x,y) θα είναι ίση με 0. Γενικότερα συμπεραίνουμε ότι όσο πιο μικρή είναι η τιμή του M(x,y) (τείνει στο μηδέν) τόσο πιο πιθανό είναι η x να μεταφράζεται στην y αλλά και αντίστοιχα η y να μεταφράζεται σε x.

3.2.2.2. Όριο Αποκοπής

Κατά την 1η φάση, επεξεργαζόμαστε τα παράλληλα κείμενα σε επίπεδο αντιστοιχισμένων περιόδων, με σκοπό να κατασκευάσουμε τις λίστες υποψήφιων μεταφράσεων κάθε λέξεως x του πηγαίου κειμένου. Ταυτόχρονα ακολουθούμε την αντίστοιχη διαδικασία και για τη γλώσσα στόχο. Έτσι για να σχηματίσουμε την λίστα με τις υποψήφιες μεταφράσεις της λέξεως x της γλώσσας πηγής, υπολογίζουμε την τιμή του M(x,y) για τη x με κάθε λέξη y της γλώσσας στόχου. Συμπεριλαμβάνουμε στη λίστα των πιθανών μεταφράσεων της x μόνο εκείνες που έχουν αρκετά χαμηλή τιμή M(x,y). Ως συχνότητες f(x) f(y) και f(x,y) χρησιμοποιούμε τις συχνότητες των x και y στο τμήμα των παράλληλων κειμένων που έχουμε επεξεργαστεί ως εκείνη τη στιγμή και όχι τις συχνότητες στο συνολικό παράλληλο κείμενο.

Το πρόβλημα που αντιμετωπίζουμε είναι η επιλογή μίας τιμής του M(x,y) ως Όριο Αποκοπής (ΟΑ), που θα μας επιτρέψει να συμπεριλάβουμε όλες τις πιθανές μεταφράσεις αλλά και ταυτόχρονα όσο το δυνατό λιγότερες άσχετες λέξεις. Από δοκιμές που έγιναν υπολογίστηκε ότι η μετάφραση μιας

λέξεως έχει συνήθως M(x,y) ≤ 0.5. Αν και η τιμή αυτή φαίνεται αρκετά ικανοποιητική για όριο αποκοπής, στην πράξη είναι πολύ αυστηρή. Η αυστηρότητα του ΟΑ=0.5 εντοπίζεται όταν οι τιμές των *f*(x), *f*(y), *f*(x,y) είναι μικρές, όταν δηλαδή έχουμε λίγες εμφανίσεις των x και y. Στον πίνακα που ακολουθεί έχουμε υπολογίσει τις πρώτες 25 τιμές που είναι δυνατό να έχει το M(x,y):

A/A	***f*(x)**	***f*(y)**	***f*(x,y)**	**M(x,y)**	Dice(x,y)	Αποδεκτή
1	1	1	1	0.000000	**1**	ΝΑΙ
2	1	0	0	0.816497	0	ΝΑΙ
3	1	1	0	0.816497	0	ΝΑΙ
4	2	0	0	1.632993	0	ΝΑΙ
5	2	1	0	1.414214	0	ΝΑΙ
6	2	1	1	0.816497	0.666667	ΝΑΙ
7	2	2	0	1.632993	0	ΝΑΙ
8	2	2	1	0.816497	0.5	ΝΑΙ
9	3	0	0	2.449490	0	
10	3	1	0	2.160247	0	
11	3	1	1	1.632993	0.5	ΝΑΙ
12	3	2	0	2.160247	0	
13	3	2	1	1.414214	0.4	ΝΑΙ
14	3	2	2	0.816497	0.8	ΝΑΙ
15	3	3	0	2.449490	0	
16	3	3	1	1.632993	0.333333	ΝΑΙ
17	3	3	2	0.816497	0.666667	ΝΑΙ
18	4	1	0	2.943920	0	
19	4	2	0	2.828427	0	
20	4	1	1	2.449490	0.4	
21	4	2	1	2.160247	0.333333	
22	4	3	0	2.943920	0	
23	4	3	1	2.160247	0.285714	
24	4	3	2	1.414214	0.571429	ΝΑΙ
25	4	3	3	0.816497	0.857143	ΝΑΙ

Πίνακας 1.

Στην τελευταία στήλη και με το πράσινο χρώμα, σημειώνουμε τις περιπτώσεις οι οποίες *θα μπορούσαν* να αναφέρονται σε ζεύγος λέξεων όπου η μία είναι μετάφραση της άλλης. Για παράδειγμα, στη γραμμή 12, αν η λέξη x έχει εμφανιστεί 3 φορές και η λέξη y έχει εμφανιστεί 2 φορές αλλά δεν έχουν εμφανιστεί ποτέ και οι δύο μαζί σε αντιστοιχισμένες προτάσεις, τότε δεν είναι πιθανό να αποτελεί η μία μετάφραση της άλλης. Τονίζουμε ότι η πρώτη αυτή αξιολόγηση γίνεται για να περιορίσει το σύνολο των υποψήφιων μεταφράσεων και όχι για να υπολογίσει την μετάφραση της λέξεως x. Σε επόμενο βήμα θα εξηγήσουμε πώς επιλέγουμε την πιθανότερη μετάφραση από το σύνολο των υποψήφιων μεταφράσεων.

Με βάση τις περιπτώσεις του πίνακα 1, επιλέγουμε την τιμή του M(x,y) η οποία θα είναι το όριο αποκοπής για τη λίστα των υποψήφιων μεταφράσεων. Η επιλογή των περιπτώσεων αυτών έγινε με εμπειρικό τρόπο μετά από δοκιμές. Παρατηρούμε ότι για τις 25 πρώτες πιθανές περιπτώσεις όπου τα *f*(x) και *f*(y) έχουν μικρές τιμές (εξαιρώντας συμμετρικούς συνδυασμούς), το M(x,y) είναι δυνατό να λάβει μια από τις παρακάτω αποδεκτές τιμές, δηλαδή τιμές που αντιστοιχούν σε πιθανές υποψήφιες μεταφράσεις:

Όριο Αποκοπής Υποψήφιων Μεταφράσεων:
M(x,y)≤**1.632993**

$$M(x,y) \in \{0, 0.816497, 1.414214, 1.632993\}$$

από τις οποίες επιλέγουμε ως όριο αποκοπής τη μεγαλύτερη. Συνεπώς για κάθε λέξη x της γλώσσας πηγής θα συμπεριλαμβάνουμε στη λίστα των υποψήφιων μεταφράσεών της όποια λέξη y της γλώσσας στόχου έχει εμφανιστεί σε αντιστοιχισμένη πρόταση με M(x,y) ≤ **1.632993.**.

Ένα δεύτερο κριτήριο αποκοπής είναι και το ίδιο το μέγεθος της λίστας των υποψηφίων μεταφράσεων. Υπολογίζοντας ότι το μέσο πλήθος λέξεων μιας πρότασης είναι περίπου 20 – 30 λέξεις και λαμβάνοντας υπόψη ότι η μετάφραση της λέξεως x της γλώσσας πηγής θα συμπεριλαμβάνεται σχεδόν σίγουρα στις πρώτες 3 αντιστοιχισμένες προτάσεις της γλώσσας στόχου που θα εξετάσουμε, μπορούμε να θέσουμε ως όριο μεγέθους της λίστας το 100, δηλαδή να περιοριστούμε στις πρώτες 100 υποψήφιες μεταφράσεις της x που θα συναντήσουμε κατά την επεξεργασία των αντιστοιχισμένων προτάσεων της γλώσσας στόχου.

3.2.2.3. Σύγκριση Μαθηματικού τύπου M(x,y) με Dice Coefficient.

Όπως αναφέραμε στο Κεφάλαιο 2, μια ευρύτατα χρησιμοποιούμενη μετρική στην αντιστοίχηση λέξεων είναι ο συντελεστής Dice (Dice Coefficient). Η μετρική M(x,y) που χρησιμοποιήθηκε στην παρούσα εργασία [3], [4] φαίνεται να έχει μεγαλύτερη διακριτική ικανότητα από το συντελεστή Dice και το γεγονός αυτό φαίνεται στον Πίνακα 1. Στις περιπτώσεις {2, 3, 4, 5, 7} και {9, 10, 12, 15, 18, 19, 22} ο συντελεστής Dice παίρνει παντού τιμή 0, παρ' όλο που οι πρώτες περιπτώσεις θα θέλαμε να περιλαμβάνονται στις υποψήφιες μεταφράσεις ενώ οι δεύτερες όχι. Αντίθετα, η M(x,y) παίρνει μικρότερες τιμές στις πρώτες περιπτώσεις και μεγαλύτερες στις δεύτερες, επιτρέποντας το διαχωρισμό τους. Επίσης, η M(x,y) ορθώς δεν διαφοροποιεί τις περιπτώσεις 8 και 11, ενώ ο συντελεστής Dice τούς δίνει πολύ διαφορετικές τιμές. Από την άλλη πλευρά, ο συντελεστής Dice δίνει τις ίδιες τιμές στις περιπτώσεις 13 και 20, ενώ θα θέλαμε να διαφοροποιούνται, κάτι που επιτρέπουν οι διαφορετικές τιμές της M(x,y). Παρόμοιο παράδειγμα μεγαλύτερης διακριτικής ικανότητας της M(x,y) είναι οι περιπτώσεις 16 και 21, που θα θέλαμε να διαφοροποιούνται και όπου ο συντελεστής Dice δίνει τις ίδιες τιμές ενώ η M(x,y) όχι. Από τις 25 περιπτώσεις, οι μόνες τις οποίες ο συντελεστής Dice διαχώρισε καλύτερα από τη M(x,y) είναι οι 6 και 14.

3.2.3. Ομαδοποίηση Λέξεων

Μία από τις σημαντικότερες λειτουργίες της 1ης φάσης του συστήματος μας είναι η ομαδοποίηση λέξεων. Στον πίνακα που ακολουθεί έχουμε ένα χαρακτηριστικό παράδειγμα για την αναγκαιότητα αυτής της διαδικασίας, όπου η Αγγλική λέξη "play" μεταφράζεται σε 4 διαφορετικές προτάσεις των Ελληνικών με 4 διαφορετικές λέξεις:

Πηγή	f(x)	Στόχος	f(y)	f(x,y)	M(x,y)[13]
play	f(play) = 4	παίζω	f(παίζω) = 1	f(play, παίζω) = 1	2.449490
play		παίζεις	f(παίζεις) = 1	f(play, παίζεις) = 1	2.449490
play		παίξω	f(παίξω) = 1	f(play, παίξω) = 1	2.449490
play		παίξει	f(παίξει) = 1	f(play, παίξει) = 1	2.449490

Πίνακας 3.

