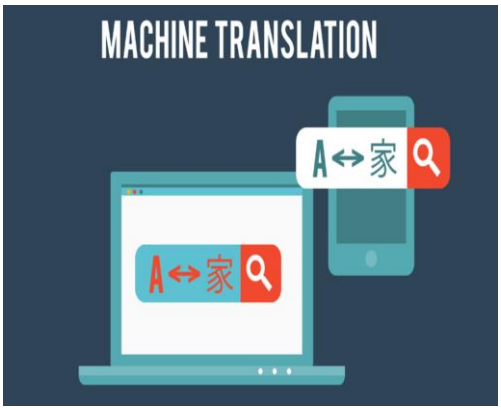
**Πανεπιστήμιο Ιωαννίνων**

**Τμήμα Πληροφορικής και Τηλεπικοινωνιών**


# ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΤΕΧΝΙΚΕΣ ΣΤΑΤΙΣΤΙΚΗΣ ΜΗΧΑΝΙΚΗΣ ΜΕΤΑΦΡΑΣΗΣ

Κωνσταντίνος Μάρας

ΑΜ: 9398

Επιβλέπων: Χρυσόστομος Στύλιος

**ΑΡΤΑ 2021**

**(ΣΕΛΙΔΑ ΕΓΚΡΙΣΗΣ)**

**ΚΑΘΗΓΗΤΗΣ 1:**

## ΠΝΕΥΜΑΤΙΚΑ ΔΙΚΑΙΩΜΑΤΑ

## ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ

Δηλώνω υπεύθυνα και γνωρίζοντας τις κυρώσεις του Ν. 2121/1993 περί Πνευματικής Ιδιοκτησίας, ότι η παρούσα πτυχιακή εργασία είναι εξ ολοκλήρου αποτέλεσμα δικής μου ερευνητικής εργασίας, δεν αποτελεί προϊόν αντιγραφής ούτε προέρχεται από ανάθεση σε τρίτους. Όλες οι πηγές που χρησιμοποιήθηκαν (κάθε είδους, μορφής και προέλευσης) για τη συγγραφή της περιλαμβάνονται στη βιβλιογραφία.

Μάρας Κωνσταντίνος

Άρτα, Απρίλιος, 2021

# ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα αρχικά να εκφράσω, τις θερμές μου ευχαριστίες στον επιβλέποντα καθηγητή μου Χρυσόστομο Στύλιο, για την εμπιστοσύνη που μου έδειξε, την πολύτιμη βοήθεια, καθοδήγηση και υπομονή σε όλη τη διάρκεια της εκπόνησης της πτυχιακής εργασίας καθώς και τις γνώσεις που μου μετέδωσε κατά τις σπουδές. Επίσης θα ήθελα να ευχαριστήσω την οικογένεια μου, για την τεράστια υποστήριξη που μου παρείχαν οποιαδήποτε στιγμή για τις σπουδές μου.

## ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία προσεγγίζεται το θέμα της στατιστικής μηχανικής μετάφρασης.

Αρχικά, γίνεται βιβλιογραφική ανασκόπηση στο θέμα της μετάφρασης και ανάλυση των εννοιών της, με τη μετάφραση να βρίσκεται διαχρονικά στην επίκεντρο της διαπολιτισμικής επικοινωνίας και είναι αναπόσπαστο κομμάτι της τεχνολογίας. Στη συνέχεια, γίνεται αναφορά και ανάπτυξη των γνωστών μεθόδων μετάφρασης, η μετάβαση στην μετάφραση του υπολογιστή και η εισαγωγή στη μηχανική μετάφραση.

Η ανάλυση των τρόπων εκμάθησης της μηχανικής μετάφρασης, οι δικλείδες ασφαλείας του υπολογιστικού συστήματος, τα προβλήματα που προκύπτουν κατά τη διαδικασία αυτή, είναι τα θέματα που αναλύονται στη συνέχεια.

Τέλος, γίνεται εισαγωγή στη στατιστική μηχανική μετάφραση, την πορεία του συστήματος μέσα από τροποποιήσεις και μεταμορφώσεις στη πάροδο του χρόνου και τις εκάστοτε τεχνικές, ενώ παράλληλα αναλύονται τα θετικά και αρνητικά που προκύπτουν από την κάθε φάση του συστήματος και την χρησιμότητα της συγκεκριμένης τεχνολογίας, σε κλάδους αιχμής.



Λέξεις κλειδιά: Μετάφραση, Μηχανική μετάφραση, Στατιστική μηχανική μετάφραση

## ABSTRACT

In the present work, the approach of statistical machine translation is accomplished.

Firstly, we have a bibliographic review on translation and its concepts analysis, whilst translation is at the center of intercultural communication and is an integral part of technology. Next, a reference is made to the known translation methods, the transition to computer translation, and the introduction to machine translation.

A reference on how machine translation is learned, the computer security flaws and the problems that arise during this process are the topics that are discussed below.

Finally, the introduction to statistical machine translation, the course of the system through modifications and transformations over time and corresponding techniques, while analyzing the advantages and disadvantages of each phase of the system and the usefulness of the technology concerned, in areas on demand.



Key words: Translation, Translate, machine translation, Word based models, Phrase based models, Statistical machine translation

## ΠΕΡΙΕΧΟΜΕΝΑ

## Κατάλογος εικόνων

## Κατάλογος πινάκων



## Συντομογραφίες

MM: Μηχανική Μετάφραση

MT: Machine Translation

CAT: Computer Assisted / Aided Translation

ALPAC: Automatic Language Processing Advisory Committee

FAHQT: Fully Automatic High - Quality Translation

TMS: Translation Memory System

ACL: Applied Computational Linguistics

MAHT: Machine-aided Human Translation

SMT: Statistical Machine Translation

EM: Expectation-Maximization

RNN: Recurrent Neutral Network

LSTM: Long Short-Term Memory

BLEU: Bi Lingual Evaluation Understudy

# Κεφάλαιο 1: Ιστορική αναδρομή της μετάφρασης

## 1.1 Η μετάφραση ως έννοια

Η μετάφραση είναι μια πράξη γραφής που μεταφέρει την έννοια ενός κειμένου γραμμένου από μία γλώσσα σε μία άλλη. Η ιστορία της μετάφρασης συμβαδίζει με την ιστορία της γραφής, της επιστήμης και της βιβλιογραφίας άλλων πολιτισμών. Υπάρχουν στοιχεία για μεταφράσεις ενός Σουμερικού έπους που χρονολογείται από τη δεύτερη χιλιετία π.Χ. Ένα ταξίδι μέσα στους αιώνες αποκαλύπτει σημαντικές μεταφραστικές δραστηριότητες στην Αρχαία Κίνα, τη Μεσοποταμία, τη Ρώμη, το Τολέδο και στην αυλή του βασιλιά Alfred the Great.

Η μετάφραση έφερε γνώση από την Ανατολή στην Ευρώπη και οι μεταφράσεις της Αγίας Γραφής στα γερμανικά και τα αγγλικά είχαν καθοριστική σημασία για την αλλαγή της πορείας της ιστορίας. Ωστόσο, τα βασικά εργαλεία του μεταφραστή για κάθε πιθανή επίδραση της δημιουργίας του ήταν και είναι ακόμα απλά: ένα όργανο γραφής και ένα μέσο για να γράφει.

Αυτό που ουσιαστικά χρειάζεται ένας μεταφραστής για να μεταφράσει το Κατά Μάρκον Ευαγγέλιον, ένα απαιτητικό δοκίμιο από τον Heidegger ή ένα σύγχρονο ιαπωνικό μυθιστόρημα σε ένα κατανοητό κείμενο στα ουαλικά (ή τα αγγλικά) είναι ένα κοφτερό μυαλό, ταλέντο, ένα τετράδιο Α4 και ένα μολύβι.

Το δεύτερο μισό του εικοστού αιώνα προήχθη η επαγγελματοποίηση της μετάφρασης. Ιδρύθηκαν ινστιτούτα και οργανώσεις, αναπτύχθηκαν πρότυπα και δεξιότητες και τα πανεπιστήμια εισήγαγαν μαθήματα που αποσκοπούσαν στην προετοιμασία των αποφοίτων τους για την είσοδο στο νέο αυτό επάγγελμα. Οι νέοι πτυχιούχοι μεταφραστές θα ξεκινήσουν τη σταδιοδρομία τους ως ελεύθεροι επαγγελματίες, οι οποίοι θα εργάζονται σε φορείς Παροχής Γλωσσικών Υπηρεσιών (Language Service Providers) (LSP) ή ως μεταφραστές προσωπικού για κυβερνήσεις και εμπορικούς οργανισμούς.

Η αυξανόμενη ζήτηση μετάφρασης παγκοσμίως έχει επικεντρώσει τις εμπορικές και κυβερνητικές οργανώσεις στο κόστος της μετάφρασης και την ανάγκη μεγιστοποίησης της παραγωγικότητας των μεταφραστών.

Η τεχνολογία της μετάφρασης έχει αναγνωριστεί ως μέσο ενίσχυσης της παραγωγικότητας. Μία σταθερά στις ποικίλες πορείες σταδιοδρομίας των επαγγελματιών μεταφραστών είναι ότι όλοι θα πρέπει να είναι ικανοί για τη χρήση μιας σειράς εργαλείων μετάφρασης. (Lewis, 2015)

## 1.2 Ιστορική αναδρομή

Αν και οι πρώτες ιδέες για αυτόματη μετάφραση συναντώνται σε επιστημονικά κείμενα του 17ου αιώνα, οι πρώτες πρακτικές προτάσεις έγιναν στον 20ο αιώνα και πιο συγκεκριμένα το 1933, όταν εμφανίστηκαν ταυτόχρονα στη Γαλλία και στη Ρωσία πατέντες για την κατασκευή μηχανικών λεξικών (Σοφιανόπουλος, 2009). Αργότερα, το 1947 επιστήμονες στον τομέα της κρυπτογραφίας, όπως οι Booth και Weaver, πρότειναν τη χρησιμοποίηση των υπολογιστών, οι οποίοι είχαν πρόσφατα εφευρεθεί, για την αυτόματη μετάφραση ανάμεσα σε δύο γλώσσες (Hutchins 1997). Μάλιστα ο Weaver την ίδια χρονιά, ως διευθυντής του ιδρύματος Rockefeller, κατέγραψε σε ένα υπόμνημα τις προτάσεις του για την αντιμετώπιση του προβλήματος της αμφισημίας (μία λέξη που μπορεί να έχει πολλές έννοιες) , χρησιμοποιώντας τις γνώσεις του στην κρυπτογραφία, στη στατιστική και στη λογική.

Στη δεκαετία του ’50 οι σχετικές μελέτες χαρακτηρίζονται από μεγάλη αισιοδοξία ως προς τα προσδοκώμενα αποτελέσματα. Οι πρώτες γενιές μεταφραστικών συστημάτων ήταν προσεγγίσεις άμεσης μετάφρασης (direct translation approaches), συστήματα από τα οποία απουσίαζαν οποιουδήποτε είδους ενδιάμεσα στάδια. Αρχικά, το κείμενο της γλώσσας πηγής υποβάλλεται σε απλή μορφολογική ανάλυση και στη συνέχεια γίνεται αναζήτηση σε ένα απλό δίγλωσσο λεξικό για τα μεταφραστικά ισοδύναμα κάθε λέξης. Το τελικό κείμενο στη γλώσσα στόχο προκύπτει με την εφαρμογή κάποιων τοπικών κανόνων αναδιοργάνωσης ώστε να τοποθετηθούν οι λέξεις στη σωστή σειρά.

Το πρώτο τέτοιο σύστημα ΜΜ παρουσιάζεται από την IBM στις ΗΠΑ το 1954 και συγκεντρώνει το ενδιαφέρον των ερευνητών. Για το πείραμα αυτό, ένα προσεκτικά επιλεγμένο δείγμα 49 ρωσικών προτάσεων μεταφράστηκε στα αγγλικά, με τη χρήση ενός περιορισμένου λεξικού 250 λέξεων και 6 γραμματικών κανόνων. Παρά το γεγονός ότι το σύστημα είχε ελάχιστη επιστημονική αξία, το μεταφραστικό αποτέλεσμα ήταν αρκετά εντυπωσιακό ώστε να τονώσει τη

χρηματοδότηση στις ΗΠΑ. Λόγω του ψυχρού πολέμου οι προσπάθειες τα επόμενα χρόνια επικεντρώθηκαν στο γλωσσικό ζεύγος ρωσικά-αγγλικά με αντικείμενο μετάφρασης κυρίως επιστημονικά και τεχνικά κείμενα.

Η ποιότητα της παραγόμενης μετάφρασης μπορεί να μην ήταν επαρκής για τη σε βάθος κατανόηση του κειμένου, αρκούσε όμως για να αναγνωριστεί κατά πόσο ήταν σημαντικό ώστε να αποσταλεί σε ένα μεταφραστή για πλήρη μετάφραση.

Στα μέσα της δεκαετίας του ’60 το θετικό αυτό κλίμα αλλάζει ριζικά. Η ΜΜ γίνεται αντικείμενο σοβαρής κριτικής και ριζικής αμφισβήτησης. Αυτό οφείλεται κυρίως σε μια αναφορά αξιολόγησης που συντάσσεται το 1966 από την επιτροπή εμπειρογνωμόνων ALPAC (Automatic Language Processing Advisory Committee), η οποία είχε συσταθεί από την Εθνική Ακαδημία Επιστημών των ΗΠΑ. Σύμφωνα με την αναφορά, τα υπέρογκα ποσά που είχε δαπανήσει η Αμερικανική Κυβέρνηση δεν είχαν οδηγήσει στα επιθυμητά αποτελέσματα, δεδομένου ότι τα συστήματα που δημιουργήθηκαν ήταν αργά και παρήγαγαν μεταφράσεις χαμηλής ποιότητας.

Ο λόγος της αποτυχίας τους συνοψίζεται στις υψηλές προσδοκίες που είχαν δημιουργηθεί, ενώ ταυτόχρονα δεν υπήρχε το βασικό θεωρητικό υπόβαθρο και η κατάλληλη τεχνική υποδομή. Η αναφορά καταλήγει στο συμπέρασμα ότι η έρευνα θα πρέπει να εστιαστεί στην επεξεργασία της φυσικής γλώσσας (Natural Language Processing – NLP) και να αποδεσμευτεί από εφαρμογές που σχετίζονται αποκλειστικά με τη μετάφραση. Η αναφορά της ALPAC οδήγησε στη μείωση των ερευνητικών κονδυλίων στον τομέα της ΜΜ στις ΗΠΑ αλλά και σε άλλες χώρες τουλάχιστον για μια δεκαετία, όμως ταυτόχρονα αυξήθηκαν οι προσπάθειες σε τομείς όπως η υπολογιστική γλωσσολογία και η τεχνητή νοημοσύνη, τομείς που αποτέλεσαν θεωρητικό υπόβαθρο για τα σύγχρονα συστήματα.

Ένα από τα ελάχιστα συστήματα ΜΜ που επιβίωσαν από την κρίση που επέφερε η αναφορά ήταν το Systran (1970), το οποίο αναπτύχθηκε από μια εταιρεία που εργαζόταν για την αμερικανική πολεμική αεροπορία, προκειμένου να μεταφράζει μεγάλο όγκο ρωσικών επιστημονικών και τεχνικών εγγράφων κατά τη διάρκεια του ψυχρού πολέμου. Οι μεταφράσεις ήταν συνήθως κατά προσέγγιση, αλλά ικανοποιητικές για μια πρόχειρη κατανόηση του περιεχομένου. Η επιτυχία του συστήματος οδήγησε την Επιτροπή των Ευρωπαϊκών Κοινοτήτων

να αγοράσει μια άγγλο-γαλλική έκδοση του συστήματος, την οποία ακολούθησαν συστήματα για τη μετάφραση των περισσότερων ζευγών γλωσσών της ΕΕ.

Κυρίως μετά το 1970, χρησιμοποιήθηκαν ευρέως συστήματα ΜΜ σε διεθνείς οργανισμούς και ιδιωτικές εταιρείες, ενώ στη δεκαετία του ’80 η έρευνα για τη ΜΜ γνωρίζει νέα άνθηση, κυρίως στον Καναδά και στην ΕΕ, αφού εκεί η ανάγκη για μετάφραση ήταν μεγαλύτερη λόγω της ύπαρξης πολλών επίσημων γλωσσών. Οι νέες προσπάθειες, επηρεασμένες από την αναφορά της ALPAC, διερευνούν νέες τεχνικές οι οποίες αξιοποιούν τις γλωσσολογικές θεωρίες της εποχής και τα υπολογιστικά εργαλεία που είχαν δημιουργηθεί.

Η αδυναμία των συστημάτων άμεσης μετάφρασης να κάνουν σωστές λεξικές και συντακτικές επιλογές, όπως αποτυπώθηκε στην αναφορά της ALPAC, οδήγησε την έρευνα προς τις έμμεσες προσεγγίσεις. Τα συστήματα της νέας αυτής γενιάς, τα οποία ανήκουν στη γενικότερη κατηγορία συστημάτων που βασίζονται σε κανόνες, ακολουθούσαν μια στρατηγική επεξεργασίας του κειμένου σε τρία στάδια: ανάλυση, μεταφορά, παραγωγή.

Ένα από τα πρώτα επιτυχημένα συστήματα μεταφοράς αναπτύχθηκε στο Μόντρεαλ. Το σύστημα μετάφρασης μετεωρολογικών δελτίων Météo, κάνοντας χρήση ενός περιορισμένου λεξιλογίου και συντακτικών κανόνων, λειτουργεί με επιτυχία από το 1976 μέχρι και σήμερα. Επιπλέον, το ενδιαφέρον στρέφεται πλέον περισσότερο προς την εξερεύνηση μεθόδων γλωσσικής ανάλυσης και σύνθεσης και προς τη μηχανική μετάφραση με ανθρώπινη υποστήριξη (HAMT).

### 1.3 Η μετάφραση ως επικοινωνία και μνήμη

Όπως όλα τα πολιτιστικά φαινόμενα, ο όρος «μετάφραση» δεν επιδέχεται μία μόνο ερμηνεία και δεν ορίζεται μονοσήμαντα. Ολόκληρη η μεταφρασεολογία - το πεδίο του θεωρητικού αναστοχασμού της μετάφρασης, διεπιστημονικό και διαρκώς εξελισσόμενο, σε διάλογο με τις ανθρωπιστικές, κοινωνικές αλλά και τις θετικές επιστήμες – προσπαθεί να ορίσει τη μετάφραση.

Ασχέτως των ποικίλων ορισμών, η «πολιτισμική στροφή στη μετάφραση» έχει αναδείξει, εδώ και πάνω από τριάντα χρόνια, τον σύνθετο χαρακτήρα της μεταφραστικής πράξης: στη μετάφραση δεν εμπλέκονται απλώς δύο γλώσσες αλλά δύο πολιτισμοί, που

συναντιούνται στο πρόσωπο (στον λόγο, για να ακριβολογούμε) ενός διαπολιτισμικού μεσολαβητή, του μεταφραστή (Pym 1997). Αυτή η πράξη είναι μια πράξη επικοινωνίας (Hatim & Mason 1997, Lotman 2000).

Η μετάφραση βρίσκεται διαχρονικά στην καρδιά της διαπολιτισμικής επικοινωνίας και συνδέεται ιστορικά και εξ ορισμού με τις τεχνολογίες γραφής-επικοινωνίας. Στην εποχή μας αυτές οι τεχνολογίες ονομάζονται Τεχνολογίες της Πληροφορίας και της Επικοινωνίας (ΤΠΕ) ή αλλιώς Νέες Τεχνολογίες, και στηρίζονται στον υπολογιστή και στα δίκτυα. Η μεταφραστική τεχνολογία ή τα μεταφραστικά εργαλεία, αφορούν τη μετάφραση, αλλά προκύπτουν ως εφαρμογές της Γλωσσικής Τεχνολογίας (Language Technology–Human Language Technology), δηλαδή των υπολογιστικών εφαρμογών επεξεργασίας της φυσικής γλώσσας, γραπτής και προφορικής. Ως εκ τούτου η μεταφραστική τεχνολογία ενσωματώνει, διαρκώς σε ευρύτερες τεχνολογικές εξελίξεις.

Η μεταφραστική τεχνολογία ξεκίνησε με την Αυτόματη μετάφραση (Traduction automatique) (Μηχανική μετάφραση [Machine Translation] στον αγγλοσαξονικό χώρο), το 1949, με το υπόμνημα"Translation" του Warren Weaver. Ως μηχανική μετάφραση οριζόταν η διαγλωσσική μετάφραση χωρίς ανθρώπινη παρέμβαση, με βάση αλγορίθμους – ένα σχέδιο που πολύ νωρίς αποδείχτηκε ανέφικτο (έκθεση ALPAC 1966 και αναστολή της χρηματοδότησης του έργου).

Η Αυτόματη Μετάφραση Υψηλής Ποιότητας (Fully Automatic High Quality Machine Translation, FAHQMT) οδηγεί στην ανάπτυξη της Μετάφρασης με τη βοήθεια υπολογιστή (Computer Assisted / Aided Translation, CAT), στην εστίαση δηλαδή σε εργαλεία που μπορούν να επιταχύνουν και να διευκολύνουν τη μεταφραστική εργασία μέσα από τη διάδραση ανθρώπου και υπολογιστή.

Η πρόοδος της υπήρξε ραγδαία μετά από τη διάδοση του προσωπικού υπολογιστή. Σε άλλες χώρες, όπως η Γαλλία, η Γερμανία, η Ρωσία αλλά και το Μεξικό και η Ουγγαρία), οι επιστήμονες, σε αντίθεση με τους Αμερικανούς που είχαν επενδύσει αποκλειστικά στην Αυτόματη μετάφραση, ασχολούνταν από τη δεκαετία του 1950 με εφαρμογές γλωσσικής τεχνολογίας στη λεξικογραφία και την ορολογία (Hutchins1978, Léon 2001, Δημητρούλια 2005).