Όπως βλέπουμε η λέξη "play" είναι πολύ πιθανό να μην αντιστοιχηθεί σε καμία από τις παραπάνω ελληνικές μεταφράσεις της, αφού η τιμή του M(x,y) για κάθε μια από αυτές είναι πολύ μεγάλη. Για να αντιμετωπιστεί το πρόβλημα θα πρέπει να γίνει μια ομαδοποίηση των λέξεων με την ίδια ρίζα. Σε κάθε σύνολο από ομαδοποιημένες μορφές, της ίδιας λέξεως, ορίζεται μια από αυτές ως αντιπρόσω-

[13] Η τιμή του M(x,y) αντιστοιχεί σε αυτή της γραμμής 20 του πίνακα 1.

πος του συνόλου. Για την επιλογή του αντιπροσώπου και την συνένωση υποσυνόλων, χρησιμοποιήθηκαν οι τεχνικές Union_by_Height και Path_Compression [18]. Δοκιμάσαμε δύο τρόπους ομαδοποίησης, έναν με γλωσσική πληροφορία και έναν χωρίς.

3.2.3.1. Ομαδοποίηση Λέξεων ΧΩΡΙΣ Γλωσσική Πληροφορία

Η ομαδοποίηση χωρίς γλωσσική πληροφορία βασίστηκε στην παρατήρηση ότι και στις δύο γλώσσες (όπως και στις περισσότερες) οι διαφορετικές μορφές των λέξεων σχηματίζονται τροποποιώντας την κατάληξη τους, ενώ το μεγαλύτερο μέρος τους (από αριστερά προς τα δεξιά), παραμένει αμετάβλητο. Έτσι για να αποφασίσουμε αν δύο λέξεις της ίδιας γλώσσας έχουν κοινή ρίζα, καταλήξαμε στον εξής αλγόριθμο:

Α. Από τις Ελληνικές λέξεις αφαιρούνται οι τόνοι.

Β. Ανάλογα με το μήκος σε γράμματα της κάθε μίας από τις δύο λέξεις, αγνοούμε έως k και k΄ γράμματα από το τέλος της κάθε μίας αντίστοιχα (όπου k, k΄ : ∈ {0,1,2,3,4,5})[14]. Το πλήθος των χαρακτήρων που αγνοούνται σε καμιά περίπτωση δεν πρέπει να ξεπερνάει το 40% του συνολικού μήκους της λέξεως.

Γ. Συγκρίνουμε τις νέες λέξεις που προκύπτουν (για τις διάφορες επιτρεπόμενες τιμές των k και k΄) και αν οποιεσδήποτε 2 από αυτές είναι ίδιες, τότε συμπεραίνουμε ότι οι δύο αρχικές λέξεις έχουν κοινή ρίζα.

Δ. Κάθε καινούργια λέξη που συναντάμε (για την οποία δεν έχουμε συναντήσει άλλη με την ίδια ρίζα) ορίζεται ως «Λέξη Αντιπρόσωπος» για όλες τις επόμενες που θα συναντήσουμε με την ίδια ρίζα.

Το πλήθος των γραμμάτων που αγνοούμε κατά την σύγκριση των 2 λέξεων είναι κατά περίπτωση:

Συνολικό μήκος λέξεως (σε χαρακτήρες)	Χαρακτήρες που αγνοούνται	% χαρακτήρων που αγνοούνται
μέχρι 2	0	0 %
3 ή 4	μέχρι 1	25% - 33.3%
5 ή 6 ή 7	μέχρι 2	28.5% - 40%
8 ή 9	μέχρι 3	33.3% - 37.5%
10 ή 11 ή 12	μέχρι 4	33.3% - 40%
περισσότερα από 12	μέχρι 5	< 38.46%

Πίνακας 4.

Η μέθοδος αυτή είναι αρκετά αποδοτική, χωρίς όμως να επιτυγχάνει απολύτως ορθά αποτελέσματα. Υπάρχουν περιπτώσεις όπου ο αλγόριθμος των N-Grams αποδίδει καλύτερα, αλλά και περιπτώσεις όπου ο αλγόριθμός μας δίνει πιο σωστά αποτελέσματα, ειδικότερα στις περιπτώσεις που οι δύο λέξεις διαφέρουν αρκετά στο πλήθος των χαρακτήρων τους. Ας θεωρήσουμε, για παράδειγμα, τις λέξεις «υποστήριξη» και «υποστηρίζεται»:

Λέξεις	2-Grams	#2-Grams	Μήκος (χαρακτήρες)
Υποστήριξη	**υπ πο οσ στ τη ηρ ρι** ιξ ξη	9	10
Υποστηρίζεται	**υπ πο οσ στ τη ηρ ρι** ιζ ζε ετ τα αι	12	13

[14] Βλ. πίνακα 4.

- Ομαδοποίηση κατά N-Grams (N=2): Υπάρχουν 7 κοινά 2-grams (C=7)

$$\text{Ομοιότητα } S = \frac{2 \cdot 7}{9+12} \approx 0.6667$$

Η τιμή του S δεν είναι αρκετά μεγάλη[15] ώστε να οδηγήσει σε συσχέτιση των 2 λέξεων.

- Ομαδοποίηση με τον αλγόριθμό μας: Σχηματίζουμε όλους τους συνδυασμούς από τις παραγόμενες (από τον αλγόριθμο μας) μορφές τους, αφαιρώντας από 0-5 χαρακτήρες για τη λέξη «υποστηρίζεται» (μήκος 13) και 0-4 για τη λέξη «υποστήριξη» (μήκος 10):

Χαρακτήρες που απορρίπτονται	«υποστηρίζεται»	«υποστήριξη»
0	υποστηριζεται	υποστηριξη
1	υποστηριζετα	υποστηριξ
2	υποστηριζετ	***υποστηρι***
3	υποστηριζε	υποστηρ
4	υποστηριζ	υποστη
5	***υποστηρι***	-

Πίνακας 4.1

Όπως σημειώνεται στον παραπάνω πίνακα, υπάρχουν δύο μορφές των λέξεων που παράγονται από τον αλγόριθμό μας οι οποίες ταιριάζουν απόλυτα, οπότε οι δύο αρχικές λέξεις θα ομαδοποιηθούν.

3.2.3.2. Ομαδοποίηση Λέξεων ΜΕ Γλωσσική Πληροφορία

Η ομαδοποίηση των λέξεων με χρήση γλωσσικής πληροφορίας έγινε χρησιμοποιώντας τα λήμματα όλων των λέξεων που συμπεριλαμβάνονται στα παράλληλα κείμενα. Το σύστημά μας δεν εκτελεί λημματοποίηση (lemmatization) λέξεων. Η εισαγωγή των λημμάτων των λέξεων γίνεται στο σύστημά μας με την μορφή επιπλέον πληροφορίας από δύο αρχεία που περιέχουν όλες τις λέξεις του κειμένου και τα λήμματα τους. Λέξεις που αποτελούν μορφές του ιδίου λήμματος ομαδοποιούνται και τα λήμματα αποτελούν τους αντιπροσώπους τους.

Το αρχείο με τα λήμματα του Ελληνικού κειμένου κατασκευάστηκε με το λημματοποιητή των Παπαγεωργίου κ.ά [2b]. Η λημματοποίηση των Τουρκικών λέξεων έγινε με εργαλεία που διαθέτει ο καθηγητής κ. Kemal Oflazer, με είσοδο ένα αρχείο που περιείχε μία λίστα με όλες τις λέξεις των Τουρκικών κειμένων (και όχι με επεξεργασία στα ίδια τα κείμενα). Τα προβλήματα που αντιμετωπίσαμε κατά την υλοποίηση της ομαδοποίησης ΜΕ γλωσσική πληροφορία αναλύονται στην § 5.4.

[15] Ένα ελαστικό κατώφλι επιλογής θα ήταν 0.7

3.3. Φάση $2^{η}$: Αντιστοίχηση (Μετάφραση) Λέξεων

Στη δεύτερη φάση επιλέγουμε ως μετάφραση μιας δεδομένης λέξης x την καλύτερη από όλες όσες περιέχονται στο σύνολο των υποψήφιων μεταφράσεών της. Η επιλογή της καλύτερης βασίζεται σε μια διαδικασία βαθμολόγησης των λέξεων της άλλης γλώσσας, με μεθόδους που θα παρουσιάσουμε στις επόμενες παραγράφους.

3.3.1. Βαθμολογία κατά Λέξη M(x,y)

Με τον όρο «Βαθμολογία κατά Λέξη» εννοούμε την εφαρμογή της μετρικής σχέσεως 3.2 M(x,y), όπου x η λέξη της γλώσσας πηγής και y η υποψήφια μετάφρασή της στη γλώσσα στόχο. Κατά τον υπολογισμό της «βαθμολογίας κατά λέξη» δεν λαμβάνονται υπόψη τα αποτελέσματα της ομαδοποίησης (αξιολογούνται δηλαδή ξεχωριστά διαφορετικές μορφές των ιδίων λέξεων). Οι υποψήφιες μεταφράσεις y που θα αξιολογηθούν είναι όσες έχουν συνεμφανιστεί σε αντιστοιχισμένες προτάσεις με την x και έχουν περάσει επιτυχώς το Όριο Αποκοπής της επιλογής υποψηφίων μεταφράσεων (ενότητα 3.2.2)[16].

3.3.2. Βαθμολογία κατά Αντιπρόσωπο : M(X,Y)

Η «Βαθμολογία κατά Αντιπρόσωπο» υπολογίζεται με τη χρήση της μαθηματικής σχέσης M(x,y) όπως παραπάνω, με τη διαφορά ότι τώρα όλες οι λέξεις αντικαθίστανται από τους αντιπροσώπους τους. Χρησιμοποιούμε τα $f(X)$, $f(Y)$, $f(X,Y)$ για να αναφερθούμε στις αντίστοιχες συχνότητες όταν χρησιμοποιούμε τους αντιπροσώπους των λέξεων και το M(X,Y) για την αντίστοιχη τιμή του μέτρου Στο παράδειγμα και το σχήμα που ακολουθεί θα υπολογίσουμε την «Βαθμολογία κατά Αντιπρόσωπο» της λέξεως x_1 από την γλώσσα πηγή:

Α. Εντοπίζεται ο αντιπρόσωπος της x_1 : η λέξη $X = x_0$.

Β. Εντοπίζονται όλες οι λέξεις $x_1, x_2, \dots, x_n$ που αντιπροσωπεύονται από την X.

Γ. Σε κάθε μία από τις λέξεις αυτές αντιστοιχεί μία λίστα S_i με τις περιόδους που την περιέχουν. Καλούμε Sx την ένωση όλων αυτών των λιστών:

$$S_X = S_1 \cup S_2 \cup \dots \cup S_n$$

Δ. Το πλήθος των περιόδων της S_X ισούται με τη συχνότητα του αντιπροσώπου X της x_1, δηλαδή είναι το $f(X)$, για το οποίο θα ισχύει $f(X) \geq f(x_1)$.

Ε. Από τις λέξεις $x_1, x_2, \dots, x_n$ που αντιπροσωπεύονται από τη X, βρίσκουμε το υπερσύνολο:

$$WC_X = WC_{x1} \cup WC_{x2} \cup \dots \cup WC_{xn}$$

όπου το κάθε WC_{xi} περιέχουν τις λέξεις της γλώσσας στόχου που εμφανίζονται σε περιόδους αντίστοιχες με αυτές που περιέχουν τη x_i.

ΣΤ. Για κάθε λέξη y_i που περιέχεται στο WC_X εντοπίζουμε την αντιπρόσωπό της, την $Y = y_0$.

[16] Το μέτρο M(x,y) χρησιμοποιείται και κατά την επιλογή των υποψηφίων μεταφράσεων (§3.2.2.), με τη διαφορά ότι σε εκείνο το στάδιο τα $f(x)$, $f(y)$, $f(x,y)$ υπολογίζονται βάσει μόνο των συχνοτήτων των λέξεων στο τμήμα των παράλληλων κειμένων που έχουμε επεξεργαστεί, ενώ εδώ χρησιμοποιούνται οι συχνότητες σε ολόκληρα τα κείμενα.

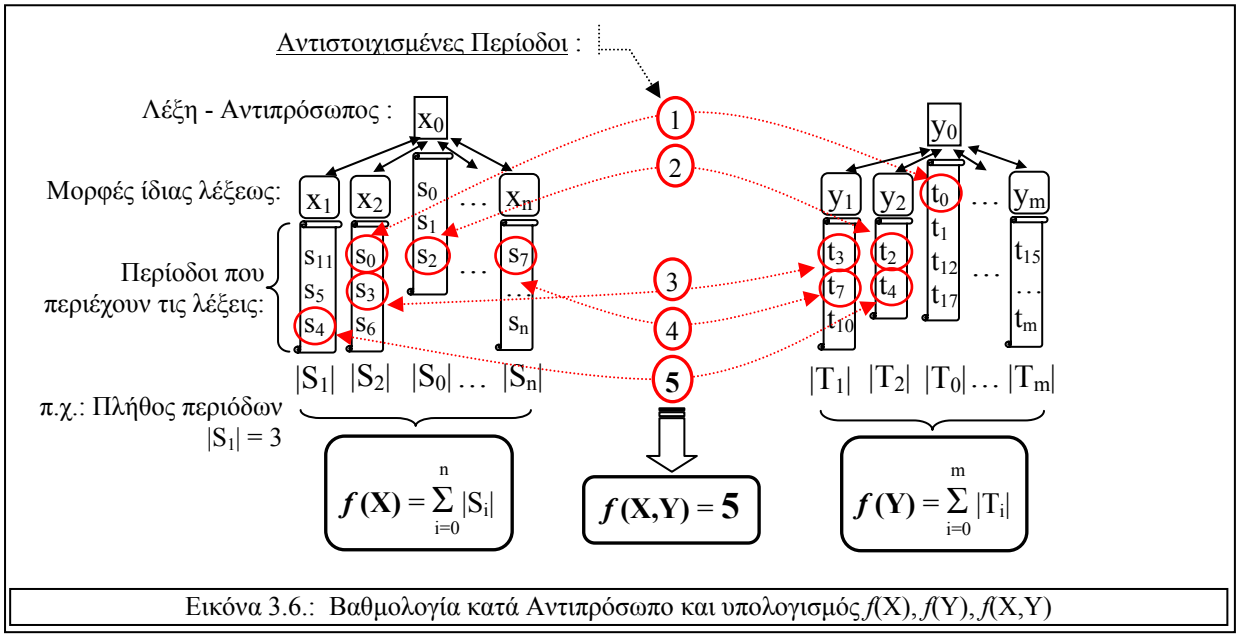
Ζ. Κατόπιν βρίσκουμε όλες τις λέξεις y_1, y_2 … y_m που αντιπροσωπεύονται από την Υ και τις περιόδους στις οποίες αυτές περιέχονται, τις οποίες εισάγουμε στις λίστες T_1, T_2, …, T_m.

Η. Το άθροισμα των πληθάριθμων των λιστών T_1, T_2, …, T_m ισούται με τη συχνότητα του αντιπρόσωπου Υ της y_i , δηλαδή είναι το $f(Y)$.

Θ. Το πλήθος των κοινών περιόδων που εμφανίζονται στα S_X και $T_Y = T_1 \cup T_2 \cup \ldots \cup T_m$ ισούται με τη συχνότητα συνεμφανίσεων λέξεων που αντιπροσωπεύονται από τις Χ και Υ:

$$f(X,Y) = |S_X \cap T_Y|$$

Ι. Υπολογίζουμε την «Βαθμολογία κατά Αντιπρόσωπο» Μ(Χ,Υ) χρησιμοποιώντας τη σχέση 3.2, αντικαθιστώντας τα $f(x)$, $f(y)$, $f(x,y)$ με τα $f(X)$, $f(Y)$, $f(X,Y)$ αντίστοιχα.



Εικόνα 3.6.: Βαθμολογία κατά Αντιπρόσωπο και υπολογισμός $f(X)$, $f(Y)$, $f(X,Y)$

3.3.3. Εντοπισμός Αντίστοιχης Λέξεως

Χρησιμοποιώντας τις βαθμολογίες «κατά Λέξη» και «κατά Αντιπρόσωπο», το σύστημά μας αντιστοιχεί τις λέξεις των δύο γλωσσών με την εξής διαδικασία:

Α. Υπολογίζουμε όλες τις «βαθμολογίες κατά Αντιπρόσωπο» μεταξύ των λέξεων που έχουν τον ίδιο αντιπρόσωπο με την εξεταζόμενη λέξη και των λέξεων που εμφανίζονται σε αντιστοιχισμένες περιόδους με αυτές.

Β. Υπολογίζουμε όλες τις «βαθμολογίες κατά Λέξη» μεταξύ της εξεταζόμενης λέξεως και των λέξεων που εμφανίζονται σε αντιστοιχισμένες περιόδους με αυτή και ανήκουν στο σύνολο των Υποψήφιων Μεταφράσεών της (ΥΜ)[17].

[17] Βλ. §3.2.2.

Γ. Από τις λέξεις-Αντιπροσώπους της γλώσσας στόχου, που εξετάσαμε στο βήμα ‘Α’, ορίζουμε ένα σύνολο Υποψήφιων Αντιπροσώπων (ΥΑ) στο οποίο θα ανήκουν εκείνη ή όλες εκείνες με την μικρότερη βαθμολογία κατά αντιπρόσωπο (π.χ. έστω τις αντιπρόσωπους Y, Z).

Δ. Από τις λέξεις που ανήκουν στο σύνολο των υποψήφιων μεταφράσεων της εξεταζόμενης λέξης, επιλέγουμε ως καταλληλότερη μετάφραση εκείνη την λέξη η οποία έχει αντιπρόσωπο μία από τις αντιπροσώπους του συνόλου ΥΑ (π.χ. μία από τις αντιπροσώπους Y ή Z) και έχει τη μικρότερη «βαθμολογία κατά λέξη» από τις υπόλοιπες λέξεις που ο αντιπρόσωπός τους ανήκει στο σύνολο ΥΑ. Στο βήμα αυτό είναι δυνατό να επιλυθούν και ισοβαθμίες οι οποίες μπορεί να προέκυψαν στο προηγούμενο βήμα της βαθμολογίας κατά Αντιπρόσωπο.

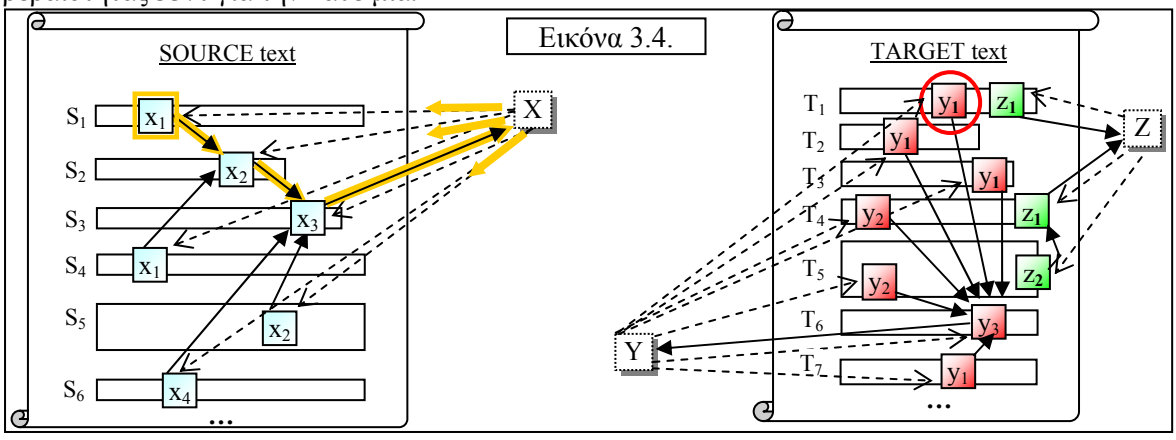
Ε. Επειδή το σύνολο των αντιπροσώπων που εξετάζουμε αποτελεί μια διεύρυνση του συνόλου των υποψήφιων μεταφράσεων της λέξεως προς μετάφραση, είναι δυνατό καμία λέξη από τις υποψήφιες μεταφράσεις της x να μην έχει αντιπρόσωπο στο σύνολο ΥΑ. Σε αυτή την περίπτωση επιλέγουμε ως πιθανότερες μεταφράσεις:

- όλες τις λέξεις-αντιπροσώπους που ανήκουν στο σύνολο ΥΑ καθώς και
- τυχόν λέξεις που είχαν βαθμολογία κατά λέξη = 0.