# Κεφάλαιο 2: Μηχανική μετάφραση

## 2.1 Συστήματα μηχανικής μετάφρασης

Η Μηχανική Μετάφραση (Machine Translation) (MT) ορίζεται από τους Hutchins και Somers ως «*ηλεκτρονικά συστήματα υπεύθυνα για την παραγωγή μετάφρασης από μια φυσική γλώσσα σε μια άλλη, με ή χωρίς ανθρώπινη βοήθεια*» (Hutchins & Somers, 1992). Η ανάπτυξη της MT χαρακτηρίζεται από πολλές μεταπτώσεις. Η πρώτη ιδέα της MT μπορεί να χρονολογηθεί από το 1903 όταν ο Couturat και ο Leau πρότειναν τον όρο "ein mechanisches Uebersetzen", που είναι η γερμανική ονομασία της "μηχανικής μετάφρασης".

Πριν από το 1960, η ανάπτυξη της ΜΤ ήταν αρκετά αισιόδοξη και οι ερευνητές πίστευαν ότι η ΜΤ θα μπορούσε να παράγει «πλήρως αυτόματη υψηλής ποιότητας μετάφραση (fully automatic high - quality translation) (FAHQT)». Ωστόσο, το 1966, μια έκθεση της ALPAC (Automatic Language Processing Advisory Committee, Συμβουλευτική Επιτροπή για την Επεξεργασία της Αυτόματης Γλώσσας) δήλωσε ότι η έννοια της FAHQT ήταν μη ρεαλιστική. Αν και η έκθεση θεωρήθηκε ευρέως βραχυπρόθεσμη αργότερα, είχε μεγάλη αρνητική επίδραση στις έρευνες της ΜΤ εκείνη την εποχή.

Στη συνέχεια, ορισμένοι ερευνητές επιμένουν στην ανάπτυξη του συστήματος ΜΤ με πιο ρεαλιστικό στόχο, ενώ άλλοι αναζητούν εναλλακτικούς τρόπους επίλυσης του προβλήματος που προκαλείται από την αντίθεση μεταξύ ζήτησης και προσφοράς της μεταφραστικής υπηρεσίας.

## 2.2 Σύστημα μεταφραστικής μνήμης

Μια από τις πιο ελκυστικές εναλλακτικές λύσεις που χρησιμοποιείται ευρέως τα τελευταία χρόνια είναι το Σύστημα Μεταφραστικής Μνήμης (Translation Memory System) (TMS). Το TMS είναι ένα εργαλείο μετάφρασης με τη βοήθεια υπολογιστή «*σχεδιασμένο για να αυξήσει την παραγωγικότητα της μετάφρασης με την αυτοματοποίηση της γλωσσικής μεταφοράς από την πηγή στο κείμενο-στόχο*» (Austermuhl, 2001). Είναι σε θέση να βοηθήσει τους μεταφραστές της ανθρώπινης γλώσσας επαναλαμβάνοντας τις προηγούμενες μεταφράσεις τους. Υπάρχουν πολλές διαφορές μεταξύ της ΜΤ και του TMS.

Σύμφωνα με τον Liang Sanyun (2004), οι διαφορές μεταξύ της ΜΤ και του TMS μπορούν να εξηγηθούν σε τουλάχιστον τρία μέρη. Το πρώτο και το πιο σημαντικό είναι ότι έχουν σχεδιαστεί για διαφορετικούς σκοπούς. Η ΜΤ αναπτύσσεται για να αντικαταστήσει την ανθρώπινη μετάφραση με τη λειτουργία μηχανήματος, με στόχο την παροχή γλωσσικής βοήθειας στους χρήστες που δεν είναι καλοί σε κάποια ξένη γλώσσα, ενώ το TMS είναι ένα εργαλείο σχεδιασμένο για επαγγελματίες μεταφραστές, με στόχο τη βελτίωση της αποτελεσματικότητας του έργου τους και της ποιότητας των μεταφρασμένων κειμένων.

Η δεύτερη διαφορά έγκειται στο πεδίο εφαρμογής τους. Έχει αναγνωριστεί γενικά ότι η ΜΤ πρέπει να χρησιμοποιείται μόνο για κείμενα από κάποιους περιορισμένους τομείς, ειδικά για μη λογοτεχνικά κείμενα. Αντίθετα, η εφαρμογή του TMS δεν έχει τέτοιο περιορισμό, δεδομένου

ότι παρέχει μόνο βοήθεια στους μεταφραστές και οι μεταφραστές εξακολουθούν να διαδραματίζουν ηγετικό ρόλο στη διαδικασία μετάφρασης. Η τελευταία αλλά όχι λιγότερο σημαντική είναι ότι έχουν διαφορετικές μηχανές μετάφρασης. Ως πλήρως αυτόματη συσκευή μετάφρασης, η MT έχει υψηλή απαίτηση τόσο για την τράπεζα γραμματικής όσο και για την τράπεζα γνώσεων της μηχανής μετάφρασης. Σε σύγκριση με την MT, ωστόσο, το TMS είναι μόνο ημιαυτόματο, καθώς εστιάζει μόνο στην αυτόματη αντιστοίχιση μεταξύ του νέου αποσπάσματος και του αποθηκευμένου αποσπάσματος.

Επομένως, η αποθήκευση μεγάλης ποσότητας μεταφρασμένου υλικού είναι πιο σημαντική από την ικανότητα γραμματικής για τη μηχανή μετάφρασης του TMS. Από την παραπάνω σύγκριση, μπορεί να φανεί σαφώς ότι η πιο σημαντική διαφορά μεταξύ της MT και του TMS είναι ότι η πρώτη είναι μηχανοκεντρική, ενώ το δεύτερο είναι ανθρωποκεντρικό.

Η δύναμη του TMS έγκειται στο ότι παρέχει καλύτερη συνεργασία με τους ανθρώπινους μεταφραστές και ενσωματώνει τις δυνάμεις των υπολογιστών με εκείνες των ανθρώπων. Με την ανακύκλωση των υφιστάμενων μεταφράσεων, είναι σε θέση να μειώσει το κόστος και το χρόνο της μετάφρασης καθώς και να βελτιώσει την ποιότητα και τη συνοχή του μεταφρασμένου κειμένου. Ως αποτέλεσμα, το TMS έχει μια καλύτερη πρακτική χρήση από την MT και είναι σήμερα ένα από τα πιο δημοφιλή εργαλεία CAT που χρησιμοποιούνται μεταξύ επαγγελματιών μεταφραστών. Ωστόσο, η αποδοχή του TMS δεν σημαίνει απαραίτητα ότι δεν υπάρχει πρόβλημα στην εφαρμογή του TMS. Παρόλο που το TMS δεν πραγματοποιεί πραγματική μετάφραση και είναι απλώς μια τράπεζα προ-μεταφρασμένων υλικών, είναι πιθανόν ο μεταφραστής και το μεταφρασμένο κείμενο να επηρεάζονται αρνητικά από τη λειτουργία του TMS.

Ένα πιθανό πρόβλημα είναι ότι ο μεταφραστής βασίζεται πάρα πολύ στο TMS και πιστεύει ό, τι προτείνεται από το TMS, ακόμη κι αν η ίδια η πρόταση μετάφρασης είναι λάθος ή ανακριβής σε ορισμένες πτυχές. Ως αποτέλεσμα, η λανθασμένη ή ανακριβής μετάφραση μπορεί να επαναλαμβάνεται και επιπλέον να ανακυκλώνεται στο μεταφρασμένο κείμενο. Ταυτόχρονα, η δημιουργικότητα του μεταφραστή μπορεί επίσης να επηρεαστεί.

Οι έρευνες σχετικά με τον περιορισμό των μελετών μετάφρασης με βάση σύνολο κειμένων μπορεί να παρέχουν ορισμένες θεωρητικές εξηγήσεις σχετικά με τη πιθανότητα αυτού του

προβλήματος. Ο Liu Kanglong (2006) αναφέρει τη γνώμη της Mona Baker στο άρθρο του, λέγοντας ότι οι ερευνητές τείνουν να προσελκύονται από το μεγάλο όγκο μεταφραστικών πληροφοριών και δεδομένων από ένα μεγάλης κλίμακας σύνολο κειμένων, κάνοντας απλά κλικ στο ποντίκι. Ως εκ τούτου, μπορεί να επικεντρωθούν περισσότερο στο Μεταφραστικό Πρότυπο, αλλά να παραμελούν τη δημιουργικότητα των μεταφραστών και ακόμη να θεωρούν μια δημιουργική έκφραση ως λάθος.

Ο Liu Jingguo (2006) αναφέρει επίσης την αδυναμία των μελετών με βάση το σύνολο κειμένων. Επισημαίνει ότι οι μελέτες με βάση το σύνολο κειμένων έχουν σαφή τάση επιστημονισμού, εστιάζοντας πάρα πολύ στην αντικειμενικότητα και τείνοντας να περιορίσουν τη δημιουργικότητα των ανθρώπων. Ως εκ τούτου, μπορεί να οδηγήσει τις μελέτες μετάφρασης σε μια άψυχη κατάσταση και ακόμη και σε παύση. Οι παραπάνω ιδέες του Liu Kanglong και του Liu Jingguo βοηθούν στην κατανόηση του δυνητικού προβλήματος μεταξύ του μεταφραστή και του TMS. Από την παραπάνω εισαγωγή, μπορεί να φανεί καθαρά ότι το TMS είναι στη φύση του ένα ευθυγραμμισμένο σύνολο κειμένων.

Συμπερασματικά, κατά τη χρήση του TMS ο μεταφραστής μπορεί επίσης να έχει την τάση να προσελκύεται από τις πληροφορίες που αποθηκεύονται σε αυτό και να παραμελεί τη δική του δημιουργικότητα, ακόμη και την κρίση του. Δηλαδή, μπορεί να τείνει να εμπιστεύεται και να βασίζεται σε όλες τις προτάσεις μετάφρασης που παρέχονται από το TMS χωρίς ή με λίγη περαιτέρω σκέψη.

Εάν υπάρχουν κάποια λάθη στη βάση δεδομένων TM, είναι επίσης πιθανό να μην βρεθούν και απλά να υιοθετούνται στο μεταφρασμένο κείμενο, ως συνήθως. Εάν υπάρχει τέτοιο πρόβλημα, μπορεί να υπάρχει κίνδυνος να επαναληφθούν αυτά τα λάθη και να ανακυκλωθούν στο TM και στο μεταφρασμένο κείμενο και συνεπώς να προκαλέσουν ζημιές σε μεγαλύτερη κλίμακα. Λαμβάνοντας υπόψη τις παραπάνω παραδοχές, ο συγγραφέας αυτής της διατριβής προσπαθεί να χρησιμοποιήσει ένα αυτοσχέδιο πείραμα και ένα ερωτηματολόγιο για να αποδείξει την ύπαρξή τους, με στόχο να διαπιστώσει εάν τα λάθη στο TMS θα οδηγήσουν στην επανάληψη των λαθών στο μεταφρασμένο κείμενο και ποιες είναι οι στάσεις των μεταφραστών κατά τη διάρκεια αυτής της διαδικασίας. Το αποτέλεσμα της έρευνας αυτής μπορεί να έχει κάποια σημασία είτε για την εκπαίδευση μετάφρασης, την υπηρεσία μετάφρασης, είτε για την ανάπτυξη του TMS.

Όσον αφορά την εκπαίδευση μετάφρασης, το αποτέλεσμα μπορεί να συμβάλει στη βελτίωση της αυτεπίγνωσης των μεταφραστών σπουδαστών στη διαδικασία μετάφρασης με το TMS, έτσι ώστε να τους βοηθήσει να αποφύγουν την πλήρη εμπιστοσύνη στις πληροφορίες που αποθηκεύονται στο ΤΜ και να διατηρήσουν την ανεξαρτησία τους στην ανάλυση και κρίση.

Όσον αφορά την υπηρεσία μετάφρασης, το αποτέλεσμα αυτής της έρευνας μπορεί να υπογραμμίσει τη σημασία της διαχείρισης της βάσης δεδομένων ΤΜ, είτε στις μεταφραστικές εταιρείες είτε στον πελάτη της μεταφραστικής υπηρεσίας. Και για την ανάπτυξη του TMS, το αποτέλεσμα αυτής της έρευνας μπορεί να προσφέρει κάποια νέα προοπτική βελτίωσης ειδικά σε ό, τι αφορά τη διαχείριση του ΤΜ. (Wu&Pan, 2013).

Στα πλαίσια της επικοινωνίας η μετάφραση θεωρείται πολύ χρήσιμο εργαλείο και διαδραματίζει έναν ουσιαστικό ρόλο. Ιδιαίτερα τα τελευταία χρόνια εξαιτίας της παγκοσμιοποίησης και της διεθνοποίησης της αγοράς, η ανάγκη για τη μετάφραση κειμένων μεγαλώνει καθημερινά. Ωστόσο, η διαδικασία της μετάφρασης είναι χρονοβόρα και αυτό έρχεται σε αντίθεση με την ανάγκη για αύξηση του όγκου των μεταφράσεων.

Είναι αυτή η ανάγκη που ώθησε τη μεταφραστική βιομηχανία να στραφεί σε εναλλακτικούς, πέραν του παραδοσιακού, τρόπους μετάφρασης. Συγκεκριμένα στη μηχανική ή αυτόματη μετάφραση, κυρίως σε είδη κειμένων όπως τα τεχνικά, με σκοπό την εξοικονόμηση χρόνου και χρημάτων καθώς και την εξασφάλιση μεγαλύτερου όγκου μεταφράσεων. Ωστόσο, τα μεταφραστικά αποτελέσματα που προέρχονται από τη χρήση συστημάτων μηχανικής μετάφρασης, παρουσιάζουν πληθώρα προβλημάτων σε επίπεδο συντακτικό, γραμματικό και λεξιλογικό. Η δημιουργία των ελεγχόμενων γλωσσών (controlled languages) και η χρήση των κανόνων που τις χαρακτηρίζουν, σε συνδυασμό πάντα με τη μηχανική μετάφραση επιφέρουν κατά ένα μεγάλο ποσοστό τη βελτίωση του μεταφραστικού αποτελέσματος.

Κατά τις τελευταίες δεκαετίες γλωσσολόγοι, μεταφραστές και άλλοι επιστήμονες πραγματοποίησαν έρευνες, σχετικές με τις δυνατότητες της μηχανικής μετάφρασης - με τον όρο μηχανική μετάφραση εννοείται η μετάφραση κειμένων από μια γλώσσα σε μια άλλη με τη βοήθεια ηλεκτρονικού υπολογιστή και προγραμμάτων επεξεργασίας κειμένων χωρίς την παρέμβαση ανθρώπινου παράγοντα. Η εισαγωγή του κειμένου γίνεται με τη βοήθεια ανθρώπου αλλά η

διαδικασία της μετάφρασης γίνεται αποκλειστικά και μόνο από τον υπολογιστή - και αυτή η συστηματική μελέτη οδήγησε στη δημιουργία συστημάτων μηχανικής μετάφρασης.

Ωστόσο, η προσπάθεια που κατέβαλαν οι επιστήμονες σε σχέση με τα αποτελέσματα που επιτεύχθηκαν με τη χρήση της μηχανικής μετάφρασης δεν είναι ανάλογα. Είναι γεγονός ότι έχει σημειωθεί πρόοδος στον τομέα αυτό, αλλά δεν πραγματοποιήθηκε σε τόσο γρήγορο ρυθμό όσο αναδενόταν. Τα αποτελέσματα της μηχανικής μετάφρασης είναι φτωχά σε ποιότητα, και σε επίπεδο ακρίβειας (accuracy) και σε επίπεδο κατανόησης κειμένου (intelligibility).

Η χρήση της Μηχανικής Μετάφρασης, είναι παρεξηγημένη ακόμη και από ανθρώπους οι οποίοι ανήκουν στο χώρο της μεταφραστικής βιομηχανίας. Πολλοί χρήστες, ειδικά οι χρήστες του Διαδικτύου, αγνοούν το τί ακριβώς είναι η Μηχανική Μετάφραση και ποιος είναι ο ρόλος της. Το τελευταίο επιφέρει τη δημιουργία λανθασμένων αντιλήψεων γύρω από τον τομέα αυτό.

Η μηχανική μετάφραση δε φτιάχτηκε για να μεταφράζει Σαίξπηρ όπως αναφέρει ο Arnold το 1994. Την ίδια στιγμή, η Μηχανική Μετάφραση προκαλεί φόβο σε μια μερίδα μεταφραστών οι οποίοι θεωρούν ότι ελλοχεύετε ο κίνδυνος να χάσουν τη δουλειά τους. Αυτό που είναι σημαντικό, είναι να γίνει κατανοητό ότι με τον όρο Μηχανική Μετάφραση δεν εννοείται το πάτημα ενός κουπιού του υπολογιστή και η παραγωγή ενός άψογα μμεταφρασμένου κειμένου. Σε καμία περίπτωση όμως δε θα μπορούσε οποιοσδήποτε να πει ότι από τα προβληματικά αποτελέσματα που έχουμε σήμερα προκύπτει το συμπέρασμα ότι ο τομέας της Μηχανικής Μετάφρασης είναι άχρηστος. Πολλά συστήματα όπως το METEO που χρησιμοποιεί το Canadian Meteorological Center ήδη από το 1977, το SYSTRAN και το LOGOS, αποδεικνύουν το αντίθετο (Arnold, 1994).

Η αυτόματη ή μηχανική μετάφραση αποτελεί έναν χώρο συνεχούς έρευνας και προσπάθειας εφαρμογών λόγω της άμεσης χρησιμότητας και της αυξημένης ζήτησης στην διεθνή αγορά, αυτού του τύπου προϊόντων. Είναι σημαντικό όμως για την καλύτερη προσέγγιση/αξιολόγηση αυτού του είδους εργαλείων, να γίνει σαφής ο διαχωρισμός ανάμεσα στις περιπτώσεις μετάφρασης που χειρίζεται ένας φυσικός μεταφραστής και σε εκείνες που αναμένεται να χειρισθεί ένα αυτόματο σύστημα επεξεργασίας. Αν σκεφθούμε την φυσική γλώσσα σαν βιολογικό σύστημα που χαρακτηρίζει τον άνθρωπο και απλουστευτικά την ονομάσουμε "Σύνολο

Γ", τότε αν απομονώσουμε από αυτό το "υποσύνολο γ", των γραμματικών/συντακτικών φαινομένων και των στοιχείων του λεξικού που χρησιμοποιεί ένας περιορισμένος θεματικά χώρος για να καλύψει τις γλωσσικές ανάγκες του (έκφραση , πληροφορία, επικοινωνία κ.λπ.), φθάνουμε στην έννοια της Υπογλώσσας (Sublanguage).

Η έννοια Υπογλώσσα θα κυριαρχήσει σε ότι θα ακολουθήσει, δεδομένου ότι οι πλέον επιτυχημένες εφαρμογές στον χώρο της Επεξεργασίας Φυσικής Γλώσσας, όπως αυτός είναι ευρύτερα γνωστός, αφορούν στην ανάλυση ή σύνθεση κειμένων ορισμένης θεματολογίας κάθε φορά. Γίνεται δε, εφικτή η απομόνωση και η συμμετοχή στα εργαλεία (modules) ανάλυσης/ σύνθεσης, των στοιχείων του λεξικού και της σύνταξης που αποτελούν την υπογλώσσα του συγκεκριμένου χώρου, στον μέγιστο δυνατό βαθμό αλγοριθμικής τους κάλυψης.

Όταν τα κείμενα μίας υπογλώσσας αποτελούν αντικείμενο αυτόματης μετάφρασης από μία αρχική γλώσσα σε μία γλώσσα στόχο, τα στάδια επεξεργασίας στον ΗΥ καθορίζονται συνήθως ως εξής:

- ✓ Στάδιο Ανάλυσης: περιλαμβάνει τα υποσύνολα της γραμματικής (σύνολο κανόνων συντακτικής αναγνώρισης φυσικής γλώσσας από τον ΗΥ) και του λεξικού που απαιτούνται για την αναγνώριση ανάλυση του συνόλου των προτάσεων της, προς μετάφραση, αρχικής γλώσσας.
- ✓ Στάδιο Μετάφρασης: επιτυγχάνει την ταύτιση των δομών της αρχικής γλώσσας με τις αντίστοιχες δομές στην γλώσσα στόχο και περιλαμβάνει το σύνολο των λημμάτων του απαιτούμενου δίγλωσσου λεξικού, καθώς και τους κανόνες εκείνους που εξομαλύνουν τις δομικές διαφορές ανάμεσα στις δύο γλώσσες.
- ✓ Στάδιο Σύνθεσης: περιλαμβάνει το υποσύνολο της γραμματικής και του λεξικού που απαιτούνται για την σύνθεση του κειμένου στην γλώσσα στόχο.

Όμως δεν είναι μόνον η βάση δεδομένων των κοινών στοιχείων του λεξικού που συμμετέχει στις διαδικασίες επεξεργασίας των προτάσεων μίας υπογλώσσας. Πολύ σημαντική θέση κατέχουν τα στοιχεία ορολογίας που διέπουν την υπογλώσσα και ορίζουν τον χαρακτήρα του θεματικού χώρου της, καθώς και την διαφοροποίηση της από άλλες υπογλώσσες διαφορετικής θεματολογίας.

Τα στοιχεία ορολογίας μίας υπογλώσσας συγκεντρώνονται στην βάση δεδομένων ορολογίας, που συμμετέχει παράλληλα με την βάση δεδομένων κοινών στοιχείων του λεξικού στην επεξεργασία των προς ανάλυση ή σύνθεση προτάσεων. Αυτό όμως, που κυρίως διαφοροποιεί μια βάση δεδομένων ορολογίας από μία βάση δεδομένων κοινών στοιχείων του λεξικού, είναι ότι ενώ στην περίπτωση του κοινού λεξικού, στην βάση δεδομένων δεν περιέχεται ποτέ συντακτική κατηγορία μεγαλύτερη από την μηδενική κατηγορία που αντιστοιχεί στη λέξη. Μια βάση δεδομένων ορολογίας μπορεί να περιέχει όχι μόνο μονολεκτικούς όρους, αλλά και όρους που αντιστοιχούν σε μεγαλύτερες συντακτικά, φραστικές κατηγορίες, χωρίς να είναι τυπικά δυνατό να περιορισθεί το είδος ή το μέγεθος των συντακτικών κατηγοριών που θα μπορούσαν, εν δυνάμει, να χαρακτηρισθούν όροι μίας υπογλώσσας (Ευθυμίου, 1992).

## 2.3 Τι είναι η προεπεξεργασία

Το πρώτο βήμα της μεταφραστικής διαδικασίας είναι η προεπεξεργασία τόσο της πρότασης της γλώσσας-πηγής όσο και του μονόγλωσσου σώματος κειμένων της γλώσσας στόχου.

Η προεπεξεργασία περιλαμβάνει βασικές εργασίες επεξεργασίας της φυσικής γλώσσας, όπως είναι η γραμματική επισημείωση (tagging), η λημματοποίηση (lemmatization) και η επιφανειακή συντακτική ανάλυση και κατάτμηση (chunking). Η προεπεξεργασία του μονόγλωσσου σώματος κειμένων της γλώσσας-στόχου γίνεται μια φορά, κατά τη δημιουργία της βάσης δεδομένων.

Στο σύστημα METIS-II χρησιμοποιήθηκε το σώμα κειμένων BNC8 (British National Corpus), το οποίο χρησιμοποιεί CLAWS5 για τη γραμματική επισημείωση των όρων του. Για την προεπεξεργασία του BNC έγινε αρχικά λημματοποίηση και γραμματική επισημείωση των όρων με κατάλληλο εργαλείο στα Αγγλικά (Carletal., 2005).

Η προεπεξεργασία της πρότασης στη γλώσσα-πηγή είναι η πρώτη εργασία που εκτελείται όταν δίνεται η πρόταση στο σύστημα για μετάφραση. Σε αυτή την περίπτωση η προεπεξεργασία περιλαμβάνει τη λημματοποίηση και γραμματική επισημείωση των όρων με βάση το σύνολο γραμματικών ετικετών του PAROLE (Labropoulou et al., 1996).

Στη συνέχεια, η πρόταση περνά από το δίγλωσσο λεξικό, το οποίο μας δίνει τις μεταφράσεις στη γλώσσα-στόχο κάθε λέξης της πρότασης (λήμμα και γραμματική ετικέτα). Οι μεταφράσεις είναι όλες ισοπίθανες, αφού το λεξικό που χρησιμοποιείται δεν περιέχει καθόλου πιθανότητες. Κατόπιν, η αρχική πρόταση μετατρέπεται σε ένα διάνυσμα το οποίο περιέχει γραμματική και συντακτική πληροφορία στη γλώσσα-πηγή, αλλά και τις μμεταφράσεις των λέξεων στη γλώσσα-στόχος.

## 2.4 Συστήματα που βασίζονται σε κανόνες

Τα συστήματα που βασίζονται σε κανόνες (Rule-Based Machine Translation - RBMT) κάνουν χρήση γλωσσολογικών κανόνων σε διάφορα επίπεδα για τη μετάφραση από μια γλώσσα πηγή σε μια γλώσσα στόχο.

Αποτελούν την πολυπληθέστερη και ωριμότερη κατηγορία συστημάτων που κυκλοφορούν στο εμπόριο, αφού είναι η παλαιότερη, και συνεχίζουν να ασκούν επιρροή ακόμη και στα σύγχρονα συστήματα ΜΜ. Αν και κάποια από αυτά, όπως το Systran, έχουν επιβιώσει μέχρι σήμερα, η πλειοψηφία των ερευνητών θεωρεί ότι η χρήση γλωσσολογικών κανόνων κρύβει πολλούς κινδύνους, διότι για κάθε μεταφραστικό φαινόμενο που μπορεί να συναντήσουμε, πρέπει να δημιουργηθεί ένας κανόνας για να το αντιμετωπίσει.

Όμως, η εκτεταμένη χρήση γλωσσολογικών κανόνων μπορεί να οδηγήσει σε μια έκρηξη του αριθμού τους σε τέτοιο βαθμό ώστε να μην είναι διαχειρίσιμα. Με την προσθήκη ενός νέου κανόνα μπορεί να διαταραχθεί το σύστημα και να οδηγήσει σε απροσδιόριστα αποτελέσματα. Ένα ακόμη πρόβλημα των συστημάτων αυτών είναι η αλληλεπίδραση των κανόνων, οπότε και πρέπει να ορίσουμε ποιος κανόνας έχει προτεραιότητα απέναντι σε ποιον. (Σοφιανόπουλος 2009).

Σχηματικά, τα συστήματα ΜΜ που βασίζονται σε κανόνες ακολουθούν μια στρατηγική επεξεργασίας του κειμένου σε τρία στάδια: ανάλυση, μεταφορά, παραγωγή.

- ✓ Στο στάδιο της ανάλυσης, το κείμενο της γλώσσας πηγής υφίσταται κατάτμηση σε μικρότερες μονάδες, αναλύεται και μετατρέπεται, τις περισσότερες φορές, σε μια ενδιάμεση αναπαράσταση με τη χρήση γλωσσολογικών εργαλείων στη γλώσσα πηγή.

- ✓ Στο στάδιο της μεταφοράς, γίνεται αντιστοίχιση της αναπαράστασης του προηγούμενου βήματος σε αναπαράσταση στη γλώσσα στόχο. Το σύστημα αναζητάει στους γλωσσικούς πόρους που διαθέτει (λεξικά, γραμματικές) τις κατάλληλες ισοδυναμίες μεταξύ των δύο γλωσσών για τις μονάδες που προέκυψαν κατά το πρώτο στάδιο (λεξικές μονάδες, συντακτικές δομές).
- ✓ Στο στάδιο της παραγωγής, εξάγεται η τελική μετάφραση στη γλώσσα στόχο, αφού το σύστημα οργανώσει τις πληροφορίες που έχει συγκεντρώσει σχετικά με τις λεξικές ισοδυναμίες και τις συντακτικές αντιστοιχίες προκειμένου να συνθέσει ένα κείμενο-στόχο που να είναι, κατά το δυνατόν, σημασιολογικά ισοδύναμο και κατανοητό και συντακτικά αποδεκτό.

Τα περισσότερα συστήματα με κανόνες βασίζονται στην περιγραφή κανόνων για συγκεκριμένα ζευγάρια γλωσσών (language-dependent). Στην κατηγορία αυτή ανήκουν τα λεξικολογικά συστήματα και τα συστήματα μεταφοράς.

## 2.5 Τα λεξικολογικά συστήματα (dictionary-based)

Θυμίζουν αρκετά τα συστήματα άμεσης μετάφρασης, όπου η μετάφραση γίνεται λέξη προς λέξη, χωρίς να παρεμβάλλεται κανένα ενδιάμεσο στάδιο. Στις πιο εξελιγμένες εκδοχές τους, ο έλεγχος του λεξικού συνοδεύεται από μορφολογική ανάλυση και λημματοποίηση, καθώς και από κανόνες οι οποίοι συσχετίζουν λέξεις της γλώσσας πηγής με λέξεις της γλώσσας στόχου.

Αποτελεί το απλούστερο είδος συστημάτων με περιορισμένες δυνατότητες και μετριότατα αποτελέσματα (μη αποδεκτές λεξικές επιλογές και ισοδυναμίες, συντακτικά έκτυπα, γραμματικά λάθη, αδυναμία επίλυσης αμφισημιών κ,τ,λ,). Ωστόσο, παραμένει καλή λύση για συγκεκριμένες εφαρμογές, όπως η μετάφραση καταλόγων με προϊόντα και γενικά λιστών με σύντομες φράσεις.

### 2.5.1 Τα συστήματα μεταφοράς (transfer-based)

Τα συστήματα μεταφοράς αποτελούν μια από τις ευρύτερα διαδεδομένες τεχνικές ΜΜ και βασίζονται στο γεγονός ότι οι περισσότερες γλώσσες έχουν σημαντικές διαφορές στη δομή τους και για την αντιμετώπισή τους εφαρμόζονται κανόνες συντακτικής και λεξιλογικής μεταφοράς, οι

οποίοι δίνουν αναλυτικές οδηγίες για το πώς πρέπει να μετατραπεί η πληροφορία στη γλώσσα πηγή ώστε να προσεγγίζει ένα σωστό κείμενο στη γλώσσα στόχο.

Η διαδικασία μετάφρασης δεν είναι άμεση (λέξη προς λέξη) αλλά έμμεση, αφού μεσολαβεί μια ενδιάμεση φάση αναπαράστασης της δομής μιας πρότασης στη γλώσσα-πηγή και η μεταφορά της στην ισοδύναμη δομή στη γλώσσα-στόχο. Τα αποτελέσματα αυτών των συστημάτων ΜΜ είναι αρκετά βελτιωμένα σε σχέση με τα απλά λεξικολογικά συστήματα και τα συστήματα άμεσης μετάφρασης. Το σημαντικότερο όμως, μειονέκτημα αυτής της μεθόδου είναι ότι ένα τέτοιο σύστημα δεν μπορεί να εφαρμοστεί αυτούσιο σε οποιοδήποτε άλλο γλωσσικό ζεύγος, αφού όλοι οι κανόνες μεταφοράς εξαρτώνται αποκλειστικά από τις συγκεκριμένες γλώσσες πηγή και γλώσσες στόχο για τις οποίες δημιουργήθηκαν. Επιπλέον, η ποιότητα της μετάφρασης εξαρτάται από την ακρίβεια των συντακτικών και λεξικών κανόνων που χρησιμοποιούνται για τη μεταφορά από τη γλώσσα πηγή στη γλώσσα στόχο. Έτσι, πολλές φορές αδυνατούν να αντιμετωπίσουν ικανοποιητικά συντακτικές και λεξικές αμφισημίες, με αποτέλεσμα όσο πιο μεγάλη είναι μια πρόταση, τόσο πιο δυσνόητη να είναι η μετάφρασή της.

Η δυσκολία μεταφοράς και εφαρμογής των συστημάτων αυτών σε νέα γλωσσικά ζεύγη, αφού για κάθε ζεύγος γλωσσών χρειάζεται διακριτό σύστημα κανόνων οδήγησε στην ανάπτυξη συστημάτων που είναι ανεξάρτητα από συγκεκριμένες γλώσσες πηγή και στόχο (language-independent) και βασίζονται σε ένα ενδιάμεσο στάδιο αφηρημένης αναπαράστασης της γλώσσας.

<u>Στην κατηγορία αυτή ανήκουν:</u>

- ✓ Τα διαγλωσσικά συστήματα, όπου μεταξύ γλώσσας πηγής και γλώσσας στόχου παρεμβάλλεται η διαγλώσσα (intelingua), η οποία είναι μια αφηρημένη γλώσσα που μπορεί να περιγράψει τα χαρακτηριστικά όλων των γλωσσών προς μετάφραση ώστε στη συνέχεια να είναι δυνατή η μετατροπή και η ανασύσταση των μορφολογικών, συντακτικών και σημασιολογικών χαρακτηριστικών σε μια γλώσσα στόχο.
- ✓ Τα οντολογικά συστήματα, τα οποία βασίζονται στην αναπαράσταση της γνώσης του κόσμου μέσω της γλώσσας (knowledge-based), χρησιμοποιούν βάσεις γνώσεων και οντολογίες (δομημένη γνώση), επαγωγικούς μηχανισμούς, καθώς και τεχνικές της τεχνητής νοημοσύνης ώστε να παραγάγουν σωστές σημασιολογικές ισοδυναμίες. Θεωρητικά, τα δύο αυτά μοντέλα

θεωρούνται πιο οικονομικά εφόσον εξασφαλίζουν την επαναχρησιμοποίηση της ίδιας αφηρημένης αναπαράστασης για περισσότερους συνδυασμούς γλωσσών. Στην πράξη, όμως, παραμένουν αρκετά δύσκολα στον σχεδιασμό και στην υλοποίησή τους, με αποτέλεσμα να βρίσκουν εφαρμογή μόνο σε ερευνητικό και πειραματικό επίπεδο.

## 2.6 Η απλούστερη προσέγγιση στη ΜΜ

Η απλούστερη προσέγγιση στη ΜΜ είναι η αυτόματη αντικατάσταση των λέξεων ενός κειμένου γραμμένου σε μια φυσική γλώσσα με λέξεις μιας άλλης γλώσσας. Αυτό μπορεί να είναι χρήσιμο σε θεματικούς τομείς με περιορισμένη και τυποποιημένη γλώσσα, όπως π. χ. τα μετεωρολογικά δελτία. Ωστόσο, η καλή μετάφραση λιγότερο τυποποιημένων και μεγαλύτερων κειμένων (φράσεων, προτάσεων ή ακόμα και ολόκληρων αποσπασμάτων) απαιτεί την αντιστοίχιση με τα πιο κοντινά τους ισοδύναμα στη γλώσσα στόχο.

Η μεγαλύτερη δυσκολία εν προκειμένω έγκειται στο γεγονός ότι η ανθρώπινη γλώσσα είναι αμφίσημη, πράγμα το οποίο γεννά προκλήσεις σε πολλά απλά επίπεδα, π.χ. την αποσαφήνιση της σημασίας των λέξεων σε λεξιλογικό επίπεδο (η λέξη τζάγκουαρ μπορεί να αναφέρεται σε αυτοκίνητο ή ζώο) ή την προσάρτηση εμπρόθετων φράσεων σε συντακτικό επίπεδο όπως παρακάτω:

- Ο αστυνομικός παρακολουθεί τη γυναίκα με τα κιάλια.
- Ο αστυνομικός παρακολουθεί τη γυναίκα με το περίστροφο.

Ένας τρόπος ανάπτυξης συστημάτων ΜΜ βασίζεται στη χρήση γλωσσολογικών κανόνων. Για μεταφράσεις ανάμεσα σε συγγενικές γλώσσες, μια λέξη προς λέξη υποκατάσταση μπορεί να είναι εφικτή σε περιπτώσεις όπως το προηγούμενο παράδειγμα. Όμως, τα συστήματα που βασίζονται σε κανόνες (ή σε γλωσσολογική γνώση) συνήθως αναλύουν το εισερχόμενο κείμενο και δημιουργούν μια ενδιάμεση, συμβολική αναπαράσταση, από την οποία παράγεται το κείμενο στη γλώσσα στόχο. Η επιτυχία αυτών των μεθόδων εξαρτάται σε μεγάλο βαθμό από τη διαθεσιμότητα εκτεταμένων λεξικών με μορφολογική, συντακτική και σημασιολογική πληροφορία και μεγάλα σύνολα γραμματικών κανόνων καταρτισμένων από εξειδικευμένους γλωσσολόγους. Αυτή όμως είναι μια πολύ μακροχρόνια και επομένως δαπανηρή διαδικασία.

Στα τέλη της δεκαετίας του 1980, καθώς η υπολογιστική ισχύς αυξήθηκε και έγινε λιγότερο δαπανηρή, παρατηρήθηκε μεγαλύτερο ενδιαφέρον για στατιστικά μοντέλα ΜΜ. Τα στατιστικά μοντέλα προέρχονται από την ανάλυση δίγλωσσων σωμάτων παράλληλων κειμένων, όπως είναι για παράδειγμα το Europarl, το οποίο περιλαμβάνει τα πρακτικά του Ευρωπαϊκού Κοινοβουλίου σε 21 ευρωπαϊκές γλώσσες. Εάν υπάρχουν αρκετά δεδομένα, η στατιστική ΜΜ αποδίδει αρκετά καλά ως προς την παραγωγή νοήματος (κατά προσέγγιση) ενός ξενόγλωσσου κειμένου, μέσω της επεξεργασίας παράλληλων δεδομένων και της ανεύρεσης πιθανών αντιστοιχίσεων λέξεων.

Εντούτοις, σε αντίθεση με τα συστήματα που βασίζονται σε γλωσσολογική γνώση, η στατιστική (ή βασισμένη σε δεδομένα) ΜΜ συχνά παράγει γραμματικά εσφαλμένα αποτελέσματα. Η βασισμένη σε δεδομένα ΜΜ υπερτερεί στο ότι απαιτεί μικρότερη ανθρώπινη προσπάθεια και στο ότι έχει τη δυνατότητα να χειρίζεται ιδιαιτερότητες της γλώσσας (π.χ. ιδιωματισμούς) που τα συστήματα που βασίζονται σε γλωσσολογική γνώση πιθανόν να αγνοήσουν (Gavrilidouetal., 2012).

## Κεφάλαιο 3: Εκμάθηση της μηχανικής μετάφρασης

### 3.1 Διδάσκοντας μηχανική μετάφραση σε σπουδαστές της υπολογιστικής γλωσσολογίας

Αν μη τι άλλο, ο πρωταρχικός στόχος του μαθήματος ACL στην MT είναι να εξοπλίσει τους σπουδαστές με επαρκές σχετικό υλικό ώστε να μπορούν να μιλάνε ικανοποιητικά και με ακρίβεια για το CAT και την MT. Παρά το γεγονός ότι ο τομέας είναι σχετικά ώριμος στις μέρες μας, παραμένει λυπηρό το γεγονός ότι πολλή παραπληροφόρηση ακούγεται ακόμη σε συνέδρια και συναντάται στο Διαδίκτυο (Kenny&Way, 2001).

Η παλιά ιστορία ότι η MT αντικαθιστά τους μεταφραστές, ακούγεται ακόμα και δυστυχώς διδάσκεται στους σπουδαστές από διδάσκοντες, οι οποίοι θα έπρεπε είναι καλύτερα ενημερωμένοι. Το λογισμικό μετάφρασης είναι πλέον τόσο ευρέως διαθέσιμο, όσο ποτέ άλλοτε, αλλά οι προγραμματιστές συνεχίζουν να παραμορφώνουν τα προϊόντα τους με παραπλανητική διαφήμιση.

Ο D.Kenny και A.Way ως λέκτορες στον τομέα, έχουνε την ευθύνη να αναφέρουν με ακρίβεια την τεχνολογική εξέλιξη στον κλάδο, ώστε οι νεοεισερχόμενοι στον τομέα - ως μεταφραστές, γλωσσικοί μηχανικοί ή υποκινητές της γλωσσικής πολιτικής τόσο στη βιομηχανία, όσο και σε κυβερνητικό επίπεδο και πέραν αυτού - να μην έρχονται προ-οπλισμένοι με τις ψευδείς προσδοκίες που έχουν βλάψει τον κλάδο στο παρελθόν. Επίσης ελπίζουν πως το αποτέλεσμα των εργαστηρίων σπουδών (workshops) όπως αυτό, να διορθώσει τέτοιες λανθασμένες εντυπώσεις και να οδηγήσει σε βελτίωση της συνολικής αντίληψης του κλάδου της MT.

Όσοι από εμάς έχουμε αναπτύξει συστήματα MT και τα παρουσιάσαμε σε διάφορα φόρουμ, μπορούμε μόνο να ελπίζουμε ότι οι εποχές που κάποιος πληκτρολογούσε μια φράση 50 λέξεων αποτελούμενη από σειρές βοηθητικών, εμπρόθετων φράσεων και που περιείχε ελλείψεις, με το σύστημα είτε να μπερδεύεται, είτε να επιστρέφει μία άσχετη «μετάφραση» μετά από μερικά λεπτά, ακολουθούμενη από τον ελεγκτή του συστήματος μας να λέει "η MT δεν είναι για μένα!", έχουν περάσει εδώ και πολύ καιρό (Kenny&Way, 2001).

### 3.2 Μέσα επικοινωνίας μεταφραστή-πελάτη

Στην ψηφιακή μας εποχή, οι ηλεκτρονικές μορφές δεν αφορούν μόνο τα κείμενα μας, αλλά και τα μέσα επικοινωνίας μας με πελάτες και άλλους μεταφραστές. Λόγω του Διαδικτύου, επαγγελματίες από όλο τον κόσμο μπορούν να επικοινωνούν τακτικά μέσω του ηλεκτρονικού ταχυδρομείου ή με άλλες μορφές άμεσων μηνυμάτων. Οι εργασίες μπορούν να αποστέλλονται και να λαμβάνονται ηλεκτρονικά, σε εθνικό και διαπολιτισμικό πλαίσιο. Ωστόσο, το ανωτέρω επιφέρει συνέπειες.

Πρώτον, μπορούμε θεωρητικά να εργαστούμε για πελάτες οπουδήποτε στον κόσμο. Η αγορά μεταφράσεων δεν χρειάζεται να είναι η πόλη ή η χώρα μας. Ένα κείμενο (πηγή) που το λαμβάνει κάποιος στις 5 μ.μ. στην Tarragona μπορεί να αποσταλεί σε μεταφραστή στη Νέα Ζηλανδία, ο οποίος θα επιστρέψει τη μετάφραση πριν από τις 9 το επόμενο πρωί, ώρα Tarragona. Επομένως, οι ζώνες ώρας μπορούν να χρησιμοποιηθούν δημιουργικά, και έτσι οι εργασίες μπορούν να προέρχονται από εταιρείες που βρίσκονται πολύ μακριά. Το μόνο που έχουμε να κάνουμε είναι να καταγράψουμε το όνομά μας, τους συνδυασμούς γλωσσών και τους τομείς εξειδίκευσης σε μία από τις πολλές ιστοσελίδες που έχουν ως στόχο να φέρνουν τους μεταφραστές και τους πελάτες σε επαφή μεταξύ τους. Θα περίμενε κανείς ότι η διαδικασία αυτή θα οδηγήσει σε μια κατάσταση όπου τα ποσά που καταβάλλονται για μεταφράσεις θα γίνουν ουσιαστικά ίδια σε όλο τον κόσμο, σύμφωνα με τις θεωρίες μιας παγκόσμιας αγοράς. Ωστόσο, βρίσκεται ακόμη μακριά από την απόλυτη εφαρμογή του.