ΣΤ. Ένα πρόβλημα που αντιμετωπίσαμε αρκετά συχνά είναι οι λέξεις με πολλές υποψήφιες μεταφράσεις που ισοβαθμούν στη 1η θέση μετά το βήμα ‘Δ’ ή ‘Ε’. Το φαινόμενο αυτό οφείλεται στο ότι τα παράλληλα κείμενα δεν έχουν αρκετά μεγάλο δείγμα από αυτές τις λέξεις, ώστε να αντιστοιχισθούν σωστά. Η αντιμετώπιση του προβλήματος αυτού είναι πολύ δύσκολη και ο μόνος τρόπος για να περιορίσουμε αυτά τα φαινόμενα είναι να επιλέγουμε όσο το δυνατό μεγαλύτερα παράλληλα κείμενα. Ωστόσο είναι δυνατό να «προειδοποιήσουμε» για την αβεβαιότητα του αποτελέσματός μας, δίνοντάς τη με μορφή ποσοστού επί τοις εκατό (%). Αν δηλαδή υπάρχουν ισοβαθμίες, θα θεωρηθούν περισσότερα από 1 ορθά αποτελέσματα με ομοιόμορφα κατανεμημένο ποσοστό βεβαιότητας (π.χ. αν ισοβαθμήσουν ως πιθανές μεταφράσεις 2 λέξεις της γλώσσας στόχου, τότε κάθε μία από αυτές θα χαρακτηριστεί από το σύστημα ως πιθανή μετάφραση με βεβαιότητα 50%). Έτσι, αν π.χ. για την λέξη «άκρως» (=highly) που εμφανίζεται μόνο 1 φορά και δεν ομαδοποιείται με καμία άλλη λέξη του κειμένου, τα αποτελέσματα από τις βαθμολογίες κάθε μεθόδου είναι τα εξής:

```
10: "άκρως"     (x 1) / 1
--------------------
 | βαθμολογία κατά Λέξη    | βαθμολογία κατά Αντιπρόσωπο
1| economy     :0          | crippled :0                |
2| crippled    :0          | highly   :0                |
3| highly      :0          | dependent:0                |
4| dependent   :0          | aftermath:0,61237          |
5| aftermath   :0,61237    | second   :0,9798           |
```

οι υποψήφιες μεταφράσεις «crippled», «highly», «dependent» έχουν τις ίδιες πιθανότητες να αποτελούν μεταφράσεις της, αφού ισοβαθμούν στην βαθμολογία κατά αντιπρόσωπο και κατά λέξη. Το αποτέλεσμα θα είναι το σύστημα να αντιστοιχήσει τη λέξη «άκρως» και στις τρεις αγγλικές λέξεις, με βαθμο βεβαιότητας 33% για την κάθε μία.



Εικόνα 3.4.

3.4. Αντιστοίχηση Πολυλεκτικών Όρων.

Η μετάφραση πολυλεκτικών όρων αποτελεί ένα δύσκολο πρόβλημα το οποίο πολλές φορές δεν αντιμετωπίζεται ούτε από τα υπάρχοντα έντυπα λεξικά. Η δυσκολία έγκειται κυρίως στο γεγονός ότι οι λέξεις που αποτελούν τους πολυλεκτικούς όρους είναι δυνατό να μεταφράζονται σε μόνο μια λέξη της άλλης γλώσσας ή ακόμα και σε πολλές, οι οποίες όμως δεν αποτελούν ακριβείς μεταφράσεις των λέξεων που αποτελούν τον πολυλεκτικό όρο στην άλλη γλώσσα.

3.4.1. Εντοπισμός των Διλεκτικών Όρων.

Η βασική διαφορά της μεθόδου μας από εκείνη των Smadja κ.ά. [6] είναι ότι οι 2-λεκτικοί όροι εντοπίζονται αυτόματα από το σύστημα χωρίς καμιά γλωσσική πληροφορία. Ο εντοπισμός τους επιτυγχάνεται με τη χρήση της μετρικής M(x,y) (ενότητα 3.4.1). Σε αυτή τη φάση, όμως, και οι δύο λέξεις x,y ανήκουν στην ίδια γλώσσα και πιο συγκεκριμένα στην ίδια πρόταση. Η λέξη y μπορεί να είναι οποιαδήποτε λέξη σε απόσταση έως ±2 λέξεις από την x. Η M(x,y) υπολογίζεται ως εξής:

- *f*(x) = το πλήθος των προτάσεων που περιέχουν την λέξη x της γλώσσας L.
- *f*(y) = το πλήθος των προτάσεων που περιέχουν την λέξη y *της ίδιας* γλώσσας (L).
- *f*(x,y) = το πλήθος των προτάσεων που περιέχουν και τις 2 λέξεις.

Τονίζεται ότι με τη μέθοδο που χρησιμοποιούμε είναι δυνατό να εντοπίσουμε 2-λεκτικούς όρους οι οποίοι δεν είναι κατ' ανάγκη συνεχόμενοι αλλά παρεμβάλλονται μεταξύ τους έως και 2 λέξεις. Για την εξάλειψη του «θορύβου» που προέρχεται από λέξεις που τυχαία συνυπάρχουν, εξαιρούμε όλες τις συνεμφανίσεις που συμβαίνουν μόνο 1 φορά. Οι πραγματικοί 2-λεκτικοί όροι που αποκλείονται με αυτό το κριτήριο είναι κατά πολύ λιγότεροι από τις άσχετες συνεμφανίσεις που αποκλείονται. Ο αλγόριθμός μας είναι ο εξής:

Α. Για κάθε λέξη **x** της συγκεκριμένης γλώσσας βρίσκουμε τις περιόδους στις οποίες εμφανίζεται, από την τηρούμενη για κάθε λέξη λίστα (ενότητα 3.2).

Β. Από τις περιόδους που εντοπίσαμε στο 1^{ο} βήμα, δημιουργείται ένα σύνολο από υποψήφιες λέξεις **y** που περιέχονται στις περιόδους αυτές. Οι λέξεις αυτές είναι πιθανό να συνδυάζονται με τη x για να σχηματίσουν ένα διλεκτικό όρο. Στο βήμα αυτό εξαιρούμε από τον έλεγχο τις λέξεις εκείνες που έχουν τον ίδιο αντιπρόσωπο με την προς εξέταση λέξη, καθώς και τις συχνές λέξεις της αντίστοιχης γλώσσας [18].

Γ. Αν X, Y είναι οι λέξεις αντιπρόσωποι των x, y αντίστοιχα, τότε για κάθε λέξη y υπολογίζουμε την μετρική M(x,y) και την M(X,Y). Για τον υπολογισμό του M(X,Y) (βαθμολογία του αντιπροσώπου της λέξεως x (X) με τον αντιπρόσωπο της y (Y)) χρησιμοποιούμε τις τιμές των :

- *f*(X) = το πλήθος των προτάσεων που περιέχουν λέξη που έχει ίδιο αντιπρόσωπο (X) με την x της γλώσσας L.
- *f*(Y) = το πλήθος των προτάσεων που περιέχουν λέξη που έχει ίδιο αντιπρόσωπο (Y) με την y *της ίδιας* γλώσσας (L).
- *f*(X,Y) = το πλήθος των προτάσεων που περιέχουν λέξεις με αντιπρόσωπο X και Y.

[18] Οι συχνές λέξεις που εξαιρούμε από τον έλεγχο έχουν αναφερθεί στην §3.4

Δ. Τα ζεύγη των λέξεων τα οποία:

- επιτυγχάνουν βαθμολογία M(x,y) μικρότερη από το εμπειρικά υπολογισμένο «Όριο 2-λεκτικών όρων» (O_2)
- ή βαθμολογία M(X,Y) μικρότερη από ένα άλλο εμπειρικά υπολογισμένο «Όριο Αντιπροσώπων 2-λεκτικών όρων» (OA_2)
- και τα οποία εμφανίζονται μαζί περισσότερες από ν φορές,

αποτελούν πιθανότατα 2-λεκτικούς όρους. Η τιμή που υπολογίσαμε μετά από πειράματα για το O_2 είναι $0.3 = OA_2$. Ο λόγος για τον οποίο λαμβάνουμε υπόψη και την βαθμολογία των αυτούσιων λέξεων (εκτός από αυτή των αντιπροσώπων τους) είναι ότι οι 2-λεκτικοί όροι συνήθως εμφανίζονται με σταθερή μορφή και σπάνια με παραλλαγές των λέξεών τους. Το όριο που θέσαμε για το ελάχιστο πλήθος εμφανίσεων του 2-λεκτικού όρου είναι ν=3, αν και ο Smaja [6] επέλεξε ως όριο ν=5. Το κριτήριο επιλογής για το αν οι λέξεις x, y αποτελούν 2-λεκτικό όρο είναι:

$$\left(M(x,y) \le O_2 \ \textbf{OR}\ M(X,Y) \le OA_2 \right) \ \textbf{AND}\ \left(M(x,y) \le L_{max} \ \textbf{AND}\ M(X,Y) \le L_{max} \right)$$

Όρια Επιλογής 2-λεκτικών Όρων:
$O_2 = OA_2 = \mathbf{0.3}$
Όρια Επιλογής πολυλεκτικών Όρων:
$O_n = \mathbf{0.4}$
$OA_n = \mathbf{0.3}$
Μέγιστη Επιτρεπτή Τιμή
$L_{max} = 1.5$
Όριο πλήθους Εμφανίσεων
$\nu \ge 3$

Επειδή το όριο επιλογής γίνεται ελαστικό λόγω του "OR" στη πρώτη παρένθεση, επιβάλαμε ένα επιπλέον περιορισμό θέτοντας ως Μέγιστη Επιτρεπτή Τιμή των M(x,y) και M(x,Y) το $L_{max} = 1.5$.

Ε. Η σχετική θέση των όρων που αποτελούν τους 2-λεκτικούς όρους αποφασίζεται με βάση την σχετική τους θέση στη πρόταση που εμφανίζονται.

3.4.2. Εντοπισμός Όρων με Περισσότερες από 2 Λέξεις

Για να εντοπίσουμε στο κείμενο μιας γλώσσας όρους αποτελούμενους από περισσότερες από 2 λέξεις, στηριχθήκαμε στο προηγούμενο βήμα όπου εντοπίσαμε τους 2-λεκτικούς όρους. Αν x κάποιος πολυλεκτικός όρος, είναι βέβαιο ότι θα αποτελείται από 2 + n λέξεις. Με την μέθοδο που περιγράψαμε στην προηγούμενη παράγραφο, οι 2 πρώτες λέξεις ή κάποιες άλλες εντός του πολύ-λεκτικού όρου θα έχουν εντοπιστεί και ως 2-λεκτικός όρος. Έτσι, βαθμολογούμε χρησιμοποιώντας την ίδια μαθηματική σχέση τη συσχέτιση του 2-λεκτικού όρου με κάθε λέξη των προτάσεων όπου εμφανίζεται ο 2-λεκτικός όρος, περιοριζόμενοι σε λέξεις που απέχουν έως ±2 λέξεις από το 2-λεκτικό όρο. Με τον τρόπο αυτό εντοπίζουμε 3-λεκτικούς όρους και ομοίως, επαναλαμβάνοντας, εντοπίζουμε 4-λεκτικούς όρους κ.ο.κ.

Η διαφοροποίηση στις μεταβλητές της μαθηματικής σχέσεως της M(x,y) είναι η εξής:

- f(x) = το πλήθος των προτάσεων που περιέχουν τον εξεταζόμενο 2-λεκτικό / πολυλεκτικό όρο x της γλώσσας L.
- f(y) = το πλήθος των προτάσεων που περιέχουν την λέξη y *της ίδιας* γλώσσας (L).
- f(x,y) = το πλήθος των κοινών εμφανίσεων: του εξεταζόμενου 2-λεκτικού / πολυλεκτικού όρου x και της λέξεως y.

Επιπλέον χρησιμοποιούμε μια άλλη παραλλαγή της σχέσεως 3.2, την M(x,Y), η οποία υπολογίζεται ως εξής:

- f(x) = το πλήθος των προτάσεων που περιέχουν τον εξεταζόμενο 2-λεκτικό / πολυλεκτικό όρο x της γλώσσας L.
- f(Y) = το πλήθος των προτάσεων που περιέχουν ***αντιπρόσωπο (Y)*** της λέξης y *της ίδιας* γλώσσας (L).
- f(x,Y) = το πλήθος των κοινών εμφανίσεων: του εξεταζόμενου 2-λεκτικού / πολυλεκτικού όρου x και της λέξεως y.

Παρακάτω περιγράφουμε τον αλγόριθμο εντοπισμού των πολυλεκτικών όρων που χρησιμοποιήσαμε.

Α. Για κάθε 2-λεκτικό όρο που εντοπίσαμε, ανακαλούμε τις προτάσεις s_1, s_2, … s_n που τον περιέχουν.

Β. Από τις παραπάνω προτάσεις ανακτούμε όλες τις λέξεις y που περιέχονται σ' αυτές και που απέχουν έως ±2 λέξεις από το 2-λεκτικό όρο.

Γ. Αξιολογούμε την πιθανότητα ο 2-λεκτικός όρος x να συνδέεται με κάθε λέξη y χρησιμοποιώντας την εξίσωση 3.2. Αν η τιμή της M(x,y) είναι μικρότερη από το όριο αποκοπής πολυλεκτικών όρων (O_n), τότε είναι πολύ πιθανό ο 2-λεκτικός όρος μαζί με την y να αποτελούν πολύ-λεκτικό όρο. Από δοκιμές που έγιναν θεσπίστηκε ως όριο αποκοπής των πολυλεκτικών όρων $O_n = 0.4$. Παράλληλα υπολογίζεται και η M(x,Y) όπου Y είναι ο αντιπρόσωπος της y και εξετάζουμε τη τιμή της σε σχέση με ένα όριο αντιπροσώπων (OA_n, με τιμή 0.3). Η εξεταζόμενη λέξη y προστίθεται στον πολυλεκτικό όρο αν :

$$\left(M(x,y) \le O_n \ \textbf{OR}\ M(x,Y) \le OA_n \right) \ \textbf{AND}$$
$$\left(M(x,y) \le L_{max} \ \textbf{AND}\ M(x,Y) \le L_{max} \right) \ \textbf{AND}$$
$$\left(f(x,y) \ge 3 \ \textbf{OR}\ f(x,Y) \ge 3 \right)$$

Δ. Στη συνέχεια, τα βήματα Α, Β, Γ, επαναλαμβάνονται με τη διαφορά ότι αντί για 2-λεκτικούς χρησιμοποιούμε τους 3-λεκτικούς όρους που εντοπίσαμε, προσπαθώντας να τους προσθέσουμε ακόμα 1 λέξη για να βρούμε τους 4-λεκτικούς όρους. Η διαδικασία αυτή θα ολοκληρωθεί όταν θα έχουν εξεταστεί όλοι οι νέοι πολυλεκτικοί όροι που θα έχουν προκύψει στο τέλος κάθε φάσης εντοπισμού n-λεκτικών όρων.

Ε. Μετά την ολοκλήρωση εντοπισμού των n-λεκτικών όρων, διαγράφουμε τους πολυλεκτικούς όρους οι οποίοι είναι πιθανό να αποτελούν θόρυβο. Οι όροι που θα διαγραφούν είναι όσοι πληρούν τουλάχιστον ένα από τα παρακάτω κριτήρια:

- Έχουν εμφανιστεί λιγότερες από **v** φορές.
- Όλες οι λέξεις τους περιέχονται στις λέξεις ενός μεγαλύτερου πολυλεκτικού όρου.

3.4.3. Μετάφραση των Πολυλεκτικών Όρων

Αφού εντοπιστούν οι πολύ-λεκτικοί όροι, αναζητάμε τη μετάφρασή τους στις αντιστοιχισμένες περιόδους της άλλης γλώσσας με την εξής απλή μέθοδο, όπου x ο προς μετάφραση πολυλεκτικός όρος:

- Αν η αντίστοιχη περίοδος της άλλης γλώσσας έχει πολυλεκτικούς όρους $\{Z_1, Z_2, \dots Z_k\}$[19] , πρώτα υπολογίζουμε τα $M(x,Z_i)$.

- Κατόπιν, για κάθε λέξη y της άλλη γλώσσας που εμφανίζεται στις αντιστοιχισμένες προτάσεις, υπολογίζουμε τα M(x,y) και M(x,Y), όπου Y ο αντιπρόσωπος της y.

- Η μετάφραση του πολύ-λεκτικού όρου x είναι η λέξη y ή ο πολυλεκτικός όρος Z_i με την μικρότερη βαθμολογία $M(x, Z_i)$ ή M(x,y) ή M(x,Y).

[19] Ο πολύ-λεκτικός όρος Z_1 αποτελείται από περισσότερες από μία λέξεις $\{ z_{11}, z_{12}, \dots z_{1n}\}$.

Κεφάλαιο 4ο

Αξιολόγηση Συστήματος

4.1. Δεδομένα Αξιολόγησης

Η ιδανική μορφή δεδομένων για εφαρμογή του συστήματός μας είναι παράλληλα κείμενα με τα εξής χαρακτηριστικά:

- Τα κείμενα να έχουν μικρές προτάσεις, ώστε να δημιουργούν μικρές λίστες υποψήφιων μεταφράσεων.

- Τα κείμενα να αναφέρονται σε μια περιορισμένη γνωστική περιοχή, ώστε να μη χρειάζεται αποσαφήνιση της ερμηνείας λέξεων που μπορούν να έχουν περισσότερες από μία ερμηνείες σε ευρύτερες γνωστικές περιοχές. Επίσης, τα λογοτεχνικά κείμενα δεν ενδείκνυνται, διότι οι συγγραφείς τους προσπαθούν να αποφύγουν την επανάληψη των ίδιων λέξεων.

- Τα κείμενα να αποτελούν το ένα ακριβή μετάφραση του άλλου και όχι να είναι και τα δύο μεταφράσεις κάποιου τρίτου κειμένου, ώστε να υπάρχει μεγαλύτερη πιστότητα μεταξύ των δύο κειμένων.

- Τέλος θα πρέπει τα κείμενα να έχουν αρκετά μεγάλο μέγεθος, ώστε να μην υπάρχουν προτάσεις που να περιέχουν περισσότερες από μια λέξεις με μόνο 1 εμφάνιση σε ολόκληρο το κείμενό τους (δηλ. με $f(x)=1$). Αν υπάρχουν περισσότερες από μία λέξεις με $f(x) = 1$ στην ίδια πρόταση, είναι σχεδόν αδύνατο να διακρίνουμε τη μετάφρασή τους.

4.1.1. Παράλληλα Κείμενα

Για την αξιολόγηση του συστήματος χρησιμοποιήθηκαν δύο κείμενα διαφορετικά σε μέγεθος αλλά και σε ακρίβεια μετάφρασης.

Α. Ως κείμενο «χαλαρής» μετάφρασης (κείμενο «Α») επιλέχθηκε ένα κείμενο του NATO που αφορά τις προκλήσεις του 21ου αιώνα.

κείμενο "Α"	Σύνολο Λέξεων	Διαφορετικές Λέξεις	Χαρακτήρες (χωρίς κενά)	URL
Ελληνικά	6.011	1.695	34.433	http://www.nato.int/docu/21-cent/21st_gr.pdf
Τουρκικά	4.227	1.896	30.534	http://www.nato.int/docu/21-cent/21st_tur.pdf
Αγγλικά	5.234	1.299	28.866	http://www.nato.int/docu/21-cent/21st_eng.pdf

Β. Ως κείμενο με μεγάλη ακρίβεια στην μετάφραση (κείμενο «Β») επιλέχθηκε το σχέδιο ΑΝΑΝ για την επίλυση του «Κυπριακού Προβλήματος» το οποίο και αποτελείται από μικρές προτάσεις πάντα σε αναλογία 1-1 ή 2-2 (URL Σχεδίου ANAN : http://www.cyprus-un-plan.org/).

κείμενο "Β"	Σύνολο Λέξεων	Διαφορετικές Λέξεις	Χαρακτήρες (χωρίς κενά)	Όνομα Εγγράφου[20]
Ελληνικά	41.870	5.158	241.950	Comprehensive_Settlement_of_the_Cyprus_Problem_Greek.pdf
Τουρκικά	31.618	6.663	217.350	Comprehensive_Settlement_of_the_Cyprus_Problem_Turkish.pdf
Αγγλικά	40.742	3.061	218.206	Comprehensive_Settlement_of_the_Cyprus_Problem.pdf

Στις δοκιμές αντιστοίχησης Ελληνικών – Αγγλικών (και αντίστροφα) δεν χρησιμοποιήσαμε καμία γλωσσική πληροφορία (Λήμματα).

4.1.2 Μορφοποίηση πριν την επεξεργασία

Κατά τη στοίχιση των προτάσεων των Ελληνικών και Τουρκικών κειμένων με το **Tr•AID Align (**ενότητα 3.1.2), προηγήθηκε μορφοποίηση των κειμένων, τα κυριότερα σημεία της οποίας ήταν τα εξής:

α. Διαμορφώθηκαν και τα δύο παράλληλα σώματα κειμένου έτσι ώστε το κείμενο κάθε γλώσσας να αποτελείται από μία μόνο παράγραφο.