Η μετάφραση εξακολουθεί να είναι μια υπηρεσία που εξαρτάται από υψηλό βαθμό εμπιστοσύνης μεταξύ του μεταφραστή και του πελάτη. Μικρή, σταθερή και υψηλά αμειβόμενη εργασία θα προέρχεται από άγνωστους πελάτες, ενώ τα ποσά που καταβάλλονται σε διάφορες χώρες θα εξακολουθούν να διαφέρουν σημαντικά. Οι καλύτερες επαφές είναι πιθανώς εκείνες που γίνονται με φυσική παρουσία. Μια δεύτερη συνέπεια των ηλεκτρονικών μέσων επικοινωνίας είναι ο αυξημένος κίνδυνος για την ασφάλεια.

Οι μεταφραστές εργάζονται συχνά με υλικό που δεν είναι στο δημόσιο τομέα, και αυτός είναι πράγματι ένας από τους λόγους για τους οποίους οι σχέσεις εμπιστοσύνης γίνονται τόσο σημαντικές. Όταν στέλνουμε και λαμβάνουμε αρχεία, θα πρέπει να γνωρίζουμε διάφορους τύπους

κωδικών, ασφαλή FTP ή άλλες μορφές κωδικοποίησης για κάθε εταιρεία, με όλους τους αντίστοιχους κωδικούς πρόσβασης. Μια τρίτη συνέπεια είναι ότι οι ηλεκτρονικές επικοινωνίες καθιστούν σχετικά εύκολη τη διανομή πολύ μεγάλων θέσεων απασχόλησης στη μετάφραση μεταξύ διαφόρων διαμεσολαβητών.

Ο πελάτης μπορεί να θέλει να κυκλοφορήσει το προϊόν του σε 15 ευρωπαϊκές γλώσσες. Προσλαμβάνουν μια εταιρεία μάρκετινγκ, η οποία προσλαμβάνει έναν πάροχο υπηρεσιών γλωσσών, ο οποίος προσλαμβάνει μια σειρά από μεσάζοντες για κάθε γλώσσα, οι οποίοι δίνουν το έργο σε μια σειρά εταιρειών μετάφρασης, οι οποίες διαβιβάζουν τα κείμενα στους μεταφραστές, συχνά ελεύθερους επαγγελματίες.

Σε αυτό το είδος συστήματος, ο πελάτης μπορεί να πληρώνει έως και τέσσερις φορές υψηλότερο ποσό από αυτό που λαμβάνουν οι μεταφραστές ανά μεταφρασμένη σελίδα. Αλλά κάθε σύνδεσμος στην αλυσίδα διορθώνει συντονίζει και παράγει τα διάφορα προϊόντα μετάφρασης προσθέτοντας αξία, καθώς η διαδικασία προχωράει. Αυτό σημαίνει ότι το κείμενο που παράγει ο μεταφραστής δεν είναι συνήθως το ίδιο κείμενο με το κείμενο που πραγματικά χρησιμοποιείται και επομένως δεν μπορεί να αμφισβητηθεί το δικαίωμα πνευματικής ιδιοκτησίας πάνω στο έργο του μεταφραστή. Σημαίνει, επίσης, ότι οι μεταφραστές απέχουν μερικές φορές πολύ από τον τελικό πελάτη και το γενικό πλαίσιο των κειμένων πάνω στα οποία εργάζονται.

Οι μεταφραστές σε εργασίες όπως ο εντοπισμός λογισμικού αρκετά συχνά βλέπουν μόνο λίστες με φράσεις, μαζί με γλωσσάρια που πρέπει να διατηρηθούν. Η εργασία που προκύπτει μπορεί να είναι αρκετά ξένη και χωρίς να περιέχει το ανθρώπινο στοιχείο. Οι ηλεκτρονικές επικοινωνίες έχουν επίσης χρησιμοποιηθεί για την ενίσχυση της επικοινωνίας μεταξύ των μεταφραστών, ιδίως μέσω των φόρουμ του Διαδικτύου για επαγγελματίες μεταφραστές. Αυτά συνήθως ταξινομούνται βάσει θεμάτων ή / και ζευγών γλωσσών. Ορισμένα μπορεί να είναι ανοιχτά, σε άλλα η συμμετοχή περιορίζεται στα εγγεγραμμένα μέλη.

Η επισκεψιμότητα (αριθμός επισκεπτών) σε κάθε ομάδα ποικίλλει από μερικά μηνύματα μηνιαίως, έως εκατοντάδες ημερησίως. Σε αυτά τα φόρουμ οι μεταφραστές είναι πολύ πρόθυμοι να ανταλλάξουν συμβουλές, να δώσουν ιδέες και γενικά να συζητήσουν για το έργο τους. Διαβάζοντας απλώς τα μηνύματα, οι σπουδαστές και οι αρχάριοι μεταφραστές μπορούν να μάθουν

για τη μετάφραση και να δουν το είδος υποστήριξης που οι επαγγελματίες δίνουν ο ένας στον άλλο.

Οι λίστες συζήτησης για επαγγελματίες έχουν συνήθως τις δικές τους οδηγίες επικοινωνίας και έτσι οι νέοι συμμετέχοντες βλέπουν έναν συγκεκριμένο τρόπο αλληλεπίδρασης μεταξύ των επαγγελματιών. Για παράδειγμα, όταν ρωτάμε για την ορολογία, οι επαγγελματίες μεταφραστές στέλνουν συνήθως ένα σύντομο μήνυμα στο οποίο δίνουν τον όρο, κάποιο πλαίσιο, προτεινόμενες μεταφράσεις και τις πηγές που συμβουλεύονται.

Αυτό το μοντέλο δίνει πολύτιμες συμβουλές για την ανάκτηση ορολογίας και τις δεξιότητες ομαδικής εργασίας. Επιπλέον, διαβάζοντας μηνύματα σχετικά με ένα συγκεκριμένο εργαλείο ηλεκτρονικού υπολογιστή, οι αρχάριοι μεταφραστές συχνά ανακαλύπτουν ότι το πρόγραμμα βρίσκεται σε συνεχή ανάπτυξη και έχει λειτουργίες που σε διαφορετική περίπτωση θα αγνοούσαν. Με αυτόν τον τρόπο τα φόρουμ δημιουργούν μια πολύτιμη «γέφυρα» μεταξύ των σπουδαστών και του επαγγελματικού κόσμου. Τελικώς, καταρρίπτεται το στερεότυπο του επαγγελματία μεταφραστή, ο οποίος εργάζεται απομονωμένος πίσω από έναν τοίχο από σκονισμένα λεξικά (Gil&Pym, 2008).

### 3.3 Η γλώσσα Turbo Prolog

Η γλώσσα Turbo Prolog είναι μια σύγχρονη γλώσσα προγραμματισμού της τελευταίας δεκαετίας. Πρόκειται να ένα ειδικό σύστημα γλώσσας που ανήκει στην ευρύτερη κατηγορία των συστημάτων Prolog

Η γλώσσα Prolog, μετά την επινόησή της στις αρχές της δεκαετίας του 1970, έγινε γρήγορα η κυριότερη γλωσσά Τεχνητής Νοημοσύνης στα ευρωπαϊκό εργαστήρια και τα πανεπιστήμια. Κατά τα τέλη της δεκαετίας του 1970 άρχισαν να εμφανίζονται οι πρώτες εκδόσεις της Prolog για μικροϋπολογιστές. Η Prolog είναι σχεδιασμένη για να χειρίζεται λογικά προβλήματα, δηλαδή προβλήματα για τα οποία πρέπει να παρθούν αποφάσεις με συστηματικό τρόπο. Με την χρήση της Prolog γίνεται προσπάθεια να αιτιολογεί ο υπολογιστής την πορεία του προς τη λύση. Είναι ιδιαίτερα κατάλληλη για επίλυση πολλών προβλημάτων Τεχνητής Νοημοσύνης.

Ο δύο σημαντικότερες ομάδες είναι τα Έμπειρα Συστήματα και η Επεξεργασία της Φυσικής Γλώσσας. Η επεξεργασία φυσικών γλωσσών είναι η τεχνική με την οποία οι υπολογιστές αντιλαμβάνονται τη φυσική γλώσσα. Η Turbo Prolog μπορεί να διαιρέσει τη φυσική γλώσσα σε τμήματα από τα οποία αποτελείται κα να ανακαλύψει τις συσχετίσεις που τα συνδέουν, με σκοπό να αντιληφθεί τη σημασία του κειμένου ή της ομιλίας. Για να το επιτύχει χρησιμοποιεί βάση δεδομένων και μηχανή συμπερασμάτων.

Η παραδοσιακή Prolog είναι γλώσσα λιτή, προσκολλημένη στις δομές των λογικών κανόνων και των βάσεων δεδομένων. Πολλές εκδόσεις της Prolog δεν έχουν τη δυνατότητα να χειριστούν προβλήματα με πολλή αριθμητική, ούτε προβλήματα επεξεργασίας κειμένου. Η Turbo Prolog όχι μόνο περιέχει τις δυνατότητες αυτές, αλλά εμπεριέχει εντολές εισόδου-εξόδου, γραφικά, ακόμα και δυνατότητες παραγωγής ήχου.

Τα προγράμματα Turbo Prolog είναι οργανωμένα σε τέσσερα κύρια τμήματα:

1) Προτάσεις (Clauses)
2) Κατηγορήματα (Pedcaies)
3) Πεδία ορισμού (Domdns)
4) Στόχοι (God) (Μαλαγάρδη, 1995)

### 3.4 Ανθρώπινη Μετάφραση Υποβοηθούμενη από Υπολογιστή

Τα μηχανικά βοηθήματα είναι ήδη διαθέσιμα στο μεταφραστή. Πιθανόν να υπάρχουν κάποιοι μεταφραστές που χρησιμοποιούν μόνο στυλό και χαρτί. Οι γραφομηχανές και τα μηχανήματα υπαγόρευσης είναι γνωστός εξοπλισμός.

Τα εργαλεία που αλλάζουν τον τρόπο εργασίας των μεταφραστών είναι τα υπολογιστικά βοηθήματα, απειλώντας σύμφωνα με ορισμένους το κύρος της μετάφρασης ως τέχνης ή ικανότητας, σε αντίθεση με το χαρακτηρισμό της ως δουλειά. Δεν υφίσταται, φυσικά, απειλή ότι τα υπολογιστικά βοηθήματα θα κάνουν τη μετάφραση λιγότερο διανοητικά ή δημιουργικά απαιτητική, όπως δεν συμβαίνει κάτι τέτοιο και με τις γραφομηχανές και τα λεξικά. Σε γενικό επίπεδο, ένα πρόγραμμα επεξεργασίας κειμένου μπορεί να θεωρηθεί μηχανικό βοήθημα. Η χρήση, όμως, αυτού του προγράμματος μετά βίας μπορεί να χαρακτηρισθεί ως ΜΑΗΤ. Ως ελάχιστη

προϋπόθεση, η ΜΑΗΤ πρέπει να περιλαμβάνει κάποιο υπολογιστικό γλωσσολογικό βοήθημα, όπως ένα πρόγραμμα ορθογραφικού ελέγχου, γραμματικής ορθότητας ή ύφους της μετάφρασης. Η διαθεσιμότητα των προγραμμάτων αυτών ποικίλει ανάλογα με τη γλώσσα.

Για την αγγλική γλώσσα υπάρχει πληθώρα επιλογών, ενώ η ελληνική υποστηρίζεται από πολύ λιγότερες εφαρμογές. Γενικά, η εικόνα μεταβάλλεται ανάλογα με την εμπορική ελκυστικότητα μιας γλώσσας. Τα περισσότερα προγράμματα ελέγχου ορθογραφίας βασίζονται σε πολύ μεγάλα λεξικά με δυνατότητες ταχύτατης αναζήτησης, αλλά δεν ενσωματώνουν καθόλου γλωσσολογική «ευφυΐα», αν και ορισμένα διαθέτουν αλγόριθμους που επιχειρούν να μαντέψουν τη σωστή μορφή ενός τυπογραφικού ή ορθογραφικού λάθους.

Τα προγράμματα ελέγχου γραμματικής και ύφους του κειμένου είναι ελαφρώς πιο περίπλοκα εργαλεία που τυπικά λειτουργούν συγκρίνοντας προϋπάρχοντα πρότυπα, αν και ορισμένα χρησιμοποιούν μεθόδους ανάλυσης της υπολογιστικής γλωσσολογίας. Τα προγράμματα ελέγχου γραμματικής αναζητούν στο κείμενο λάθη, όπως ασυμφωνία υποκειμένων και ρημάτων, επανάληψη λέξεων (πχ. «το το»), προτάσεις χωρίς κύριο ρήμα κ.λπ. Τα προγράμματα ύφους ψάχνουν για βασικές μορφές που είναι υφολογικά απορριπτέες, όπως τυποποιημένες εκφράσεις, προτάσεις που ξεκινούν με σύνδεσμο ή τελειώνουν με πρόθεση, προτάσεις που είναι πολύ μεγάλες ή πολύ μικρές κ,τ,λ. Φυσικά, αυτά τα βοηθήματα δεν είναι απαραίτητα για όλους τους μεταφραστές.

Μια κατηγορία βοηθημάτων πρακτικής αξίας και αυξανόμενης δημοτικότητας είναι τα on-line εργαλεία, όπως λεξικά, θησαυροί, εγκυκλοπαίδειες και άλλες γενικές πηγές πληροφορίας, στις οποίες μπορεί να ανατρέξει ο μεταφραστής. Οπτικά μέσα αποθήκευσης όπως CD-ROM, που αποθηκεύουν μεγάλες ποσότητες πληροφορίας, μπορούν να ενσωματωθούν σε περιβάλλοντα επεξεργασίας κειμένου και να προσπελαστούν on-line. Ιδιαίτερα ελκυστικά προς χρήση είναι τα δίγλωσσα λεξικά που διατίθενται στο εμπόριο. Η συσσωμάτωση των διάφορων πηγών και εργαλείων έχει οδηγήσει στην ανάπτυξη της έννοιας που είναι γνωστή ως «εργαστήριο του μεταφραστή».

Χαρακτηριστικά, τα συστήματα αυτά είναι βασισμένα σε προσωπικούς υπολογιστές και κάνουν χρήση πολλαπλών παραθυρικών απεικονίσεων ή ακόμα και πολλαπλών οθονών. Ένα

τμήμα της οθόνης είναι ο χώρος όπου δουλεύεται το μετάφρασμα και ο υπόλοιπος χώρος χρησιμοποιείται για on-line λεξικά ή άλλες αναφορικές πηγές, για αναζήτηση σε παλαιότερες αποθηκευμένες μεταφράσεις με κοινό θέμα κ.α. Η πλήρης συσσωμάτωση σημαίνει ότι η πληροφορία μπορεί εύκολα να μεταφερθεί από το ένα παράθυρο στο άλλο.

Τα πηγαία κείμενα μπορούν να εισαχθούν με άμεση πληκτρολόγηση, να εισαχθούν με τη μορφή αρχείου από κάποιο φορητό μέσο αποθήκευσης ή μέσω δικτύου, αλλά και να μετατραπούν σε ηλεκτρονική μορφή από τυπωμένο χαρτί με τη χρήση ενός συστήματος οπτικής αναγνώρισης χαρακτήρων (optical character recognition - OCR). Συχνά υπάρχουν εργαλεία για τη δημιουργία λεξιλογίων, καταλόγων των λέξεων ορολογίας ενός συγκεκριμένου κειμένου με προτεινόμενα αντίστοιχα σε μια γλώσσα-στόχο.

Τα αντίστοιχα μπορούν να προέρχονται είτε από online πηγές, είτε από άλλα λεξιλόγια του μεταφραστή-χρήστη. Συχνά υπάρχει η εναλλακτική επιλογή της αυτόματης αναζήτησης όρων, όπου το σύστημα ανατρέχει σε τοπικές και απομακρυσμένες ορολογικές βάσεις δεδομένων για τεχνικούς και εξειδικευμένους όρους σε ένα κείμενο. Όλες αυτές οι παροχές του εργαστηρίου του μεταφραστή μπορούν να μειώσουν πολύ τον κόπο του μεταφραστή. Έχει υπολογιστεί ότι οι μεταφραστές τεχνικών κειμένων ξοδεύουν ως και 60% του χρόνου εργασίας στην διερεύνηση σε λεξικά και άλλες πηγές αναφοράς για ορολογική αναζήτηση (Hutchins & Somers, 1992).

### 3.5 Λεξικά προβλήματα της αυτόματης μετάφρασης

Το έργο των Reifler και Oettinger δείχνει σαφώς ότι δεν θα μπορούσε να υπάρξει δικαιολογία για την αναμονή βελτίωσης των αναμνήσεων πριν ξεκινήσει το βασικό γλωσσικό έργο. Το μόνο που χρειάζεται είναι να διασφαλιστεί ότι η λεξική έρευνα θα διεξαχθεί με τη σειρά που θα υπολογίζεται καλύτερα για να αξιοποιήσει τις ιδιότητες των υπαρχόντων μηχανών, χωρίς να παραβλέπει τις μελλοντικές δυνατότητες.

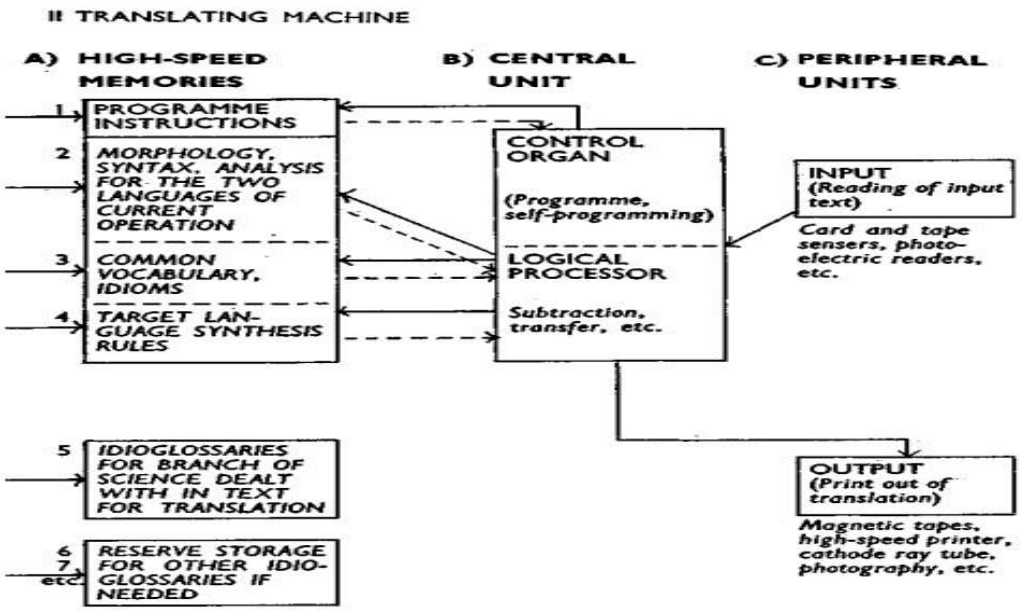
Μπορούμε λοιπόν να ελπίζουμε ότι πολλές από τις καθαρά γλωσσικές πτυχές της οργάνωσης της λεξικής εργασίας θα έχουν αντιμετωπιστεί από τη στιγμή που οι ηλεκτρονικοί τεχνικοί θα παρουσιάσουν τις βέλτιστες λύσεις τους. Εάν το μέσο μήκος μιας λέξης είναι έξι γράμματα (στα οποία πρέπει να προστεθούν γραμματικά σύμβολα και οδηγίες προγράμματος που

αντιστοιχούν σε κάθε λέξη) πρέπει να δεχτούμε έξι bits ανά χαρακτήρα + τα σύμβολα και τις οδηγίες, δηλαδή περίπου 250 bits ανά λήμμα.

Ορισμένα προγράμματα ενδέχεται να απαιτούν έως και 1.000 bits ανά λέξη λεξικού. Τα μαγνητικά τύμπανα του τύπου που χρησιμοποιείται στον υπολογιστή 704 και που λειτουργούν τώρα στο γραφείο IBM Paris αποθηκεύουν μεταξύ τους 294.912 bits - δηλαδή, ανάλογα με τον τύπο του προγράμματος, μεταξύ 300 και 1.200 λέξεων, το τύμπανο του Gamma 60, που περιέχει 786.432 bits, έχει χωρητικότητα από 785 έως 3.200 λέξεις. Συνεπώς για να εκτελεστεί ένα ελάχιστο πρόγραμμα μια μηχανή μετάφρασης πρέπει να έχει χωρητικότητα τουλάχιστον τόση όση αυτή του IBM 704. Να μπορεί να απαιτεί διπλάσια χωρητικότητα, εκτός αν είναι ρυθμισμένη ώστε η μηχανή να έχει γρήγορη πρόσβαση σε άλλους τύπους μνήμης που τίθενται σε λειτουργία μόνο σε ορισμένες περιπτώσεις, για ορισμένα τμήματα του προγράμματος.