β. Αφαιρέθηκαν όλες οι μορφοποιήσεις, ώστε να αποτελούνται και τα δύο από την ίδια γραμματοσειρά και χωρίς έντονη **(Bold)** ή πλάγια (*Italics*) γραφή.

γ. Παρατηρήθηκαν λάθη εντοπισμού αλλαγής προτάσεων στην Τουρκική γλώσσα, ιδιαίτερα σε περιπτώσεις όπου, λόγω ιδιομορφίας της γλώσσας, χρησιμοποιείται τελεία αμέσως μετά από κάθε αριθμό του κειμένου που αναφέρεται σε αριθμό άρθρου ή κεφαλαίου. Οι τελείες αυτές αφαιρέθηκαν.

δ. Ως Γλώσσα – Πηγή (Source) επιλέχθηκε η Ελληνική.

ε. Επειδή στο **Tr•AID Align** δεν υπήρχε επιλογή Τουρκικών για τη Γλώσσα – Στόχο, ορίστηκαν ως γλώσσα στόχος τα Φιλανδικά , που έχουν πολλούς κοινούς χαρακτήρες με τα Τουρκικά.

4.2. Μέθοδος Αξιολόγησης

Σε αντίθεση με αξιολογήσεις άλλων συστημάτων, δεν αξιολογούμε την αντιστοίχηση μόνο κάποιου υποσυνόλου των λέξεων των κειμένων, επιλέγοντας εκείνες με τις καλύτερες βαθμολογίες, αλλά επιχειρούμε την αντιστοίχηση, χωρίς καμία διάκριση, όλων των λέξεων των κειμένων και αξιολογούμε όλες τις αντιστοιχήσεις.

Μια επιλεκτική αξιολόγηση των «καλύτερων» αντιστοιχήσεων δεν θα συμπεριλάμβανε λέξεις για τις οποίες δεν είμαστε βέβαιοι για την αντιστοίχησή τους. Η βεβαιότητα για την ορθότητα της επιλογής μας μειώνεται όταν :

- Η βαθμολογία κατά αντιπρόσωπο (ή κατά λέξη) είναι αρκετά μεγαλύτερη από 0.

- Όταν προσπαθούμε να αντιστοιχήσουμε λέξεις που εμφανίζονται πολύ λίγες φορές στο κείμενο και ειδικότερα όταν υπάρχουν περισσότερες από μία τέτοιες λέξεις στην ίδια περίοδο.

[20] URL εγγράφων = URL Σχεδίου ANAN + όνομα εγγράφου.

Σε αρκετές από τις παραπάνω περιπτώσεις, θα ήταν δυνατό να παρουσιάζουμε περισσότερες από μία υποψήφιες αντιστοιχήσεις, μαζί με το ποσοστό βεβαιότητάς μας, με τρόπο που θα περιγράψουμε παρακάτω.

4.3. Αξιολόγηση Στοίχισης Προτάσεων

Α. Σε κείμενα «χαλαρής» μετάφρασης:

Συνολικά το **Tr•AID Align** χώρισε τα κείμενα σε **171** περιόδους, από τις οποίες μόνο οι **23** χρωματίστηκαν **Μπλέ** (δηλαδή το πρόγραμμα ήταν απόλυτα σίγουρο για τη στοίχιση τους, αν και έσφαλε σε μία από αυτές). Αναφέρονται παρακάτω ορισμένες ενδιαφέρουσες περιπτώσεις που το Tr•AID Align εντόπισε επιτυχώς:

- 6 περιπτώσεις αντιστοίχησης 2:1
- 3 περιπτώσεις αντιστοίχησης 1:2
- 2 περιπτώσεις αντιστοίχησης 2:2

Υπήρξαν ωστόσο και περιπτώσεις που το πρόγραμμα έσφαλε. Πολλές φορές τα λάθη δεν ήταν μεμονωμένα, αλλά ένα λάθος είχε ως συνεπακόλουθο άλλο 1, 2 ή και περισσότερα λάθη. Συνολικά εντοπίστηκαν 6 βασικά λάθη, τα οποία επηρεάζοντας και τις επόμενες αντιστοιχίσεις είχαν ως αποτέλεσμα 21 λάθος αντιστοιχίσεις, δηλαδή ορθότητα 87,7 %. Δύο από τα βασικά λάθη οφείλονταν σε αντιστοιχήσεις 1:3, ένα σε αντιστοίχιση 1:0 και ένα σε 0:2, περιπτώσεις που ο αλγόριθμος των Gale & Church αδυνατεί να εντοπίσει. Λαμβάνοντας υπόψη ότι οι Gale & Church σημειώνουν ορθότητα 96% αλλά χωρίς να καταφέρνουν να εντοπίσουν τις αντιστοιχήσεις 1:0 και 0:1 (ενώ ούτε καν προβλέπει αντιστοιχήσεις 1:3, 3:1, 2:0, 0:2, 3:2, 2:3), μπορούμε να θεωρήσουμε ικανοποιητική τη στοίχιση Ελληνικών-Τουρκικών προτάσεων που έγινε από το **Tr•AID Align**.

Β. Σε κείμενα «ακριβούς» μετάφρασης:

Στα κείμενα στα οποία το ένα ήταν πιστή μετάφραση του άλλου, τα αποτελέσματα ήταν πολύ καλύτερα αφού το ποσοστό ορθότητας ήταν 100%.

4.4. Αξιολόγηση Ομαδοποίησης Λέξεων

4.4.1. Ομαδοποίηση ΧΩΡΙΣ Γλωσσική πληροφορία

Ενδιαφέρον παρουσιάζει η περίπτωση της ομαδοποίησης χωρίς γλωσσική πληροφορία. Στον πίνακα που ακολουθεί παρουσιάζουμε τα αποτελέσματα από την εφαρμογή του αλγόριθμου που περιγράψαμε στην §3.2.3.1. Στον πίνακα που καθορίζουμε τα εξής :

- «ομαδοποιήσεις» καλούμε τις συσχετίσεις που έγιναν μεταξύ δύο μορφών της ίδιας λέξεως.

- «ελλιπείς ομαδοποιήσεις» καλούμε τις συσχετίσεις, που αν είχαν γίνει, θα ομαδοποιούσαν σε ένα σύνολο όλες τις μορφές της ίδιας λέξεως. Για παράδειγμα έστω ότι στο κείμενο υπάρχουν οι x_1, x_2, x_3, x_4, x_5, x_6, x_7 και x_8, μορφές της ίδιας λέξεως, οι οποίες έχουν ομαδοποιηθεί σε 3 υποσύνολα: $\{x_1, x_2, x_3\}$, $\{x_4, x_5\}$, $\{x_6, x_7, x_8\}$. Στο παράδειγμα αυτό θα έπρεπε να είχαν γίνει τουλάχιστον 2 ομαδοποιήσεις επιπλέον, ώστε η συνολική ομαδοποίηση να ισοδυναμεί με ένα σύνολο που θα περιέχει όλες τις μορφές της λέξεως x. Οι ομαδοποιήσεις που λείπουν είναι οποιεσδήποτε δύο που θα αρκούσαν για να ενώσουν τα τρία υποσύνολα σε ένα : π.χ. (x_3, x_4) και (x_4, x_7).

- «Λάθος ομαδοποιήσεις» είναι το πλήθος των λάθος (ξένων) λέξεων που περιέχονται σε ένα σύνολο ομαδοποιημένων λέξεων.

- «Μη Εφαρμογή Ομαδοποίησης» συμβαίνει όταν μια λέξη δεν έχει συσχετιστεί με καμία άλλη (ενώ θα έπρεπε).

		ΕΛΛΗΝΙΚΑ	ΤΟΥΡΚΙΚΑ
α΄	Διαφορετικές λέξεις	1695	1895
β΄	Συνολικές Ομαδοποιήσεις	658	723
γ΄	Λάθος Ομαδοποιήσεις[21]	22	22
δ΄	Ελλιπείς Ομαδοποιήσεις	28	182
ε΄	Μη Εφαρμογή Ομαδοποίησης	89	85
στ΄	**Ακρίβεια** (β-γ/β)	**95,1368 %**	**96,9571 %**
ζ΄	**Ανάκληση** (β-γ/(β-γ)+δ+ε)	**84,2530 %**	**72,4174 %**

Πίνακας 4: Αποτελέσματα ομαδοποίησης χωρίς γλωσσική πληροφορία

4.4.2. Ομαδοποίηση ΜΕ Γλωσσική πληροφορία

Αν και στην περίπτωση της ομαδοποίησης με τη χρήση γλωσσικής πληροφορίας περιμέναμε καλύτερα αποτελέσματα από εκείνα της μεθόδου χωρίς γλωσσική πληροφορία, εν τούτοις τα αποτελέσματα από τις βαθμολογίες κατά Λήμμα, ήταν κατώτερα των προσδοκιών μας.

Τα περισσότερα προβλήματα παρατηρήθηκαν στην Τουρκική γλώσσα και οφείλονταν στους εξής λόγους:

- Οι αμφιλεγόμενες συσχετίσεις λέξεων με τα λήμματά τους δεν ήταν εύκολο να επιλυθούν, αφού η λημματοποίηση έγινε σε λίστα λέξεων και όχι σε κείμενο επισημειωμένο με τα μέρη του λόγου[22].

- Κάποιοι χαρακτήρες, του Τουρκικού αλφάβητου (“ğ”, “ı”, “ş”) χάθηκαν κατά τη στοίχιση με το **Tr•AID Align**, αφού αυτό δεν υποστηρίζει την Τουρκική γλώσσα. Οι χαρακτήρες αυτοί αντικαταστάθηκαν από άλλους (“g”, “i”, “s”) κάνοντας έτσι αδύνατο τον συσχετισμό με το (ορθά γραμμένο) λήμμα. Επίσης είναι δυνατό δυο εντελώς διαφορετικές λέξεις (ως προς το νόημα και τη γραφή) μετά από αυτές τις αντικαταστάσεις να γράφονται με τον ίδιο τρόπο και κατά συνέπεια να ομαδοποιούνται λανθασμένα με το ίδιο λήμμα. Η υποκατάσταση των 2 αυτών χαρακτήρων είχε ως αποτέλεσμα να μην συσχετιστούν με κανένα λήμμα 700 λέξεις του κειμένου «Β» και 200 του κειμένου «Α». Αυτή εκτιμούμε ότι είναι η αιτία που ευθύνεται για τα περισσότερα λάθη που παρατηρήθηκαν κατά την ομαδοποίηση και βαθμολογία κατά «λήμμα».