Μπορούν να χρησιμοποιηθούν διάφοροι τύποι λεξικής μνήμης. Είναι κατανοητό ότι ένα ή περισσότερα λεξιλόγια που περιλαμβάνουν έναν πολύ μεγάλο αριθμό λέξεων θα μπορούσαν να καταγραφούν σε μαγνητικές ταινίες, απεριόριστης χωρητικότητας, αλλά διαδοχικά, και επομένως αργά, στην πρόσβαση. Είναι κατανοητό ότι, για τις ειδικές απαιτήσεις οποιασδήποτε μετάφρασης, ένα τέτοιο λεξιλόγιο (ή ένα τμήμα ενός πολύ μεγάλου λεξιλογίου) θα μπορούσε να μεταφερθεί, μόνο κατά τη διάρκεια της λειτουργίας.

Κατά τη διάρκεια μίας μετάφρασης, θα είναι δυνατόν με την παραλαβή ενός δεδομένου σήματος, να τεθεί σε λειτουργία ένα τέτοιο εξειδικευμένο λεξιλόγιο εγγεγραμμένο σε μαγνητική ταινία, για να μεταφερθεί για λίγα λεπτά μόνο σε τύμπανο ή φερρίτες και να αντικατασταθεί αργότερα με άλλο παρόμοιο λεξιλόγιο.




*Εικόνα 1: Μηχανή μετάφρασης (Πηγή: Thames and Hudson, 1960)*

### 3.6 Αυτόματη αξιολόγηση

Με την αύξηση του αριθμού των διαφορετικών τεχνικών αυτόματης μετάφρασης και των εμπορικών/ερευνητικών συστημάτων, γεννήθηκε η ανάγκη για γρήγορες, αντικειμενικές και οικονομικές σε υλοποίηση τεχνικές αξιολόγησης, ώστε να μπορεί ο καθένας να αξιολογήσει ένα μεταφραστικό σύστημα και να το συγκρίνει με άλλα. Αυτή η ανάγκη οδήγησε στη δημιουργία των αυτόματων μετρικών αξιολόγησης των συστημάτων μηχανικής μετάφρασης.

Οι μετρικές αυτές δίνουν στην παραγόμενη μετάφραση ένα βαθμό, ο οποίος συσχετίζεται με την ανθρώπινη κρίση δίνοντας υψηλότερο βαθμό στις μεταφράσεις, τις οποίες ένας ειδικός θα θεωρούσε καλύτερες. Όσο μεγαλύτερος είναι ο βαθμός της συσχέτισης με την ανθρώπινη κρίση, τόσο καλύτερη και αντικειμενικότερη θεωρείται η αυτόματη μετρική. Η απόδοση μιας καλής μετρικής πρέπει να συμβαδίζει με την ανθρώπινη κρίση σε πολλά είδη κειμένων και όχι μόνο σε συγκεκριμένους τομείς και υπογλώσσες. Επίσης, μια καλή μετρική θα πρέπει να είναι συνεπής και να δίνει παρόμοια αποτελέσματα για παρόμοια κείμενα που έχουν μεταφραστεί από το ίδιο μεταφραστικό σύστημα. Τα πλεονεκτήματα των μετρικών αυτόματης αξιολόγησης σε σχέση με την ανθρώπινη αξιολόγηση, είναι η ευκολία στην εφαρμογή και η αντικειμενική βαθμολόγηση, αφού η ο αλγόριθμος λειτουργεί πάντα με τον ίδιο τρόπο. Από την άλλη, μια αυτόματη μμετρική

δεν μπορεί να είναι πλήρως αντικειμενική, αφού συγκεκριμένες επιλογές στις παραμέτρους της, όπως για παράδειγμα στις μμεταφράσεις αναφοράς, μπορεί να επηρεάσουν το τελικό αποτέλεσμα και να ευνοήσουν συγκεκριμένα είδη κειμένων.

Για την αξιολόγηση ενός συστήματος μηχανικής μετάφρασης μπορεί κάποιος να επιλέξει ανάμεσα από ένα πλήθος αυτόματων μετρικών. Καθεμία από τις μετρικές αυτές προσφέρει διάφορες λειτουργίες, αλλά επειδή καμία μετρική δεν είναι απόλυτα αντικειμενική και δεν καλύπτει όλο το φάσμα της διαδικασίας μετάφραση, δεν γίνεται να συγκριθούν μεταξύ τους. Οι ερευνητικές προσπάθειες για την αξιολόγηση των μετρικών αξιολόγησης εστιάζονται στην ποσοτικοποίηση του βαθμού συσχέτισης με την ανθρώπινη κρίση, χωρίς όμως κάποια σημαντικά αποτελέσματα. Έτσι για την αξιολόγηση των συστημάτων μηχανικής μετάφρασης συνήθως χρησιμοποιούνται περισσότερες της μιας μετρικές αξιολόγησης (Banerjee&Lavie, 2005).

## 3.7 Μεταφράσεις στη μηχανική μετάφραση

Το μεταφρασμένο από άνθρωπο κείμενο θεωρείται ότι εμφανίζει χαρακτηριστικά που αποκλίνουν σε κάποιο βαθμό από εκείνα του κειμένου που συντάχθηκε αρχικά στη γλώσσα αυτή. Ο Baker και άλλοι (1993) αναφέρουν ότι το μεταφρασμένο κείμενο μπορεί: «*να είναι πιο σαφές από την αρχική πηγή, λιγότερο διφορούμενο, απλοποιημένο (λεξιλογικά, συντακτικά και στιλιστικά), να δείχνει προτίμηση σε συμβατική γραμματικότητα, να αποφεύγει την επανάληψη, να μεγαλοποιεί τα χαρακτηριστικά της γλώσσας στόχου, καθώς και να παρουσιάζει χαρακτηριστικά της γλώσσας πηγής*».

Ο όρος «translationese» έχει χρησιμοποιηθεί συχνά για να περιγράψει την παρουσία τέτοιου φαινομένου στο μεταφρασμένο κείμενο. Το πρότυπο πρωτόκολλο αξιολόγησης στη Μηχανική Μετάφραση (ΜΜ) περιλαμβάνει δοκιμές συστήματος σε δείγμα μεταφρασμένου από άνθρωπο κειμένου. Δεδομένου ότι η δημιουργία αυτού του μεταφρασμένου κειμένου είναι δαπανηρή, η επαναχρησιμοποίηση των συνόλων δοκιμών και για τις δύο κατευθύνσεις μετάφρασης είναι συνηθισμένη, ανεξάρτητα από το αν η πηγή ή ο στόχος περιέχουν χαρακτηριστικά translationese. Για παράδειγμα, οι κοινές εργασίες μετάφρασης στη Συνδιάσκεψη για την Μηχανική Μετάφραση (WMT) εξετάζουν γενικά τη μετάφραση μεταξύ ενός

συγκεκριμένου ζεύγους γλωσσών για τον έλεγχο της μετάφρασης από την Κινεζική στην Αγγλική, όπως απεικονίζεται στο Σχήμα 1

**α)** Τμήμα των δοκιμών δεδομένων (που αντιπροσωπεύουν περίπου το 50% των προτάσεων) αποτελείται από κείμενο που προέρχεται από Κινέζικα και μεταφράστηκε από άνθρωπο στα Αγγλικά, ενώ

**β)** Το υπόλοιπο τμήμα (δηλ. το 50%), μεταφράστηκε προς την αντίθετη κατεύθυνση, από τα Αγγλικά στα Κινέζικα με μετάφραση με συμβατικά μέσα. Το κίνητρο για τη δημιουργία των δοκιμών δεδομένων με αυτόν τον τρόπο είναι να δημιουργηθούν σύνολα δοκιμών και για τις δύο κατευθύνσεις ταυτόχρονα (έτσι χωρίς επιπλέον κόστος).

Παρόλο που ο όρος «translationese» έχει αναφερθεί ότι προκάλεσε πιθανώς σύγχυση στα αποτελέσματα της αξιολόγησης της ΜΤ κατά το παρελθόν, εξ όσων γνωρίζουμε, δεν έχει αναφερθεί μέχρι σήμερα καμία λεπτομερής έρευνα σχετικά με την επίδραση της translationese στην ακρίβεια της αξιολόγησης της ΜΤ. Με αυτό το σκοπό, εξετάζουμε το βαθμό στον οποίο τα φαινόμενα translationese μπορούν να επηρεάσουν τα ανθρώπινα και τα αυτόματα αποτελέσματα αξιολόγησης στην ΜΤ. Αρχικά εξετάζουμε τα προηγούμενα αποτελέσματα των κοινών εργασιών WMT, ένας κύριος χώρος για την αξιολόγηση της ΜΤ και αποκαλύπτουμε ότι παρόλο που το σύστημα κατάταξης είναι γενικά πολύ παρόμοιο για την ανθρώπινη αξιολόγηση των κανονικών δοκιμών δεδομένων και των αντίστροφων δοκιμών, σε λίγες περιπτώσεις το σύστημα κατάταξης αποκλίνει σε έναν πιο σοβαρό βαθμό. Για παράδειγμα, για την τουρκική-αγγλική μετάφραση στο WMT-18, το κανονικό και το αντίστροφο σύστημα κατάταξης συσχετίζονται μόνο σε $r = 0.703$ σε μία περίπτωση.

Εκτός από την ανθρώπινη αξιολόγηση, πολύ περισσότερο ανησυχητική είναι η απόκλιση στην κανονική και αντίστροφη κατάταξη όταν χρησιμοποιείται η BLEU για την αξιολόγηση των συστημάτων, όπου ο συσχετισμός μπορεί να είναι τόσο χαμηλός όσο το 0,106 στη χειρότερη περίπτωση. Στη συνέχεια, προσφέρουμε μια επανεκτίμηση μιας ανθρώπινης αξιολόγησης που είχε προηγουμένως επικριθεί για τη συμπερίληψη δοκιμών δεδομένων που δημιουργήθηκαν αντίστροφα, τα οποία απαιτούσαν την ανθρώπινη ισοτιμία από την κινεζική προς την αγγλική ΜΤ. Ανακαλύπτουμε πληροφορίες σχετικά με πρόσθετες πιθανές πηγές ανακρίβειας των

συμπερασμάτων πέρα από την παρουσία του translationese με σκοπό την πρόληψη μελλοντικών ανακριβειών. Προς το σκοπό αυτό, παρέχουμε έναν συνοπτικό και σαφή κατάλογο μελετών που πρέπει να ληφθούν υπόψη κατά τον προγραμματισμό ή την αναθεώρηση των αξιολογήσεων MT (Grahametal., 2019).

### 3.8 Επίπεδο ανάγνωσης της μηχανικής μετάφρασης

Αν και ο στόχος της μηχανικής μετάφρασης είναι να παράγει σημασιολογικά ακριβείς μεταφράσεις από τη μία γλώσσα στην άλλη, υπάρχουν και άλλοι παράγοντες που επηρεάζουν το αν μία μετάφραση είναι "καλή". Ένας παράγοντας που συχνά παραλείπεται είναι το επίπεδο ανάγνωσης της μετάφρασης. Διαφορετικά πλαίσια απαιτούν διαφορετικά επίπεδα ανάγνωσης. Κατά τη μετάφραση για λιγότερο ικανούς αναγνώστες, κάποιος μπορεί να επιθυμεί μια μετάφραση με κοινό λεξιλόγιο και απλές δομές προτάσεων. Σε ένα επαγγελματικό περιβάλλον, ωστόσο, απαιτείται συχνά συνοπτική γλώσσα με προηγμένο λεξιλόγιο και συντακτική δομή. Για παράδειγμα, κατά τη μετάφραση μιας ισπανικής ιστοσελίδας σχετικά με τη μηχανική μετάφραση προς ένα αγγλόφωνο ηλικίας 7 ετών, μπορεί κανείς να αποδώσει, "η μηχανική μετάφραση είναι ένας τρόπος για να πάρετε μια πρόταση από μια γλώσσα και να την μετατρέψετε σε πρόταση σε άλλη γλώσσα".

Κατά τη διαφήμιση ενός νέου λογισμικού μηχανικής μετάφρασης σε έναν πιθανό επενδυτή, θα μπορούσε κανείς να εξηγήσει ότι "η μηχανική μετάφραση είναι η αυτοματοποιημένη διαδικασία με την οποία μια πρόταση από μια γλώσσα προέλευσης μπορεί να μετατραπεί σε μια πρόταση σε άλλη γλώσσα". Και οι δύο προτάσεις έχουν το ίδιο νόημα και δεν απαιτούν εξειδικευμένες τεχνικές γνώσεις, αλλά η μείωση της πολυπλοκότητας στην πρώτη καθιστά ευκολότερη για το παιδί την κατανόηση και η αύξηση της πολυπλοκότητας στη δεύτερη την κάνει να ακούγεται πιο επαγγελματική και εξελιγμένη.

Επιπλέον, για εκείνους που μιλούν τη μητρική γλώσσα του χαμηλού επιπέδου κειμένου, όπου η ποιότητα της μεταφραστικής μηχανής ενδέχεται να είναι κακή, αλλά που μπορούν να διαβάσουν τις βασικές φράσεις σε μια δεύτερη γλώσσα όπου η ποιότητα της μετάφρασης είναι υψηλή, μπορεί να προτιμούν να διαβάζουν (κείμενο) με μικρότερη πολυπλοκότητα αλλά μια σημασιολογικά ακριβή μετάφραση στη δεύτερη γλώσσα τους αντί για ένα ανακριβές, αλλοιωμένο

μήνυμα στην μητρική τους γλώσσα. Σε αυτό το άρθρο, παρουσιάζουμε την εργασία του ελέγχου του επιπέδου ανάγνωσης (έλεγχος αναγνωσιμότητας) στη μηχανική μετάφραση. Αναπτύσσουμε δύο μεθοδολογίες που ελέγχουν το επίπεδο ανάγνωσης μιας μετάφρασης από την ισπανική στην αγγλική γλώσσα, εστιάζοντας στη λεκτική πολυπλοκότητα ως ένα πρώτο βήμα.

Για το επαγγελματικό περιβάλλον, στοχεύουμε στην παραγωγή προηγμένου λεξιλογίου. Για τους λιγότερο ικανούς αναγνώστες, η μετάφραση πρέπει να χρησιμοποιεί απλές λέξεις ενώ διατηρεί την έννοια της πρότασης πηγής. Συνεπώς, χτίζουμε ένα σύστημα όπου ο χρήστης μπορεί να καθορίσει το επίπεδο ανάγνωσης ("απλό" ή "σύνθετο") της μετάφρασης που επιθυμεί να αποδώσει. Μελλοντικές εργασίες θα πρέπει να εξετάσουν τον έλεγχο άλλων παραγόντων που επηρεάζουν την αναγνωσιμότητα μιας πρότασης, όπως η συντακτική δομή (Marchisio et al., 2019).

# Κεφάλαιο 4: Στατιστική μηχανική μετάφραση

## 4.1 Μηχανή στατιστική μετάφρασης

Μια στατιστική μηχανή μετάφρασης είναι μια μηχανή μετάφρασης όπου οι μεταφράσεις δημιουργούνται με βάση τις διάφορες παραμέτρους στατιστικών μοντέλων των οποίων οι παράμετροι καθορίζονται από την ανάλυση δίγλωσσων σωμάτων κειμένων. Η συγκεκριμένη προσέγγιση έρχεται σε αντίθεση με αυτή της μετάφρασης βάσει κανόνων, όπως και αυτής βάσει παραδειγμάτων.

Η στατιστική μηχανή μετάφρασης ξαναήρθε στο προσκήνιο το 1991 από ερευνητές της IBM και συνέβαλε στην αναγέννηση του ενδιαφέροντος για τις μηχανές μετάφρασης τα τελευταία χρόνια. Στις μέρες μας, είναι μακράν η πιο διαδεδομένη και μελετημένη μέθοδος μηχανής μετάφρασης. Η ιδέα στην οποία βασίζεται η στατιστική μετάφραση μηχανών είναι η θεωρία των πληροφοριών.

Ένα έγγραφο μεταφράζεται σύμφωνα με την κατανομή των πιθανοτήτων p (e|f) ότι ένα αντικείμενο e στην γλώσσα στόχο είναι η μετάφραση ενός αντικειμένου f στην πηγαία γλώσσα. Το πρόβλημα της μοντελοποίησης της κατανομής των πιθανοτήτων p (e|f) προσεγγίστηκε με πολλούς τρόπους.

Για μια αποδοτική αναζήτηση πρέπει να χρησιμοποιηθεί ένας αποκωδικοποιητής ο οποίος να χρησιμοποιεί ξένα αντικείμενα και τεχνοτροπίες για να περιορίσουν τον χρόνο αναζήτησης και την ίδια στιγμή να διατηρούν αποδεκτή ποιότητα. Αυτή η «ανταλλαγή» μεταξύ ποιότητας και χρόνου μπορεί να βρεθεί και στην φωνητική αναγνώριση (Geer, 2005).

## 4.2 Google Translate με παράδειγμα μετάφρασης

Το Google translate δεν μπορεί να μεταφράσει όπως ο άνθρωπος. Το ευρύτερο κοινό που χρησιμοποιεί το Διαδίκτυο πιστεύει πως τα συστήματα μηχανικής μετάφρασης μπορούν να μεταφράσουν όποια κείμενα και να εισαγάγουμε. Το σύστημα μεταφράζει κάποια κείμενα με επιτυχία, μέχρι ενός σημείου όμως. Κυρίως μέχρι του σημείου όπου η μετάφραση εξαντλείται από την αντικατάσταση λέξης από λέξη και με τις προϋποθέσεις ότι:

- ✓ οι γλώσσες που έρχονται σε επαφή έχουν στο κείμενο που εισάγεται ίδια (περίπου) δομή και
- ✓ το πρωτότυπο κείμενο και το κείμενο που παρήγαγε το σύστημα χρησιμοποιούνται σε αντίστοιχα επικοινωνιακά πλαίσια.

Αλλά πόσο συχνά συμβαίνει αυτό, να συναντώνται κείμενα που προέρχονται από φυσικές γλώσσες με παρόμοια συντακτική δομή και ίδια επικοινωνιακή αποτελεσματικότητα στα δεδομένα συγκείμενα; Δεν είναι συχνό φαινόμενο. Για παράδειγμα, ο Γερμανός λέει Guten Tag μετά το μεσημέρι. Χρησιμοποιεί, δηλαδή, μια φράση που αποτελείται από τις λέξεις gut (καλός) και der Tag (μέρα), για να χαιρετήσει μετά τις δώδεκα το μεσημέρι. Εμείς χρησιμοποιούμε το καλημέρα μάλλον μέχρι τις δώδεκα.

Όντως, εάν εισάγουμε τον γερμανικό χαιρετισμό στο σύστημα, μας το αποδίδει ως «Καλημέρα». Εάν εισάγουμε το Guten Morgen, τον αντίστοιχο γερμανικό για τον ελληνικό χαιρετισμό, το Google translate πάλι θα παραγάγει το «Καλημέρα». Αυτό, όμως, που δεν γνωρίζει το λογισμικό είναι ποια δομή ταιριάζει σε ποια επικοινωνιακή συγκυρία. Συνεπώς, ακόμη και αν πετύχει η γραμμική απόδοση της πρωτότυπης φράσης, το διαφοροποιημένο επικοινωνιακό πλαίσιο στο οποίο θα ενταχθεί το προϊόν δεν μπορεί να αξιολογηθεί από τη μηχανή. Ο μη ειδικός, που κάνει χρήση οποιουδήποτε συστήματος μηχανικής μετάφρασης, δεν μπορεί να αξιολογήσει την αξιοπιστία του κειμένου που παρέλαβε ως προϊόν, και την αποτελεσματικότητά του στο δεδομένο πλαίσιο επικοινωνίας. Ο αδαής χρήστης του συστήματος το χρησιμοποιεί θεωρώντας, συχνά, πως η μετάφραση είναι συνώνυμο της γλωσσομάθειας. Το ακόλουθο παράδειγμα θα φωτίσει αυτό που εννοούμε, ακόμη πιο έντονα. Η πρόταση που παρατίθεται προέρχεται από γερμανική οικονομική εφημερίδα, και μεταφράστηκε με τη βοήθεια του Google translate:

- Zum Jahresanfangstartete der Index bei 3.576 Punkten, mittlerweilenotierter gut 20 Prozenthöherbei 4300 Punkten und damitnahe am Jahreshoch. (Handelsblatt 9.1.2015)

<u>Το google translate μάς έδωσε το παρακάτω αποτέλεσμα:</u>

- Στις αρχές του δείκτη ξεκίνησε στις 3.576 μονάδες, τώρα αναφέρονται περίπου 20 τοις εκατό υψηλότερο σε 4.300 μονάδες, κοντά στο ετήσιο υψηλό.

Το ελληνικό κείμενο δεν είναι κατανοητό. Τι σημαίνει Στις αρχές του δείκτη ξεκίνησε …; Εμείς θα προτείναμε την ακόλουθη απόδοση της γερμανικής πρότασης:

- Στις αρχές του έτους ο δείκτης (του χρηματιστηρίου) ξεκίνησε στις 3.576 μονάδες, εν τω μεταξύ κυμαίνεται περίπου 20% υψηλότερα στις 4300 μονάδες και έτσι πλησιάζει το υψηλότερο σημείο του έτους.

Το google translate αγνόησε πολλές παραμέτρους. Έκανε μια απλή ανάλυση, αλλά, ως μηχανή, δεν είχε την κρίση του ανθρώπου να αξιολογήσει τα χαρακτηριστικά της γλώσσας, το νόημα και τα συμφραζόμενα. Προφανώς δεν ανέλυσε το σύνθετο γερμανικό ουσιαστικό Jahresanfang στα συστατικά του: das Jahr (έτος) και der Anfang (αρχή). Επίσης, δεν διαθέτει τη δυνατότητα να κρίνει το συγκείμενο. Το συγκεκριμένο απόσπασμα πάρθηκε από τη γερμανική οικονομική εφημερίδα Handelsblatt και αναφερόταν στις μεταβολές του δείκτη του χρηματιστηρίου της Φρανκφούρτης – κάτι που το σύστημα αδυνατεί να αξιολογήσει. Ιδιαίτερο ενδιαφέρον έχει και το παρακάτω μικρό πείραμα που έγινε.

Πέρα από τις αδυναμίες του ως σύστημα, το google translate, όπως και πολλοί μη ειδικοί, συγχέει τη μετάφραση με τη διερμηνεία. Συχνά ακούμε συνανθρώπους μας να εξομοιώνουν τις δύο αυτές δεξιότητες, που, όπως θα φανεί και στη συνέχεια, είναι διαφορετικές μεταξύ τους. Το google translate μεταφράζει τον γερμανικό όρο Dolmetschen (das) ως «ερμηνεύω». Θα περίμενε κανείς από ένα σύστημα μηχανικής μετάφρασης τουλάχιστον σωστή απόδοση των όρων που έχουν σχέση με τη διαγλωσσική επικοινωνία. Αλλά το Google translate δεν αναγνώρισε, πρώτον, πως ο γερμανικός όρος χαρακτηρίζει τη διερμηνεία ως δεξιότητα και όχι την ερμηνεία, όπως μας προτείνει.

Ο Κεντρωτής (2000: 117-118) ορίζει τη μετάφραση περιγραφικά ως εξής:

*Αν ονομάσουμε τη γλώσσα από την οποία μεταφράζουμε ένα κείμενο, γλώσσα αφετηρίας και τη γλώσσα προς την οποία μεταφράζουμε το κείμενο αυτό γλώσσα αφίξεως, μετάφραση είναι γενικώς η μεταφορά ενός μηνύματος από τη γλώσσα αφετηρίας στη γλώσσα αφίξεως ή, πιο αναλυτικά, η μεταφορά ενός σταθερά συγκεκριμένου και, ως εκ τούτου, μονίμως προσφερομένου ή κατ'αρέσκειαν επαναλαμβανόμενου κειμένου, που έχει συνταχθεί σύμφωνα με τους κανόνες της γλώσσας αφετηρίας, από τη γλώσσα αφετηρίας στη γλώσσα αφίξεως, και δη σε ένα κείμενο, που*

*α) συντάσσεται σύμφωνα με τους κανόνες της γλώσσας αφίξεως*

*β) διατηρεί το νόημα του πρωτοτύπου.*(Κεντρώτης, 2000)

### 4.3 Στατιστικά συστήματα

Το 1949, ο Warren Weaver παρουσίασε τη σκέψη της στατιστικής μηχανικής μετάφρασης. Σε αυτή τη μεθοδολογία, στατιστικές μέθοδοι χρησιμοποιούνται για τη δημιουργία μεταφρασμένης πρότασης χρησιμοποιώντας δίγλωσσο σώμα κειμένου. Η στατιστική μηχανική μετάφραση χρησιμοποιεί πραγματικά μεταφραστικά μοντέλα των οποίων οι παράμετροι προέρχονται από την εξέταση των μονογλωσσικών και διγλωσσικών σώμα κειμένων. Το να δημιουργηθούν νέα μοντέλα στατιστικής μηχανικής μετάφρασης είναι μια γρήγορη διαδικασία, ωστόσο η καινοτομία εξαρτάται έντονα από τα υπάρχοντα πολυγλωσσικά σώματα κειμένων.

Χρειάζονται τουλάχιστον 2 εκατομμύρια λέξεις για ένα συγκεκριμένο χώρο και σημαντικά περισσότερα για μία γενική διάλεκτο. Θεωρητικά είναι κατανοητό το πώς μπορεί να επιτύχει καλή ποιότητα, ωστόσο οι περισσότερες οργανώσεις δεν έχουν πολλά πολυγλωσσικά σώματα κειμένων για να κατασκευάσουν σημαντικά μεταφραστικά μοντέλα. Επίσης, η στατιστική μηχανική μετάφραση απαιτεί μεγάλη δύναμη από CPU κάνοντας λίγο πιο δύσκολο να συγκεντρωθεί το hardware.

Επίσης, η στατιστική μηχανική μετάφραση αναλύει της πιθανότητες οι οποίες συνοδεύουν τα δεδομένα του σώματος κειμένων. Για τον υπολογισμό της πιθανότητας πρέπει να υπολογιστεί ένα σύνολο παραμέτρων, μέσα από μια διαδικασία εκπαίδευσης του συστήματος. Τα πρώτα στατιστικά συστήματα εφαρμόζονταν μόνο σε λέξεις, τα σύγχρονα στατιστικά συστήματα όμως χρησιμοποιούν και σε φράσεις, δηλαδή ακολουθίες λέξεων με μεταβλητό μήκος, ως βασικά στοιχεία της μετάφρασης. Οι φράσεις που χρησιμοποιούνται δεν προκύπτουν από κάποια γλωσσολογική επεξεργασία του σώματος κειμένων, αλλά εξάγονται από τα παράλληλα σώματα των κειμένων με στατιστικές μεθόδους. Με την εισαγωγή φράσεων στα στατιστικά μοντέλα, οι ερευνητές κατάφεραν να πετύχουν συστηματική βελτίωση της ποιότητας της μετάφρασης (Koehn, 2005),

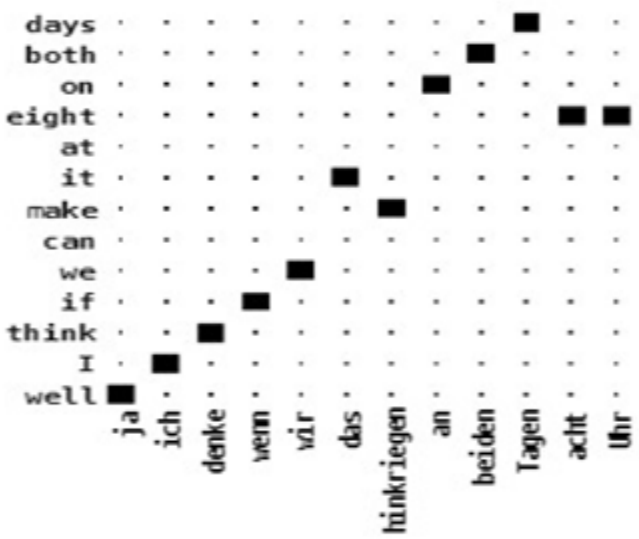

*Εικόνα 2: Μηχανική και στατιστική μετάφραση (Πηγή: Gavrilidouetal., 2012)*

Παρά το γεγονός ότι οι πρώτες ιδέες για τη χρήση στατιστικών μεθόδων εκφράστηκαν στην απαρχή της μηχανικής μμετάφρασης, δεν ήταν παρά μόλις στην αρχή της δεκαετίας του 1990 όταν και ερευνητές της IBM επανάφεραν την ιδέα, προτείνοντας τη χρήση στατιστικών μοντέλων για την μετάφραση φυσικής γλώσσας. Τα συστήματα SMT προσεγγίζουν με διαφορετικό τρόπο το πρόβλημα της μηχανικής μετάφρασης. Αντί να προσπαθούν να απαντήσουν σε ερωτήματα όπως ποια αναπαράσταση θα χρησιμοποιήσουμε και σε πόσα βήματα θα εκτελέσουμε τη μετάφραση, εστίασαν στο αποτέλεσμα της μετάφρασης και προσπάθησαν να καταλάβουν τι ακριβώς σημαίνει να αποτελεί μια πρόταση μετάφραση μιας άλλης πρότασης. Τα στατιστικά συστήματα δεν εξάγουν τη μετάφραση από το σώμα των κειμένων με άμεσο τρόπο, χρησιμοποιώντας την αντιστοιχία μεταξύ της πρότασης στη γλώσσα-πηγή και των παραδειγμάτων, αλλά μέσα από μια διαδικασία ανάλυσης των πιθανοτήτων οι οποίες συνοδεύουν τα δεδομένα του σώματος κειμένων. Ένα σύστημα SMT θεωρεί ότι μια ακολουθία λέξεων e μιας γλώσσας-στόχος είναι μια πιθανή μετάφραση μιας ακολουθίας λέξεων f μιας γλώσσας-πηγή, και αναθέτει σε κάθε τέτοιο ζεύγος (e, f) έναν αριθμό Pr (e|f). Ο αριθμός αυτός αναπαριστά την πιθανότητα ένας μεταφραστής να παράγει την ακολουθία λέξεων e σαν μετάφραση της ακολουθίας f. Η καλύτερη επιλογή ακολουθίας λέξεων στη γλώσσα-πηγή ê είναι αυτή για την οποία η πιθανότητα Pr (e|f) είναι η υψηλότερη (Σοφιανόπουλος, 2009).

## 4.4 Τεχνικές στατιστικής μηχανικής μετάφρασης

### 4.4.1 Basic Alignment Models

Ένα βασικό ζήτημα στη μοντελοποίηση της μετάφρασης της πιθανότητας $P_r\left(f_1^J \middle| e_1^I\right)$ σε στοιχειοσειρά είναι το ερώτημα πώς ορίζουμε την αντιστοιχία μεταξύ των λέξεων της πρότασης-στόχου και των λέξεων της πρότασης-πηγής. Σε τυπικές περιπτώσεις, μπορούμε να υποθέσουμε ένα είδος εξάρτησης ζεύγους εξετάζοντας όλα τα ζεύγη λέξεων $(f_J, e_i)$ για ένα δεδομένο ζεύγος προτάσεων $[f_1^J : e_1^I]$.Εδώ, θα περιορίσουμε περαιτέρω αυτό το μοντέλο, αντιστοιχίζοντας κάθε λέξη πηγής σε μια συγκεκριμένη λέξη-στόχο.




*Εικόνα 3: Μοντέλο αντιστοίχισης (Πηγή: Vogel, 2000)*

Αργότερα, η απαίτηση αυτή θα γίνει πιο χαλαρή. Τα μοντέλα που περιγράφουν αυτούς τους τύπους εξαρτήσεων αναφέρονται ως μοντέλα στοίχισης. Όταν στοιχίζουμε τις λέξεις σε παράλληλα κείμενα (για ζεύγη Ινδο-Ευρωπαϊκών γλωσσών όπως Ισπανικά-Αγγλικά, Γαλλικά-Αγγλικά, Ιταλικά-Γερμανικά, ...), παρατηρούμε συνήθως ένα ισχυρό φαινόμενο τοπικής προσαρμογής. Η εικόνα 3 απεικονίζει αυτό το φαινόμενο για το γερμανικό - αγγλικό γλωσσικό

ζεύγος. Σε πολλές περιπτώσεις, αν και όχι πάντα, υπάρχει ακόμη πιο ισχυρός περιορισμός: σε μεγάλα τμήματα της στοιχειοσειράς πηγής, η στοίχιση είναι μονότονη.

Το άθροισμα της τελευταίας εξίσωσης μπορεί να ερμηνευτεί σαν ένας συνδυασμός τύπων κατανομής με συνδυασμό φορτίων $P(i|j, I, J)$ και με κατανομές στοιχείων που μοντελοποιούν τις ζεύξεις εξαρτήσεων μεταξύ τους. Εκτός από την ”κενή λέξη” που λείπει, αυτό το μοντέλο είναι πανομοιότυπο με το λεγόμενο μοντέλο IBM-2 (Vogel, 2000).

### 4.4.2 Word-Based models

Εάν ανοίξουμε ένα κοινό δίγλωσσο λεξικό, ας πούμε, Γερμανικά-Αγγλικά, μπορεί να βρούμε ένα λήμμα όπως

- ✓ Haus — house, building, home, household, shell.

Οι περισσότερες λέξεις έχουν πολλαπλές μεταφράσεις. Μερικές είναι πιο πιθανές από άλλες. Σε αυτό το παράδειγμα, η μετάφραση σπίτι θα είναι συχνά σωστή κατά τη μετάφραση του Haus στα αγγλικά. Άλλες είναι επίσης συνηθισμένες – κτήριο, οικεία - ενώ κάποιες χρησιμοποιούνται μόνο υπό ορισμένες συνθήκες. Για παράδειγμα, το σπίτι ενός σαλιγκαριού είναι το κέλυφος του. Η έννοια της στατιστικής μηχανικής μετάφρασης συνεπάγεται τη χρήση στατιστικών στοιχείων. Τι είδους στατιστικές θα ήταν χρήσιμες για να αποφασιστεί πώς να μεταφραστεί το Haus;

Εάν είχαμε μια μεγάλη συλλογή γερμανικών κειμένων, σε συνδυασμό με μεταφράσεις στα αγγλικά, θα μπορούσαμε να μετρήσουμε πόσο συχνά μεταφράζεται το Haus σε καθεμία από τις συγκεκριμένες επιλογές. Στον πίνακα 1παρουσιάζεται το πιθανό αποτέλεσμα μιας τέτοιας άσκησης στη συλλογή δεδομένων. Η λέξη Haus εμφανίζεται 10.000 φορές στην υποθετική μας συλλογή κειμένων. Μεταφράζεται 8000 φορές σε σπίτι, 1600 φορές σε κτίριο, και ούτω καθεξής. Σημειώστε ότι αυτή η άσκηση απλοποιεί εντυπωσιακά τα δεδομένα. Αγνοούμε εντελώς το πλαίσιο των περιστατικών αυτών των περιπτώσεων.

Αναγάγουμε το πρόβλημα της μετάφρασης της λέξης Haus σε μια πολύ απλή ερώτηση. Δεδομένου ότι δεν υπάρχουν εξωτερικές γνώσεις, ποιες είναι οι πιθανές μεταφράσεις και πόσο

συχνά συμβαίνουν; Τώρα θέλουμε να υπολογίσουμε μια λεξική μετάφραση κατανομής πιθανοτήτων από αυτές τις μετρήσεις. Αυτή η συνάρτηση θα μας βοηθήσει να απαντήσουμε σε μια ερώτηση που θα προκύψει όταν πρέπει να μεταφράσουμε ένα νέο γερμανικό κείμενο. Ποια είναι η πιο πιθανή αγγλική μετάφραση για μια ξένη λέξη όπως το Haus; Για να το θέσουμε πιο επίσημα, θέλουμε να βρούμε μια συνάρτηση η οποία, δεδομένης μιας ξένης λέξης f (εδώ η Haus), αποδίδει μια πιθανότητα για κάθε επιλογή της αγγλικής μετάφρασης e, που υποδεικνύει πόσο πιθανή είναι η μετάφραση αυτή, Δηλαδή: $pf: e \rightarrow pf(e)$

Η συνάρτηση θα πρέπει να αποδώσει υψηλή τιμή εάν μια αγγλική υποψήφια λέξη e είναι μια συνηθισμένη μετάφραση. Αποδίδει χαμηλή τιμή εάν μια αγγλική υποψήφια λέξη e είναι μια σπάνια μετάφραση. Αποδίδει 0 εάν η αγγλική μετάφραση e είναι αδύνατη. Ο ορισμός της κατανομής πιθανότητας απαιτεί η συνάρτηση pf να έχει δύο ιδιότητες:

$$\sum_{e} p_f(e) = 1$$

$$\forall e: 0 \leq p_f(e) \leq 1$$

Πώς παράγουμε μια κατανομή πιθανοτήτων λαμβάνοντας υπόψη τις μετρήσεις στον πίνακα 1; Ένας απλός τρόπος είναι να χρησιμοποιήσετε την αναλογία των μετρήσεων. Έχουμε 10.000 περιστατικά της λέξης Haus στη συλλογή κειμένων μας. Σε 8000 περιπτώσεις, μεταφράζεται ως σπίτι. Ο διαχωρισμός αυτών των δύο αριθμών δίνει μια αναλογία 0,8, οπότε θέτουμε pHaus (σπίτι) = 0,8.

$$P_f(e) = \begin{cases} 0.8 & \text{if e = house} \\ 0.16 & \text{if e = building} \\ 0.02 & \text{if e = home} \\ 0.015 & \text{if e = household} \\ 0.005 & \text{if e = shell} \end{cases}$$

Αυτή η μέθοδος απόκτησης μιας κατανομής πιθανοτήτων από τα δεδομένα δεν είναι μόνο πολύ ευκολονόητη, αλλά έχει και ισχυρό θεωρητικό κίνητρο. Υπάρχουν πολλοί τρόποι με τους οποίους θα μπορούσαμε να δημιουργήσουμε ένα μοντέλο για τα υπάρχοντα δεδομένα (για παράδειγμα, κρατώντας κάποια μάζα πιθανοτήτων για αόρατα συμβάντα). Αυτός ο τύπος εκτίμησης ονομάζεται *μέγιστη εκτίμηση πιθανότητας*, διότι μεγιστοποιεί την πιθανότητα των δεδομένων (Koehn, 2010).

Έχοντας τις κατανομές πιθανότητας για μία λεξική μετάφραση, μπορούμε να δούμε το πρώτο μοντέλο της στατιστικής μηχανικής μετάφρασης, το οποίο χρησιμοποιεί μόνο πιθανότητες λεξικής μετάφρασης. Για παράδειγμα, ο **πίνακας 1** δείχνει τις κατανομές με βάση πιθανότητας για τη μετάφραση των τεσσάρων γερμανικών λέξεων στα αγγλικά. Δηλώνουμε τώρα την πιθανότητα της μετάφρασης μιας ξένης λέξης «f» σε μια αγγλική λέξη e με τη συνάρτηση πιθανότητας t (e | f), για να κάνουμε πιο σαφή αυτή την κατανομή πιθανότητας ως πιθανότητα μετάφρασης. Οι αντίστοιχοι πίνακες μετάφρασης ονομάζονται συχνά T-πίνακες.

*Πίνακας 1: Πιθανότητα μετάφρασης/λέξη*

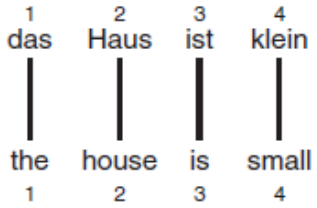
| **das** | | **Haus** | | **ist** | | **klein** | |
|---|---|---|---|---|---|---|---|
| *e* | *t(e\| f)* | *e* | *t(e\| f)* | *e* | *t(e\| f)* | *e* | *t(e\| f)* |
| *the* | 0.7 | *house* | 0.8 | *is* | 0.8 | *small* | 0.4 |
| *that* | 0.15 | *building* | 0.16 | *'s* | 0.16 | *little* | 0.4 |
| *which* | 0.075 | *home* | 0.02 | *exists* | 0.02 | *short* | 0.1 |
| *who* | 0.05 | *household* | 0.015 | *has* | 0.015 | *minor* | 0.06 |
| *this* | 0.025 | *shell* | 0.005 | *are* | 0.005 | *petty* | 0.04 |

Η γερμανική πρόταση: *das Haus ist klein*

Ξεκινά η μετάφραση της πρότασης λέξη προς λέξη στα αγγλικά, Μία πιθανή μετάφραση είναι η εξής: *The house is small.*

Σε αυτή τη μετάφραση υπάρχει μία αντιστοιχία των λέξεων, μια απεικόνιση από Γερμανικές σε Αγγλικές λέξεις. Μεταφράστηκε η Γερμανική λέξη das στην Αγγλική λέξη the.

Αυτή η ευθυγράμμιση μεταξύ των λέξεων εισόδου και εξόδου απεικονίζεται με το παρακάτω διάγραμμα:




*Εικόνα 4: Παράδειγμα 1 αντιστοίχισης λέξεων*

Η ευθυγράμμιση μπορεί να επισημοποιηθεί με μια συνάρτηση ευθυγράμμισης a.

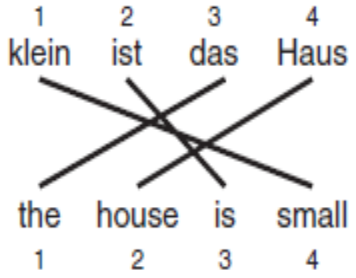
Αυτή η συνάρτηση αντιστοιχίζει, στο παράδειγμά, κάθε αγγλική λέξη εξόδου σε μία θέση i από μια γερμανική λέξη εισόδου στη θέση j: $a : j \rightarrow i$

Παρατηρείτε ότι παρόλο που μεταφράζεται από τα Γερμανικά στα Αγγλικά, η συνάρτηση ευθυγράμμισης αντιστοιχεί τις Αγγλικές θέσεις των λέξεων σε θέσεις Γερμανικών λέξεων.