[21] Περιέχονται στις συνολικές ομαδοποιήσεις.
[22] βλ. παράδειγμα §2.2. που αφορά τη Τουρκική λέξη “yaz”

4.5. Αξιολόγηση Αντιστοίχησης Λέξεων

Για την αξιολόγηση των αποτελεσμάτων της αντιστοίχησης λέξεων, αναζητάμε την αντίστοιχη λέξη της άλλης γλώσσας στα πρώτα 5 καλύτερα αποτελέσματα. Επεκτείνουμε την αναζήτηση σε περισσότερα αποτελέσματα μόνο όταν υπάρχει ισοβαθμία περισσότερων από 5 λέξεων στην 1^{η} θέση. Για την αξιολόγηση της αντιστοίχισης των λέξεων ορίζουμε τα παρακάτω:

- «*Λέξεις που ΔΕΝ μεταφράστηκαν*» (Πιν. 6, Γραμμή γ΄) είναι εκείνες των οποίων η μετάφραση δεν συμπεριλαμβανόταν μέσα στα 5 πρώτα αποτελέσματα καμίας μεθόδου αξιολόγησης ενώ οι υπόλοιπες χαρακτηρίζονται ως «λέξεις που μεταφράστηκαν» (Πιν. 6, Γραμμή δ΄).
- «*Λέξεις που μεταφράστηκαν σωστά*» (Πιν. 6, γραμμή ε΄) είναι εκείνες που η μετάφρασή τους ήταν η πρώτη στη βαθμολογία *κατά αντιπρόσωπο* (ακόμα και αν ισοβάθμησε στην πρώτη θέση μαζί με άλλες).

- $$\text{ΟΡΘΟΤΗΤΑ} = \frac{\text{Λέξεις που μεταφράστηκαν σωστά}}{\text{Λέξεις που εξετάστηκαν}}$$

Μετά από αξιολόγηση των αποτελεσμάτων έχουμε τους παρακάτω πίνακες:

	Κείμενο "Β"	Ελληνικά	Τουρκικά
α	Διαφορετικές λέξεις κειμένων	5.215	6.663
β	Δείγμα λέξεων που εξετάστηκαν	919	962
γ	*Λέξεις που ΔΕΝ μεταφράστηκαν*	*300*	*371*
δ	*Λέξεις που μεταφράστηκαν*	*619*	*591*
ε	Λέξεις που μεταφράστηκαν σωστά	390	410
στ	**Ορθότητα (ε/β)**	**42.4374 %**	**42.6195 %**

Πίνακας 6.α: Αποτελέσματα αντιστοίχησης λέξεων για Ελληνικά – Τουρκικά

Το μεγαλύτερο πρόβλημα στην αντιστοίχηση των λέξεων Ελληνικών – Τουρκικών (και αντίστροφα) είναι ότι πολλές ελληνικές λέξεις δεν αντιστοιχίζονται σε κάποια τουρκική λέξη αλλά σε κατάληξη τουρκικής λέξεως.

Απόλυτη ορθότητα είναι πολύ δύσκολο να επιτευχθεί αν αναλογιστούμε και τις διαφορές που έχουν οι δύο γλώσσες, όπως η ύπαρξη συνώνυμων λέξεων μόνο στην μία από τις δύο γλώσσες. Για παράδειγμα, οι συνώνυμες, Τουρκικές λέξεις "birinci" και "ilk" θα μεταφραστούν και οι δύο στο Ελληνικό κείμενο ως «πρώτος». Το αποτέλεσμα, στη χειρότερη περίπτωση, θα είναι *f*("birinci") = *f*("ilk") = *f*("πρώτος")/2 με συνέπεια η λέξη «πρώτος» να μην αντιστοιχηθεί με καμία από αυτές καθώς είναι πολύ πιθανό κάποια άλλη λέξη να έχει μεγαλύτερη συχνότητα από τις 2 αυτές Τουρκικές λέξεις.

Ο παρακάτω πίνακας δείχνει τα αντίστοιχα αποτελέσματα στην περίπτωση Ελληνικών – Αγγλικών.

		Ελληνικά	Αγγλικά
α	Διαφορετικές λέξεις κειμένων	1.695	1.299
β	Δείγμα λέξεων που εξετάστηκαν	1.000	1.000
γ	Λέξεις που ΔΕΝ μεταφράστηκαν	193	185
δ	Λέξεις που μεταφράστηκαν	807	815
ε	Λέξεις που μεταφράστηκαν σωστά	575	600
στ	**Ορθότητα (ε/β)**	**57,5 %**	**60 %**

Πίνακας 6.β: Αποτελέσματα αντιστοίχησης λέξεων για Ελληνικά – Αγγλικά

4.6. Αξιολόγηση Αντιστοίχησης Πολυλεκτικών Όρων

Τα αποτελέσματα από τον εντοπισμό και τη μετάφραση των πολυλεκτικών όρων των δύο παράλληλων κειμένων φαίνονται στους παρακάτω πίνακες. Οι πολυλεκτικοί όροι που εντοπίστηκαν παρατίθενται στο Παράρτημα 'Α':

		Κείμενο «Α»		Κείμενο «Β»	
		Ελληνικά	Τουρκικά	Ελληνικά	Τουρκικά
α	Σύνολο Πολυλεκτικών Όρων[23]	30	30	48	48
β	Εντοπίσθηκαν	15	13	64	50
γ	Εντοπίσθηκαν Ορθά	14	11	43	30
δ	Μεταφράστηκαν Ορθά	6	5		
ε	**Ανάκληση Εντοπισμού (γ/α)**	46,6667 %	36,6667 %	89,5833 %	62,5 %
στ	**Ακρίβεια Εντοπισμού (γ/β)**	93,3333 %	84,6154 %	87,5 %	62,5 %
ζ	**Ορθότητα Μεταφράσεων (δ/γ)**	42,8571 %	45,4545 %		

Πίνακας 7.α: Αποτελέσματα εντοπισμού και μετάφρασης πολυλεκτικών όρων στα Ελληνικά–Τουρκικά

		Κείμενο «Α»	
		Ελληνικά	Αγγλικά
α	Σύνολο Πολυλεκτικών Όρων	30	30
β	Εντοπίσθηκαν	15	12
γ	Εντοπίσθηκαν Ορθά	14	7
δ	Μεταφράστηκαν Ορθά	7	5
ε	**Ανάκληση Εντοπισμού (γ/α)**	46.6667 %	23.3334 %
στ	**Ακρίβεια Εντοπισμού (γ/β)**	93,3333 %	58,3333 %
ζ	**Ορθότητα Μεταφράσεων (δ/γ)**	50 %	71 %

Πίνακας 7.β: Αποτελέσματα εντοπισμού και μετάφρασης πολυλεκτικών όρων στα Ελληνικά–Αγγλικά

Το μεγαλύτερο πρόβλημα του αλγόριθμου *εντοπισμού* πολυλεκτικών όρων είναι ο θόρυβος που υπάρχει από λέξεις που δεν αποτελούν πολύ-λεκτικούς όρους και απλά τυχαίνει να συνυπάρχουν πολλές φορές (τουλάχιστον 2) σε ίδια πρόταση. Για να αντιμετωπίσουμε αυτό το πρόβλημα, θέσαμε πιο αυστηρά κριτήρια επιλογής, αλλά αναπόφευκτα υπήρξε επίδραση στην ανάκληση, όπως βλέπουμε στους παραπάνω πίνακες. Ένα άλλο πρόβλημα που αντιμετωπίσαμε κατά τον εντοπισμό πολυλεκτικών όρων είναι η επανάληψη κάποιων προτάσεων του κειμένου. Σε αυτές τις περιπτώσεις, το σύστημά μας θεωρεί τις επαναλαμβανόμενες προτάσεις ως πολυλεκτικούς όρους.

Τα χαμηλά ποσοστά στην ακρίβεια των *μεταφράσεων* των πολυλεκτικών όρων οφείλονται στα χαμηλά ποσοστά της ανάκλησης κατά τον εντοπισμό των πολυλεκτικών όρων. Συχνά, δηλαδή, ένας πολυλεκτικός όρος δεν είναι δυνατόν να μεταφραστεί σωστά, επειδή δεν έχει εντοπιστεί ο αντίστοιχος πολυλεκτικός όρος της άλλης γλώσσας.

[23] Ο υπολογισμός του συνόλου των πολυλεκτικών όρων έγινε χειρωνακτικά. Το ίδιο και στον επόμενο πίνακα.

Κεφάλαιο 5ο

Συμπεράσματα – Μελλοντικές Κατευθύνσεις

5.1. Συμπεράσματα

Από τα αποτελέσματα της αντιστοίχησης προτάσεων με τον αλγόριθμο Gale&Church συμπεραίνουμε ότι είναι δυνατή η εφαρμογή του και σε γλώσσες που δεν έχουν Ευρωπαϊκή προέλευση.

Τα αποτελέσματα της αντιστοίχησης λέξεων δεν είναι τα ίδια στην περίπτωση Ελληνικών – Τουρκικών και Ελληνικών – Αγγλικών (βλ. πίνακες 6.α και 6.β). Αυτό οφείλεται κατ' αρχάς στην διαφορετική στοίχιση προτάσεων που έγινε στην φάση προεπεξεργασίας των παράλληλων κειμένων αλλά και στην πιο σύνθετη μορφολογία των Αγγλικών από τα Τουρκικά. Χαρακτηριστικό της πιο σύνθετης μορφολογίας της Τουρκικής είναι το γεγονός ότι σπάνια μια Αγγλική λέξη εμφανίζεται στο κείμενο με περισσότερες από 3 μορφές. Αντίθετα, μια Τουρκική μπορεί να εμφανιστεί με πολύ περισσότερες μορφές, αφού οι καταλήξεις που είναι δυνατό να συγκολληθούν είναι πολύ περισσότερες.

Η κυριότερη αιτία αποτυχίας στην μετάφραση μιας λέξης είναι η λάθος ή η ελλιπής ομαδοποίηση των λέξεων. Σημαντική βελτίωση στην συνολική απόδοση του συστήματος θα υπάρξει αν βελτιώσουμε τον αλγόριθμο ομαδοποίησης των λέξεων χωρίς γλωσσική πληροφορία.