Στο παράδειγμά, αυτή η συνάρτηση θα παρέχει τις εξής αντιστοιχίσεις:

$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4\}$

Αυτή είναι μια πολύ απλή ευθυγράμμιση, δεδομένου ότι οι Γερμανικές λέξεις και οι αντίστοιχες αγγλικές είναι ακριβώς στην ίδια σειρά. Πολλές γλώσσες έχουν παρόμοια σειρά λέξεων, μια ξένη γλώσσα μπορεί να έχει προτάσεις με διαφορετική σειρά από ότι είναι δυνατή στα Αγγλικά. Αυτό σημαίνει ότι οι λέξεις πρέπει να αναδιαρθρωθούν κατά τη μετάφραση, όπως απεικονίζει το ακόλουθο παράδειγμα:




*Εικόνα 5: Παράδειγμα 2 αντιστοίχισης λέξεων*

Οπότε έχουμε $a : \{1 \rightarrow 3, 2 \rightarrow 4, 3 \rightarrow 2, 4 \rightarrow 1\}$

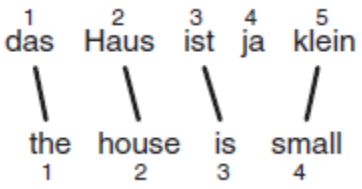
Εκτός από τη διαφορετική σειρά λέξεων, οι γλώσσες μπορεί επίσης να διαφέρουν στο πόσες λέξεις είναι απαραίτητο για να εκφραστεί η ίδια έννοια. Στο Επόμενο παράδειγμα μια Γερμανική λέξη απαιτεί δύο Αγγλικές λέξεις για να έχουν το ίδιο νόημα:


*Εικόνα 6: Παράδειγμα 3 αντιστοίχισης λέξεων*

Οπότε: $a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4, 5 \rightarrow 4\}$

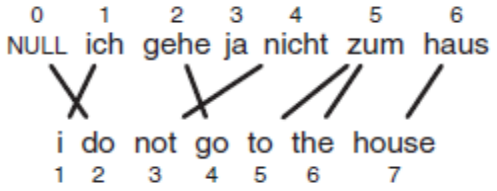
Οι γλώσσες μπορεί να έχουν λέξεις που δεν έχουν ισοδύναμες λέξεις στα Αγγλικά

και για αυτό πρέπει απλά αυτή η λέξη να μην χρησιμοποιηθεί κατά τη μετάφραση. Ένα παράδειγμα αυτό είναι το γερμανικό ja, όπως στο παρακάτω ζεύγος προτάσεων:




*Εικόνα 7: Παράδειγμα 4 αντιστοίχισης λέξεων*

Αντίθετα, μερικές λέξεις στην αγγλική έξοδο της μετάφρασης τους μπορεί να μην έχουν καμία σχέση με οποιαδήποτε από τις γερμανικές λέξεις εισόδου. Για να το μοντελοποιήσουμε, εισάγουμε ένα ειδικό «NULL token» που αντιμετωπίζεται ακριβώς όπως οποιαδήποτε άλλη λέξη εισόδου. Χρειάζεται τε το «NULL token» γιατί πρέπει να ευθυγραμμιστεί κάθε αγγλική λέξη με μία γερμανική λέξη εισόδου, έτσι ώστε η συνάρτηση ευθυγράμμισης να είναι ορισμένη. Στο παρακάτω παράδειγμα, η αγγλική λέξη "do" δεν έχει κάποιο σαφές ισοδύναμο στα γερμανικά:




*Εικόνα 8: Παράδειγμα 5 αντιστοίχισης λέξεων*

Από την εικόνα 8, προκύπτει ότι : $a : \{1 \rightarrow 1, 2 \rightarrow 0, 3 \rightarrow 4, 4 \rightarrow 2, 5 \rightarrow 5, 6 \rightarrow 5, 7 \rightarrow 6\}$

Αυτά είναι τα πρώρα βήματα ή τα θεμέλια για ένα μοντέλο ευθυγράμμισης βασισμένο σε λέξεις που επιτρέπουν την διαγραφή, την προσθήκη ή και την επανάληψη των λέξεων κατά τη μετάφραση. Παρατηρείται ότι στο μοντέλο ευθυγράμμισης, κάθε λέξη στην έξοδο συνδέεται σε μία λέξη εισόδου (συμπεριλαμβανομένου του «token NULL»), αυτά ορίζονται από τη συνάρτηση ευθυγράμμισης. Το ίδιο όμως δεν ισχύει από την άλλη κατεύθυνση: μια λέξη-είσοδος μπορεί να συνδέεται με πολλές λέξεις εξόδου ή με καμία.

#### 4.4.2.1 IBM Model 1

Οι πιθανότητες μετάφρασης και η ιδέα της ευθυγράμμισης μας επιτρέπουν να ορίσουμε ένα μοντέλο το οποίο δημιουργεί έναν αριθμό διαφορετικών μεταφράσεων για μία πρόταση, η κάθε μία με διαφορετική πιθανότητα. Αυτό το μοντέλο ονομάζεται IBM Μοντέλο 1.

Δεν γίνετε να δημιουργηθεί άμεσα ένα μοντέλο για κατανομή πιθανότητας μετάφρασης για πλήρεις προτάσεις, δεδομένου ότι είναι πολύ δύσκολο να εκτιμηθεί μια τέτοια κατανομή. Οι περισσότερες προτάσεις εμφανίζονται μόνο μία φορά, ακόμη και σε τεράστιες συλλογές κειμένων. Οπότε, γίνετε αυτή η διαδικασία σε μικρότερα βήματα, και σε αυτή την περίπτωσή γίνετε σε

μετάφραση κάθε λέξης ξεχωριστά. Είναι πολύ πιο ελπιδοφόρο να συλλέγονται επαρκείς στατιστικές για την εκτίμηση των κατανομών πιθανότητας για τη μετάφραση της κάθε λέξης.

Αυτή η μέθοδος μοντελοποίησης , δηλαδή η διάσπαση της διαδικασίας παραγωγής των δεδομένων σε μικρότερα βήματα, μοντελοποιώντας τα μικρότερα βήματα με μία πιθανότητα κατανομής και συνδυάζοντας τα βήματα σε μια λογική σειρά ονομάζεται γενετική μοντελοποίηση (generative modeling).

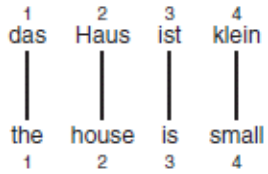
Παρατηρώντας το IBM Μοντέλο 1, είναι ένα γενετικό μοντέλο για μετάφραση φράσεων βασισμένο αποκλειστικά σε κατανομές πιθανοτήτων λεξικών μεταφράσεων. Για κάθε λέξη εξόδου e που παράγεται από το μοντέλο αυτό, από μια λέξη εισόδου f, ο σκοπός είναι να υπολογιστεί η πιθανότητα μετάφρασης P (e|f), και τίποτα άλλο.

Ορίζεται η πιθανότητα μετάφρασης για μια ξένη πρόταση f = (f1, ..., $F_{lf}$) με μήκος $L_f$, σε μια αγγλική πρόταση e = (e1, ..., $e_{le}$) με μήκους $l_e$, με ευθυγράμμιση κάθε αγγλικής λέξης $e_j$ με μια ξένη λέξη $f_i$ σύμφωνα στη λειτουργία ευθυγράμμισης a: j → i ως εξής:

$$P(e,a|f) = \frac{\epsilon}{(l_f+1)^{le}} \prod_{J=1}^{le} t(e_J|f_{a(j)})$$

Ο πυρήνας του είναι ένα προϊόν από τις πιθανότητες της λεξικής μετάφρασης για όλες τις $l_e$ λέξεις σε εξόδου $e_j$ που δημιουργούνται. Το κλάσμα πριν από το προϊόν είναι απαραίτητο για την κανονικοποίηση. Εφόσον υπάρχει το ειδικό σύμβολο NULL, υπάρχουν στην πραγματικότητα $L_f$ + 1 λέξεις εισόδου. Οπότε και υπάρχουν $(lf + 1)^{le}$ διαφορετικές ευθυγραμμίσεις που χαρτογραφούν $L_f$ + 1 λέξεις εισόδου στις λέξεις εξόδου $l_e$. Η παράμετρος ϵ είναι μια σταθερά κανονικοποίησης, έτσι ώστε το p (e, a | f) να είναι μια σωστή κατανομή πιθανότητας, που σημαίνει ότι οι πιθανότητες όλων των πιθανών αγγλικών μεταφράσεων e και ευθυγραμμίσεις a έχουν ένα άθροισμα: $\Sigma_{e,a} p(e,a|f) = 1$

Οπότε εάν αυτό το μοντέλο εφαρμοστεί στο αρχικό μας παράδειγμα:




*Εικόνα 9:Παράδειγμα 6 αντιστοίχισης λέξεων*

Μεταφράζονται τέσσερεις λέξεις και κατά συνέπεια τέσσερεις λεξικές πιθανότητες μετάφρασης πρέπει να λαμβάνεται υπόψη:

$$P(e,a|F) = \frac{\epsilon}{5^4}\ x\ t(the|das)\ x\ t(house|Haus) x\ t(is|ist) x\ t(small|klein)$$
$$= \frac{\epsilon}{5^4} \times 0.7 \times 0.8 \times 0.4\ =\ 0.0029\ \epsilon$$

Οπότε, η πιθανότητα να μεταφραστεί η γερμανική φράση ως "το σπίτι είναι μικρό" είναι 0,0029 ϵ. Αυτό είναι ένα παράδειγμα μετάφρασης χρησιμοποιώντας μοντέλο στατιστικής μηχανικής μετάφρασης.

#### 4.4.2.2 Μαθαίνοντας τα μοντέλα λεξικών μεταφράσεων

Έγινε εισαγωγή ένας μοντέλου για τη μετάφραση προτάσεων με βάση τις διανομής πιθανοτήτων των λεξικών μεταφράσεων. Επίσης υποτίθεται πως ήδη αυτές οι πιθανότητες διανομής μετάφρασης υπάρχουν. Παρακάτω υπάρχει μια μέθοδος που δείχνει πως γίνονται οι κατανομές πιθανοτήτων από παράλληλο κείμενο που έχει ευθυγραμμιστεί κατά πρόταση (δηλαδή ένα κείμενο που συνοδεύεται από μια μετάφραση, πρόταση προς πρόταση). Αυτή η μέθοδος είναι ο αλγόριθμος μεγιστοποίησης προσδοκίας.

#### 4.4.2.3 Το πρόβλημα με την ελλιπή πληροφορία

Η μέθοδος για την εκτίμηση της διανομής των πιθανοτήτων των λεξικών μεταφράσεων σημαίνει ότι, για οποιαδήποτε είσοδο λέξη f, ελέγχετε μία τεράστια συλλογή κειμένων, καταγράφετε πως μεταφράζετε σε κάθε περίπτωση, γίνετε ένα άθροισμα όλων αυτών των

περιπτώσεων και στη συνέχεια υπολογίζετε μια διανομή πιθανοτήτων από αυτό το άθροισμα χρησιμοποιώντας μία εκτίμηση μέγιστης πιθανοφάνειας (maximum likelihood estimation).

Σε αυτή τη μέθοδο όμως υπάρχει ένα μειονέκτημα. Παρόλο που είναι λογικό να συμπεραίνουμε ότι μπορούμε να κάνουμε μεγάλη συλλογή από κείμενα τα οποία είναι ευθυγραμμισμένα κατά πρόταση, δεν σημαίνει ότι είναι ευθυγραμμισμένα και κατά λέξη. Για κάθε είσοδο-λέξη f, δεν ξέρουμε ποια από τις λέξεις της πρότασης είναι η μετάφραση της. Υπάρχει δηλαδή έλλειψη μια συνάρτησης ευθυγράμμισης για αυτό το σκοπό.

Αυτό είναι ένα χαρακτηριστικό πρόβλημα μηχανικής μάθησης. Πρέπει να γίνει εκτίμηση του μοντέλου από ελλιπή δεδομένα. Μια πτυχή του μοντέλου αυτού είναι κρυμμένη, αυτή είναι η ευθυγράμμιση των λέξεων. Για το λόγο αυτό, η ευθυγράμμιση θεωρείται μια κρυφή μεταβλητή του μοντέλου.

Είναι ένα αντιστρέφον πρόβλημα. Αν είχαμε την αντιστοίχιση των λέξεων (word alignment) μαζί με τα δεδομένα μας, θα ήταν ασήμαντο να υπολογίσουμε το μοντέλο λεξικής μετάφρασης. Συλλέγονται τιμές και έπειτα εκτελείτε μία εκτίμηση μέγιστης πιθανοφανείς (maximum likelihood estimation). Όμως αντίστοιχα, αν το μοντέλο ήταν δεδομένο, θα ήταν εφικτό να υπολογιστεί η πιο πιθανή αντιστοίχιση των λέξεων για κάθε ζευγάρι προτάσεων. Με άλλα λόγια: Έχοντας το μοντέλο, θα μπορούσε να καλυφθεί το κενό στα δεδομένα. Λαμβάνοντας υπόψη τα πλήρη δεδομένα, θα μπορούσαμε να εκτιμήσουμε το μοντέλο. Δυστυχώς, το πρόβλημα είναι ότι δεν είναι κανένα δεδομένο.

#### 4.4.2.4 Αλγόριθμος προσδοκίας - μεγιστοποίησης

Ο αλγόριθμος προσδοκίας - μεγιστοποίησης (EM), βοηθάει στο πρόβλημα με τα ελλιπή δεδομένα. Είναι μια επαναληπτική μέθοδος μάθησης η οποία συμπληρώνει τα κενά στα δεδομένα και εκπαιδεύει το μοντέλο σε διαδοχικά βήματα.

Ο αλγόριθμος EM έχει τα εξής βήματα:

1. Αρχικοποίηση του μοντέλου, συνήθως με ομοιόμορφες κατανομές ·

2. Εφαρμογή του μοντέλου στα δεδομένα (βήμα προσδοκίας-expectation) ·

3. Εκμάθηση του μοντέλου από τα δεδομένα (βήμα μεγιστοποίησης-maximazation) ·

4. Επανάληψη των βημάτων 2 και 3 έως τη σύγκλιση.

Αρχικά, γίνετε αρχικοποίηση του μοντέλου. Χωρίς προηγούμενη γνώση, οι ομοιόμορφες κατανομές βάση πιθανότητας είναι ένα καλό σημείο εκκίνησης. Σε αυτήν την περίπτωση εδώ , δηλαδή μίας λεξικής μετάφρασης αυτό σημαίνει ότι για κάθε λέξη εισόδου f , αυτή μπορεί να μεταφραστεί με την ίδια πιθανότητα σε οποιαδήποτε λέξη εξόδου e. Μια άλλη επιλογή είναι να γίνει αρχή με τυχαίες πιθανότητες μετάφρασης.

Στο βήμα προσδοκίας, εφαρμόζουμε το μοντέλο στα δεδομένα μας. Συμπληρώνονται τα κενά που υπάρχουν στα δεδομένα με τις πιθανότερες τιμές. Σε αυτήν την περίπτωση εδώ, αυτό που λείπει είναι η αντιστοίχιση μεταξύ των λέξεων. Επομένως, πρέπει να βρεθούνε οι πιο πιθανές αντιστοιχίσεις. Αρχικά, όλες οι αντιστοιχίσεις είναι εξίσου πιθανές, αλλά επίσης, θα προτιμούμε αντιστοιχίσεις όπου για παράδειγμα η γερμανική λέξη Haus έχει ως αντιστοίχιση πιθανός με την μετάφραση house.

Στο βήμα μεγιστοποίησης, γίνετε κατανοητό το μοντέλο από τα δεδομένα. Στα δεδομένα τώρα έχουν προστεθεί εικασίες για τα κενά. Θα μπορούσε πολύ απλά να ληφθεί υπόψη η καλύτερη εικασία σύμφωνα με το μοντέλο μας, αλλά είναι καλύτερα να εξεταστούν όλες τις πιθανές εικασίες και να τις αριθμήσουμε με τις αντίστοιχες πιθανότητες.

Μερικές φορές είναι αδύνατον να υπολογιστούν αποτελεσματικά όλες οι πιθανές εικασίες, επομένως πρέπει να γίνει κάποια δειγματοληψία. Μαθαίνεται το μοντέλο με εκτίμηση μέγιστης πιθανοφάνειας, χρησιμοποιώντας μερικές από τις τιμές που έχουν συγκεντρωθεί από τα αριθμημένα παραδείγματα.

Επαναλαμβάνονται αυτά τα δύο βήματα μέχρι να γίνει σύγκλιση. Υπάρχουν κάποιες μαθηματικές εγγυήσεις για τον αλγόριθμο ΕΜ. Αρχικά, perplexity (μέτρηση ενός μοντέλου πιθανοτήτων κατά πόσο καλή πρόβλεψη κάνει για ένα δείγμα και επίσης εάν ο αριθμός είναι μικρός δείχνει ότι η κατανομή πιθανότητας είναι καλή στην πρόβλεψη του δείγματος) του μοντέλου είναι εγγυημένη ότι δεν αυξάνεται σε κάθε επανάληψη. Σε ορισμένες περιπτώσεις, όπως

για παράδειγμα στο IBM Model 1, ο αλγόριθμος EM είναι εγγυημένο ότι θα φτάσει στο γενικό ελάχιστο.

#### 4.4.2.5 Εξασφάλιση άπταιστης μετάφρασης

Παραπάνω αναφέρθηκε ένα μοντέλο μετάφρασης βασισμένο σε λέξεις, το μοντέλο IBM 1. Όταν γίνετε η μετάφραση μίας λέξης αυτό το μοντέλο αγνοεί εντελώς της γειτονικές λέξεις. Στην πραγματικότητα, τα περισσότερα μοντέλα που είναι βάση - λέξεων το κάνουν αυτό. Αλλά συχνά το συγκείμενο έχει σημασία. Μπορεί κατά την διάρκεια της αντιστοίχισης να έχουμε ίσες πιθανότητες στην μετάφραση μίας λέξης σε μία άλλη. Οπότε, ανάλογα με τα συμφραζόμενα μία μετάφραση μπορεί να είναι πιο κατάλληλη ή και πιο καλή από την άλλη.

Όσον αφορά το συγκείμενο, χρησιμοποιώντας τη μηχανή αναζήτησης της Google, μπορούμε να πάρουμε αρκετές πληροφορίες για το ποια μετάφραση είναι πιο σωστή για μία λέξη, όπως για παράδειγμα τη συχνότητα που εμφανίζονται 2 φράσεις.

#### 4.4.2.6 Language Model

Αναζητώντας μέσω της μηχανής αναζήτησης της Google είναι μία καλή μέθοδος, αλλά μέχρι και το τεράστιο μέγεθος του παγκόσμιου ιστού έχει τα όρια του. Όσο οι προτάσεις μεγαλώνουν, τόσο πιο δύσκολο θα είναι να βρούμε αντιστοιχίσεις σε σημείο που θα φτάσει στο 0. Έτσι πρέπει να εφαρμοστούν γενετικά μοντέλα, τα οποία θα μειώνουν τους υπολογισμούς μεγάλων προτάσεων σε μικρότερα βήματα και με τη σειρά αυτό θα βοηθήσει να μαζευτούνε στατιστικά για τα μικρότερα κομμάτια. Η πιο συχνή μέθοδος για γλωσσικά μοντέλα είναι η χρήση των n-gram μοντέλων. Παρακάτω φαίνεται ο τύπος του trigram γλωσσικού μοντέλου. Χρησιμοποιείτε n-gram όπου το n = 3.

$$P(e) = P(e_1, e_2, \ldots, e_n)$$

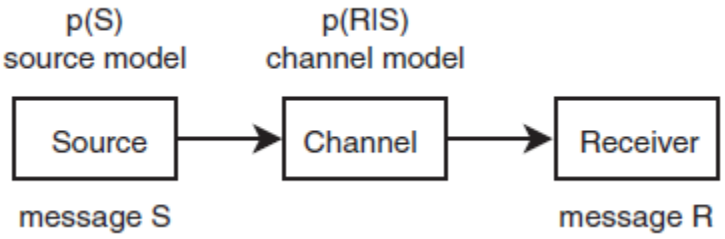
$$= P(e_1)P(e_2|e_1)\ldots P(e_n| e_1, e_2, \ldots, e_{n-1})$$

$$\approx P(e_1)P(e_2|e_1)\ldots P(e_n|e_{n-2}, e_{n-1})$$

Γίνεται διάσπαση της πιθανότητας όλης της πρότασης σε πιθανότητες μίας λέξης, χρησιμοποιώντας των κανόνα αλυσίδας. Μετέπειτα γίνετε μία υπόθεση ότι μόνο οι προηγούμενες δύο λέξεις έχουν σημασία στην πρόβλεψη μίας λέξης. Για να γίνει εκτίμηση πιθανοτήτων του trigram γλωσσικού μοντέλου πρέπει να συγκεντρωθούνε δεδομένα για προτάσεις τριών λέξεων από ένα τεράστιο όγκο δεδομένων. Στην στατιστική μηχανική μετάφραση χρησιμοποιείτε η Αγγλική πλευρά των παράλληλων δεδομένων.

#### 4.4.2.7 Noisy-Channel Model

Το μοντέλο αυτό αρχικά αναπτύχθηκε από τον Claude Shannon το 1948 με σκοπό την διόρθωση τυχόν λαθών λόγω θορύβου.




*Εικόνα 10: Noisy-Channel Model*

Στο μοντέλο θορυβώδους καναλιού ένα μήνυμα μεταφέρεται μέσα από ένα θορυβώδες κανάλι στον παραλήπτη. Το μήνυμα μπορεί να μην είναι ολόκληρο ή να έχει αλλοιωθεί. Το μήνυμα ανακατασκευάζεται με το μοντέλο της πηγής p(S) και το μοντέλο του καναλιού που μεταδόθηκε p(R/S).

Το noisy-channel μοντέλο μπορούμε να το χρησιμοποιήσουμε και στο πρόβλημα της μετάφρασης. Κάνοντας μία υπόθεση ότι ο ομιλητής της ξένης γλώσσας θέλει να εκφέρει μία αγγλική πρόταση αλλά έγινε αλλοίωση σε ένα 'θορυβώδες κανάλι' και το αποτέλεσμα ήταν μία ξένη πρόταση.

Ο δέκτης λαμβάνει μόνο ένα κατεστραμμένο μήνυμα “R”. Η πρόκληση είναι τώρα να ανακατασκευαστεί το αρχικό μήνυμα χρησιμοποιώντας στοιχεία σχετικά με το πιθανό μήνυμα της πηγής και στοιχεία σχετικά με τις παραμορφώσεις που προκαλούνται από το θόρυβο του καναλιού. Επίσης αυτό το μοντέλο μπορεί να χρησιμοποιηθεί για την διόρθωση των λαθών που προκύπτουνε στην οπτική αναγνώριση χαρακτήρων (OCR).