Ο εντοπισμός των πολυλεκτικών όρων έγινε με τη χρήση της ίδιας μετρικής με αυτή που χρησιμοποιήσαμε για την αντιστοίχηση λέξεων. Μεγαλύτερο πρόβλημα στους πολυλεκτικούς όρους είναι η είσοδος θορύβου, δηλαδή λέξεων άσχετων μεταξύ τους, που τυχαία βρίσκονται συχνά στην ίδια πρόταση αρκετές φορές. Αν και τα χαμηλά ποσοστά ανάκλησης στον εντοπισμό των πολυλεκτικών όρων επηρεάζουν την ορθότητα των μεταφράσεών τους, το στάδιο της μετάφρασης πολυλεκτικών όρων αποδίδει ικανοποιητικά, αφού καταφέρνει να μεταφράσει σωστά τους πολυλεκτικούς όρους υπό την προϋπόθεση ότι έχουν εντοπιστεί και οι πολυλεκτικοί όροι που αποτελούν τις σωστές τους μεταφράσεις.

Εκτιμούμε ότι τα αποτελέσματα του συστήματος θα βελτιωθούν ακόμα περισσότερο αν χρησιμοποιηθούν πολύ μεγαλύτερα σώματα κειμένου.

5.2. Μελλοντικές Κατευθύνσεις

Σε πρώτη φάση μάς απασχολεί η μελέτη της συμπεριφοράς του συστήματος και σε άλλα ζεύγη γλωσσών, καθώς και σε μεγαλύτερα παράλληλα σώματα κειμένου (με τουλάχιστον 250.000 λέξεις).

Ενδιαφέρον παρουσιάζει η βελτίωση της μεθόδου εντοπισμού και μετάφρασης πολυλεκτικών όρων με τη χρήση γλωσσικής πληροφορίας. Να χρησιμοποιηθούν, δηλαδή, κάποιες χαρακτηριστικές ακολουθίες μερών του λόγου (π.χ. δύο συνεχόμενα ουσιαστικά) ως πρότυπα (patterns) για το φιλτράρισμα των ενδιάμεσων «άσχετων λέξεων» που εμφανίζονται ως «θόρυβος» στους εντοπισμένους πολυλεκτικούς όρους.

Ένα σημαντικό πρόβλημα που θα πρέπει να επιλυθεί είναι η απόκτηση από το σύστημα δυνατότητας κατάταξης ισοβαθμούντων μεταφράσεων, καθώς και η ικανότητα διάκρισης των λέξεων που δεν ήταν δυνατό να μεταφραστούν (πιθανότατα λόγω μικρού μεγέθους των παράλληλων κειμένων). Σε αυτή τη κατεύθυνση θα παρουσίαζε ενδιαφέρον η δυνατότητα επίλυσης των ισοβαθμιών λαμβάνοντας υπόψη τα συμφραζόμενα.

Θα μπορούσαν, τέλος, να χρησιμοποιηθούν αλγόριθμοι μηχανικής μάθησης για την εκμάθηση των βέλτιστων τιμών διαφόρων παραμέτρων (π.χ. ορίων αποκοπής) που χρησιμοποιούμε.

Αναφορές

[1] Piperidis, S., Papageorgiou, H., Demiros, I., Malavazos, C., Triantafyllou, I, (1998). A Framework for Example-based Translation-Aid Tools", Proceedings of the Panhellenic Conference on New Information Technology-(NIT'98) 8-10 October 1998, Athens, Greece, 269-278.

[2a] Piperidis, S., Malavazos, C., Triantafyllou, Y., (1999). A Multi-level Framework for Memory-Based Translation Aid Tools, Aslib, Translating and the Computer 21, 10-11 November 1999, London.

[2b] Papageorgiou, H., Prokopidis, P., Giouli, V., Piperidis, S., (2000) "A Unified Tagging Architecture and its Application to Greek", Proceedings of Second International Conference on Language Resources and Evaluation-LREC2000, 31 May- 2 June2000, Athens, Greece, 1455-1462.

[3] Piperidis, S., Boutsis, S., Demiros, I., (1997). Automatic Translation Lexicon Generation from Multilingual texts, Workshop on Multilinguality in the Software Industry: the AI Contribution (MULSAIC'97), Fifteenth International Joint Conference on Artificial Intelligence (IJCAI'97), 25 August 1997, Nagoya, Japan, 57-62.

[4] Piperidis, S., Papageorgiou H., Boutsis, S. (2000) From sentences to words and clauses. In Veronis, J. (Ed) Parallel Text Processing, Alignment and use of translation corpora, Kluwer Academic Publishers, Text Speech and Language Technology Series, pp. 117-138.

[5] [Gale and Church, 1991b] Gale W.A. and Church K.W., "A Program for Aligning Sentences in Parallel Corpora", *Proceedings of the 29th Annual Meeting of the ACL.* pp.177-184*,* 1991.

[6] [Smadja *et al*, 1996] Smadja F., McKeown K.R., Hatzivassiloglou V. "Translating Collocations for Bilingual Lexicons: A Statistical Approach" *Computational Linguistics*, 22(1): 1-38, 1996.

[7] Kupiec, J. "An algorithm for finding noun phrase correspondences in bilingual corpora." *Proceedings of the 31st Annual Meeting of the ACL*, Columbus, Ohio, 1993.

[8] Van der Eijk, Pim. "Automating the Acquisition of Bilingual Terminology." In *Proceedings, Sixth Conference of the European Chapter of the Association for Computational Linguistics,* Utrecht, The Netherlands, 113-119. Association for Computational Linguistics. 1993.

[9] Kenji Yamada & Kevin Knight : "A Syndax-based Statistical Translation Model", Meeting of the Association for Computational Linguistics, p.523-530, 2001.

[10] Brown, P., J. Lai, and R. Mercer. "Aligning sentences in parallel corpora." In Proc. 29th Annual Meeting of the ACL, 169-176. 18-21 June, Berkley, Calif. 1991.

[11] Serhiy Kosinov : "Evaluation of N-GRAMS Conflation Approach in text-based information retrieval" , 2001.

[12] R. Krovetz, 1993: "Viewing morphology as an inference process," in R. Korfhage κ.ά., *Proc. 16th ACM SIGIR Conference*, Pittsburgh, pp. 191-202, June 27-July 1, 1993.

[13] Martin Porter : "An algorithm for suffix stripping", M.F., 1980.
(http://www.tartarus.org/~martin/index.html)

[14] Dan Tufiş, Ana-Maria Barbu : "Automatic construction of translation lexicons.".

[15] Ralf D. Brown, Jaime Carbonell, Yiming Yang : “Automatic Dictionary Extraction for Cross-Language Information Retrieval”, Dec 1998.

[16] Kyo Kageura, Keita Tsuji, Akiko N. Aizawa : “Automatic Thesaurus Generation through Multiple Filtering.”, 2000.

[17] Jörg Tiedemann : “Recycling Translations. *Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*”. Acta Universitatis Upsaliensis. *Studia Linguistica Upsaliensia* 1. 130 pp. Uppsala. ISBN 91-554-5815-7, 2003.

[18] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein : “Introduction to Algorithms”, ISBN 0-07-013151-1.

[19] Michel Simard, George F. Foster, and Pierre Isabelle : “Using cognates to align sentences in bilingual corpora.” In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pages 67-81, Montreal, Canada, 1992.

[20] Éric Gaussier : “Flow network models for word alignment and terminology extraction from bilingual corpora.”, 1998.

[21] Frege, G. “On Sense and Nominatum, in A.P.Martinich (ed) The Philisophy of Language.”, Oxford, pp 186-198, 1892.

[22] WEKA : http://www.cs.waikato.ac.nz/~ml/weka/index.html

Παράρτημα:

Μετάφραση Εντοπισμένων Ελληνικών Πολυλεκτικών Όρων Στο Κείμενο 'Α'

	ΕΛΛΗΝΙΚΑ			ΤΟΥΡΚΙΚΑ		ΑΓΓΛΙΚΑ	
	Εντοπισμένοι Πολύ-λεκτικοί Όροι	πλήθος	**ΟΡΘΟ**	Αντιστοίχιση	**ΟΡΘΟ**	Αντιστοίχιση	**ΟΡΘΟ**
1	σοβιετικής ένωσης	4	**ΝΑΙ**	sovyetler+ birligi'nin	**ΝΑΙ**	soviet	
2	ίδια στιγμή	4	**ΝΑΙ**	Anilan		Δεν Εντοπίστηκε	
3	βόρεια αμερική	4	**ΝΑΙ**			america	
4	ευρω-ατλαντικό χώρο	5	**ΝΑΙ**	avrupa-atlantik		euro-atlantic+ area	**ΝΑΙ**
5	ηνωμένες πολιτείες	7	**ΝΑΙ**	ABD'ye	**ΝΑΙ**	united+ states	**ΝΑΙ**
6	ηνωμένων πολιτειών	3	**ΝΑΙ**			united+ states (2ο)	
7	κράτη μέλη	7	**ΝΑΙ**	Edilen		Δεν Εντοπίστηκε	
8	κρατών μελών	7	**ΝΑΙ**			Δεν Εντοπίστηκε	
9	ψυχρού πολέμου	8	**ΝΑΙ**	soguk (κρύο)		cold	
10	διάσκεψη κορυφής πράγας	4	**ΝΑΙ**	prag zirvesinde	**ΝΑΙ**	summit(1) , prague(2)	
11	αριθμός τοις εκατό	3	**ΝΑΙ**	sayisi		cut+ per+ cent	**ΝΑΙ**
12	τρομοκρατικές επιθέσεις σεπτεμβρίου	4	**ΝΑΙ**	ll+ eylül+ terörist	**ΝΑΙ**	terrorist+ attacks+ september	**ΝΑΙ**
13	όπλων μαζικής καταστροφής	9	**ΝΑΙ**	kitle+ imha+ silahlari	**ΝΑΙ**	weapons+ mass+ destruction	**ΝΑΙ**
14	όπλα μαζικής καταστροφής	9	**ΝΑΙ**	kitle+ imha+ silahlarinin	**ΝΑΙ**	weapons+ mass+ destruction	**ΝΑΙ**
15	πρώτη φορά επίκληση αρθρου	3		Maddesini (άρθρου)		first(1) , article(2), invoked(3)	
	Εντοπίστηκαν συνολικά		**14**	Μεταφράστηκαν Ορθά	**6**	Μεταφράστηκαν Ορθά	**6**
	Ακρίβεια (εντοπισμού): 93.3333%			**Ακρίβεια Μετάφρασης** [24] **: 45,4545%**		**Ακρίβεια Μετάφρασης**[25] **: 71%**	

[24] Από Ελληνικά σε Τουρκικά
[25] Από Ελληνικά σε Αγγλικά