Παραπάνω αναφέρθηκε το IBM Model 1, ένα μοντέλο μηχανικής μετάφρασης. Επειδή όμως έχει αρκετά προβλήματα, δημιουργήθηκαν και άλλα μοντέλα όπως:

IBM Model 1: λεξική μετάφραση

IBM Model 2: προσθήκη απόλυτου μοντέλου αντιστοίχισης

IBM Model 3: προσθήκη μοντέλου παραγωγικότητας

IBM Model 4: προσθήκη σχετικού μοντέλου αντιστοίχισης

IBM Model 5: διορθώσεις σε ατέλειες

### 4.4.3 Phrase-Based models

Η στατιστική μηχανική μετάφραση βασισμένη σε φράσεις έχει αναδειχθεί ως το κυρίαρχο πρότυπο στην έρευνα μηχανικής μετάφρασης. Ωστόσο, μέχρι σήμερα, οι περισσότερες εργασίες στον τομέα αυτό έχουν διεξαχθεί σε ιδιόκτητα και εσωτερικά ερευνητικά συστήματα. Αυτή η έλλειψη ανοίγματος δημιούργησε ένα υψηλό εμπόδιο για την είσοδο των ερευνητών, καθώς πολλά από τα απαιτούμενα στοιχεία χρειάστηκαν να αναπαραχθούν. Αυτό έχει επίσης παρεμποδίσει την αποτελεσματική σύγκριση των διαφόρων στοιχείων των συστημάτων. Παρέχοντας ένα δωρεάν και ολοκληρωμένο εργαλείο, ελπίζουμε ότι αυτό θα τονώσει την ανάπτυξη του τομέα. Για να υιοθετηθεί αυτό το σύστημα από την κοινωνία, πρέπει να καταδείξει απόδοση συγκρίσιμη με τα καλύτερα διαθέσιμα συστήματα.

Ο Moses έχει δείξει ότι αυτό επιτυγχάνει αποτελέσματα συγκρίσιμα με τα πιο ανταγωνιστικά και ευρέως χρησιμοποιούμενα συστήματα στατιστικής μηχανικής μετάφρασης σε

ποιότητα μετάφρασης και χρόνο εκτέλεσης (Shenetal., 2006). Διαθέτει όλες τις δυνατότητες του αποκωδικοποιητή των κλειστών πηγών του Pharaoh (Koehn 2004). Εκτός από την παροχή ενός εργαλείου ανοιχτών πηγών για το SMT, ένα ακόμα κίνητρο για το Moses είναι να επεκτείνει τη μετάφραση βασισμένη σε φράσεις με παράγοντες και με αποκωδικοποίηση δικτύου σύγχυσης.

Η τρέχουσα προσέγγιση που βασίζεται σε φράσεις για τη στατιστική μηχανική μετάφραση περιορίζεται στη χαρτογράφηση μικρών κομματιών κειμένου χωρίς ρητή χρήση γλωσσικών πληροφοριών, είτε μορφολογικών, συντακτικών είτε σημασιολογικών. Αυτές οι πρόσθετες πηγές πληροφοριών έχουν αποδειχθεί πολύτιμες όταν ενσωματώνονται σε στάδια προεπεξεργασίας ή μεταεπεξεργασίας. Ο Moses ενσωματώνει επίσης την αποκωδικοποίηση δικτύου σύγχυσης, η οποία επιτρέπει τη μετάφραση ασαφών καταχωρήσεων. Αυτό επιτρέπει, για παράδειγμα, την πληρέστερη ενσωμάτωση της αναγνώρισης ομιλίας και της μηχανικής μετάφρασης.

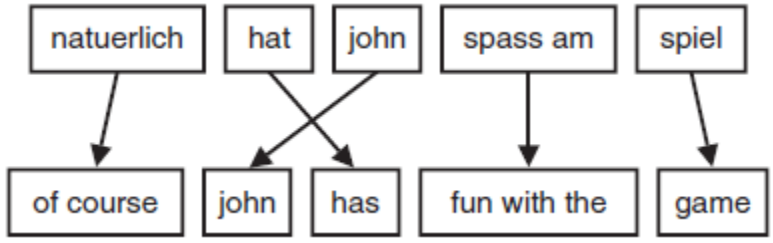
Αντί να μεταφέρει το μοναδικό αποτέλεσμα του αναγνωριστικού, ένα δίκτυο διαφορετικών επιλογών λέξεων μπορεί να εξεταστεί από το σύστημα μηχανικής μετάφρασης. Οι αποδοτικές δομές δεδομένων στο Moses για το εξονυχιστικής μνήμης μοντέλο μετάφρασης και το γλωσσικό μοντέλο επιτρέπουν την αξιοποίηση πολύ μεγαλύτερων πόρων δεδομένων με περιορισμένο υλικό (Koehn et al, 2007).

Τα συστήματα στατιστικής μηχανικής μετάφρασης βασισμένα σε φράσεις είναι: μοντέλα που μεταφράζουν μικρές ακολουθίες λέξεων τη κάθε φορά. Σαν βασικά στοιχεία στη μετάφραση βάση φράσεων είναι: το μεταφραστικό μοντέλο και το μοντέλο αναδιάταξης.

Αρχικά, θα δούμε το βασικό μοντέλο για τη μετάφραση που βασίζεται σε φράσεις. Ενώ υπάρχουν πολλές παραλλαγές, όλα αυτά μπορούν να θεωρηθούν ως επεκτάσεις αυτού του μοντέλου. Είδαμε προηγουμένως μία μέθοδο μηχανικής μετάφρασης η οποία ήταν βασισμένη πάνω σε λέξεις, όμως μπορεί να μην είναι η καλύτερη λύση για μικρές προτάσεις ή μονάδες μετάφρασης.

Μερικές φορές μια λέξη από μια γλώσσα μεταφράζεται σε δύο αγγλικές λέξεις ή αντίστροφα. Τα μοντέλα που βασίζονται σε λέξεις συχνά καταρρέουν σε αυτές τις περιπτώσεις. Στην εικόνα 15, φαίνονται τα μοντέλα που βασίζονται σε φράσεις. Η γερμανική πρόταση εισόδου χωρίζετε αρχικά σε μικρότερες φράσεις. Στη συνέχεια, κάθε φράση μεταφράζεται σε μια Αγγλική

φράση. Τέλος, οι φράσεις μπορεί να αναδιαταχθούν. Στην εικόνα 15, οι έξι γερμανικές λέξεις και οι οκτώ αγγλικές λέξεις αντιστοιχίζονται ως πέντε ζεύγη φράσεων.




*Εικόνα 11: Παράδειγμα 7 αντιστοίχισης λέξεων*

Οι αγγλικές φράσεις πρέπει να αναδιαταχθούν, έτσι ώστε να ακολουθεί το ρήμα. Η γερμανική λέξη natuerlich μεταφράζεται πιο σωστά σε «φυσικά. Ένας μεταφραστικός πίνακας προτάσεων από αγγλικές μεταφράσεις για το γερμανικό natuerlich είναι ο εξής:

*Πίνακας 2: Μεταφραστικός πίνακας*

| Translation | Probability $p(e\|f)$ |
|---|---|
| of course | 0.5 |
| naturally | 0.3 |
| of course , | 0.15 |
| , of course , | 0.05 |

Τα τρέχοντα μοντέλα βασισμένα σε φράσεις δεν είναι βασισμένα σε οποιαδήποτε γλωσσολογική ιδέα. Μία από τις φράσεις στην εικόνα 15 είναι «διασκεδάζει με» (fun with the). Αυτή είναι μια ασυνήθιστη ομαδοποίηση. Οι περισσότερες συντακτικές θεωρίες θα χώριζαν την πρόταση στο ονοματικό σύνολο “fun” και στην προθετική φράση «with the game».

Ωστόσο, η εκμάθηση της μετάφρασης του «spass am» σαν «fun with the» είναι πολύ χρήσιμη. Οι γερμανικές και οι αγγλικές προθέσεις δεν ταιριάζουν πολύ καλά. Αλλά τα συμφραζόμενα μας παρέχουν χρήσιμες ενδείξεις για το πώς πρέπει να μεταφραστούν. Το

γερμανικό «am» έχει πολλές πιθανές μεταφράσεις στα Αγγλικά. Μεταφράζοντας το με το “with the” είναι αρκετά σπάνιο ( πιο συχνό είναι να μεταφραστεί ως “on the” ή “at the” ) αλλά με το συγκείμενο και ακολουθώντας η λέξη «spass» είναι μια πολύ καλή μετάφραση.

Έχουμε δει μέχρι τώρα δύο πλεονεκτήματα της μετάφρασης με βάση φράσεων, αντί για λέξεων. Αρχικά, οι λέξεις μπορεί να μην είναι οι καλύτερες μονάδες για μετάφραση, λόγω συχνών πολλαπλών αντιστοιχίσεων (και αντίστροφα). Επίσης, η μετάφραση ομάδων λέξεων αντί για μεμονωμένες λέξεις μας βοηθούν στην επίλυση σε διάφορες μεταφραστικές ασάφειες. Υπάρχει ένα τρίτο όφελος: αν έχουμε μεγάλο σώμα κειμένων για εκπαίδευση, μπορούμε να μάθουμε όλο και μεγαλύτερες ή και χρήσιμες φράσεις, μερικές φορές ακόμη μπορεί και να «απομνημονεύσει» τη μετάφραση ολόκληρων προτάσεων.

Τελικά, εννοιολογικά, το μοντέλο είναι πολύ απλούστερο. Καταργούμε τις πολύπλοκες έννοιες της γονιμότητας, της εισαγωγής και της διαγραφής που είδαμε στην μετάφραση βάση λέξεων. Δηλαδή, ένα μοντέλο που δεν επιτρέπει την αυθαίρετη προσθήκη ή διαγραφή λέξεων έχει περισσότερη λογική.

### 4.4.4 System Integration

Η στατιστική προσέγγιση για τη μηχανική μετάφραση περιέχεται στη μονάδα stat trans (stat trans module) που ενσωματώνεται στο σύστημα Verbmobil. Η εφαρμογή υποστηρίζει τις οδηγίες μετάφρασης από τα γερμανικά στα αγγλικά και από τα αγγλικά στα γερμανικά. Σε κανονική λειτουργία επεξεργασίας, το stat trans module λαμβάνει την καταχώρησή του από τη μονάδα επισκευής (repair module).

Εκείνη τη στιγμή, τα πλέγματα λέξεων και οι καλύτερες υποθέσεις από τα συστήματα αναγνώρισης ομιλίας έχουν ήδη προσωδιακά επισημανθεί, δηλαδή προστίθενται πληροφορίες σε κάθε άκρη του πλέγματος λέξεων σχετικά με τα όρια του προσωδικού τμήματος, τη λειτουργία της πρότασης και τις τονισμένες συλλαβές. Η μετάφραση πραγματοποιείται με βάση την υπόθεση της καλύτερης πρότασης του αναγνωριστικού. Τα προσωδικά όρια και οι πληροφορίες λειτουργίας της πρότασης χρησιμοποιούνται από τη μονάδα stat trans (stat trans module) ως εξής. Εάν υπάρχει ένα σημαντικό όριο φράσης, εισάγεται μια τελεία ή ερωτηματικό στην πρόταση των

λέξεων, ανάλογα με τη λειτουργία της πρότασης όπως υποδεικνύεται από τη μονάδα προσωδίας. Πρόσθετα κόμματα εισάγονται για άλλους τύπους ορίων του τμήματος (της πρότασης).

Η μονάδα προσωδίας υπολογίζει τις πιθανότητες για τα όρια του τμήματος (τη πρότασης) και χρησιμοποιούνται κατώτατα όρια για να αποφασιστεί εάν θα εισαχθούν τα σημάδια προτάσεων. Αυτά τα κατώτατα όρια έχουν επιλεγεί με τέτοιο τρόπο ώστε, κατά μέσο όρο, για κάθε αλλαγή συνομιλητή του διαλόγου, να επιτυγχάνεται μια καλή κατάτμηση. Τα όρια του τμήματος περιορίζουν την πιθανή αναδιάταξη λέξεων μεταξύ της γλώσσας πηγής και γλώσσας στόχου. Αυτό όχι μόνο βελτιώνει την ποιότητα μετάφρασης, αλλά περιορίζει επίσης τον χώρο αναζήτησης και επομένως επιταχύνει τη διαδικασία μετάφρασης.

Το αποτέλεσμα της μονάδας stat trans είναι η ακολουθία των λέξεων στη γλώσσα στόχο μαζί με ένα σύστημα μέτρησης εμπιστοσύνης που απαιτείται από τη μονάδα επιλογής. Το σύστημα μέτρησης εμπιστοσύνης για τη μετάφραση βασίζεται στις λογαριθμικές πιθανότητες που υπολογίζονται στη διαδικασία αναζήτησης, ομαλοποιημένες σε σχέση με το μήκος της πρότασης (Wahlster, 2000).

## 4.5 Νευρωνικά Δίκτυα

Τα βαθιά νευρωνικά δίκτυα έχουν μεγάλη επιτυχία σε διάφορες εφαρμογές, όπως η αναγνώριση προτύπων και η αναγνώριση ομιλίας. Επιπλέον, πολλά πρόσφατα έργα έδειξαν ότι τα νευρωνικά δίκτυα μπορούν να χρησιμοποιηθούν με επιτυχία σε μια σειρά εργασιών στην επεξεργασία φυσικής γλώσσας (NLP). Αυτές περιλαμβάνουν, αλλά δεν περιορίζονται σε, γλωσσική μοντελοποίηση, ανίχνευση παραφράσεων και εξαγωγή λέξεων. Στον τομέα της στατιστικής μηχανικής μετάφρασης (SMT), τα βαθιά νευρωνικά δίκτυα έχουν αρχίσει να παρουσιάζουν πολλά υποσχόμενα αποτελέσματα. Συνοψίζει μια επιτυχή χρήση των τροφοδοτικών νευρωνικών δικτύων στο πλαίσιο του συστήματος SMT που βασίζεται σε φράσεις.

Κατά τη διάρκεια αυτής της γραμμής έρευνας για τη χρήση νευρωνικών δικτύων για το SMT, το παρόν πόμνημα επικεντρώνεται σε μια καινοτόμο αρχιτεκτονική νευρωνικού δικτύου που μπορεί να χρησιμοποιηθεί ως μέρος του συμβατικού συστήματος SMT με βάση τις φράσεις. Η προτεινόμενη αρχιτεκτονική Νευρωνικού Δικτύου, στην οποία θα αναφερθούμε ως

Κωδικοποιητή-Αποκωδικοποιητή RNN, αποτελείται από δύο επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) που λειτουργούν ως ένα ζεύγος κωδικοποιητή και αποκωδικοποιητή.

Ο κωδικοποιητής χαρτογραφεί μια αλληλουχία πηγών μεταβλητού μήκους σε ένα διάνυσμα σταθερού μήκους και ο αποκωδικοποιητής χαρτογραφεί την αναπαραγωγή του διανύσματος πίσω σε μια αλληλουχία στόχου μεταβλητού μήκους. Τα δύο δίκτυα εκπαιδεύονται από κοινού για να μεγιστοποιήσουν την υποθετική πιθανότητα της ακολουθίας στόχων δεδομένης μιας ακολουθίας πηγών. Επιπλέον, προτείνουμε να χρησιμοποιηθεί μια μάλλον εκλεπτυσμένη κρυφή μονάδα για να βελτιωθεί τόσο η ικανότητα μνήμης όσο και η ευκολία της εκπαίδευσης. Ο προτεινόμενος Κωδικοποιητής-Αποκωδικοποιητής RNN με μια καινοτόμο κρυφή μονάδα αξιολογείται εμπειρικά για το έργο της μετάφρασης από τα αγγλικά στα γαλλικά. Εκπαιδεύουμε το μοντέλο για να μάθουμε την πιθανότητα μετάφρασης μιας αγγλικής φράσης σε μια αντίστοιχη γαλλική φράση. Το μοντέλο στη συνέχεια χρησιμοποιείται ως μέρος ενός τυπικού συστήματος SMT που βασίζεται σε φράσεις, βαθμολογώντας κάθε ζεύγος φράσεων στον πίνακα φράσεων.

Η εμπειρική αξιολόγηση αποκαλύπτει ότι αυτή η προσέγγιση βαθμολόγησης ζευγών φράσεων με Κωδικοποιητή-Αποκωδικοποιητή RNN βελτιώνει την απόδοση μετάφρασης. Αναλύουμε ποιοτικά τον εκπαιδευμένο Κωδικοποιητή-Αποκωδικοποιητή RNN συγκρίνοντας τις βαθμολογίες των φράσεων του με αυτές που δίνονται από το υπάρχον μοντέλο μετάφρασης.

Η ποιοτική ανάλυση δείχνει ότι ο Κωδικοποιητής-Αποκωδικοποιητής RNN είναι καλύτερο να καταγράφει τις γλωσσικές τακτικότητες στον πίνακα φράσεων, εξηγώντας έμμεσα τις ποσοτικές βελτιώσεις στη συνολική απόδοση μετάφρασης. Η περαιτέρω ανάλυση του μοντέλου αποκαλύπτει ότι ο Κωδικοποιητής-Αποκωδικοποιητής RNN μαθαίνει μια συνεχή απεικόνιση του διαστήματος μιας φράσης που διατηρεί τόσο τη σημασιολογική όσο και τη συντακτική δομή της φράσης (Cho et al, 2014).

Η Google παρουσίασε το GNMT (Google Neural Machine Translation) σύστημα το Νοέμβριο του 2016, *Wu, Y. et al. (2016)*. Για την υλοποίησή αυτού του συστήματος χρησιμοποιήθηκαν, τα επαναλαμβανόμενα δίκτυα είναι Long Short-Term Memory (LSTM) RNNs, Hochreiter(1997). Τα LSTM RNN έχουν 8 στρώματα, με υπολειπόμενες συνδέσεις μεταξύ των στρωμάτων για να ενθαρρύνουν τη ροή κύλισης.

Για παραλληλισμό, συνδέεται το attention από το τελευταίο στρώμα του δικτύου αποκωδικοποιητή στο πρώτο επίπεδο του δικτύου κωδικοποιητή. Για την βελτίωση του χρόνου συμπερασμάτων, χρησιμοποιείται αριθμητική χαμηλής ακρίβειας για συμπεράσματα, η οποία επιταχύνεται περαιτέρω από ένα ειδικό hardware (Tensor Processing Unit ή TPU) της Google. Για την αποτελεσματική αντιμετώπιση των σπάνιων λέξεων, χρησιμοποιούνται μονάδες sub-word (επίσης γνωστά ως "wordpieces") για την είσοδο και έξοδο του συστήματος. Η χρήση των wordpieces δίνει καλή ισορροπία μεταξύ της ευελιξίας των μεμονωμένων χαρακτήρων και της αποτελεσματικότητας των πλήρων λέξεων για την αποκωδικοποίηση, επίσης παρακάμπτει την ανάγκη για την ειδική μεταχείριση άγνωστων λέξεων. Η τεχνική beam search περιλαμβάνει μια διαδικασία κανονικοποίησης μήκους για την αποτελεσματική αντιμετώπιση του προβλήματος της σύγκρισης υποθέσεων διαφορετικών μηκών κατά την διαδικασία της αποκωδικοποίηση και μια coverage penalty για να ενθαρρύνει το μοντέλο να μεταφράσει όλες τις παρεχόμενες εισόδους.

Με την χρήση νευρωνικών δικτύων πλέον το επίπεδο της μετάφρασης έχει βελτιωθεί σημαντικά, χωρίς να χρησιμοποιούνται σχεδόν καθόλου εξωτερικά βοηθήματα γλωσσολογίας. Οι Sennrich and Haddow(2016) πρότειναν μία λύση για χρήση γλωσσικής μεθόδου μαζί με ένα NMS , γενικεύοντας το πρόσθετο επίπεδο του κωδικοποιητή στο attentional κωδικοποιητή-αποκωδικοποιητή μοντέλο να υποστηρίζει την χρήση αυθαίρετων χαρακτηριστικών, μαζί με το αρχικό σετ λέξεων. Προθέτονται μορφολογικά χαρακτηριστικά, μέρος του λόγου και ετικέτες συντακτικής εξάρτησης ως χαρακτηριστικά εισόδου στα ζεύγη Αγγλικά ↔ Γερμανικά και Αγγλικά → Ρουμανικά. Στα πειράματα που διεξάγονται υπάρχει βελτίωση του μοντέλου στην ποιότητα σύμφωνα με τρείς μετρήσεις: perplexity, BLEU and CHRF.

Τα πειράματα χρησιμοποιώντας γλωσσικά χαρακτηριστικά δείχνουν βελτίωση, στο newstest2016 με 1.5 μονάδα BLEU για Γερμανικά→Αγγλικά ,0.6 μονάδα BLEU για Αγγλικά→Γερμανικά και 1 μονάδα BLEU για Αγγλικά→Ρουμάνικα.

### 4.5.1 Το πολυγλωσσικό NMT σύστημα (Neural Machine Translation system) της GOOGLE

Αυτή η μέθοδος είναι μια λύση που έδωσαν (Johnson, M. et al, (2017) για ένα NMT σύστημα, το οποίο παρουσιάζει μια μέθοδο η οποία δεν παρεμβαίνει στην αρχιτεκτονική ενός κλασσικού NMT συστήματος αλλά προσθέτει ένα διακριτικό στην αρχή της πρότασης για να προσδιορίσει την ζητούμενη γλώσσα-στόχο. Χρησιμοποιώντας ένα κοινό wordpiece λεξιλόγιο, αυτή η μέθοδος επιτρέπει ένα πολυγλωσσικό NMT σύστημα να λειτουργεί χρησιμοποιώντας μόνο ένα μοντέλο. Υπάρχουν πολλές αναφορές για End-to-End NMT μεταφράσεις αλλά όλα αυτά τα συστήματα είναι σχεδιασμένα για να λειτουργούνε μόνο για ένα ζευγάρι γλωσσών. Σε αυτή την απλή μέθοδο εκμεταλλεύονται τα πολύγλωσσα δεδομένα που υπάρχουν για να βελτιώσει το NMT για όλες τις γλώσσες οι οποίες συμπεριλαμβάνονται. Χωρίς καμία αλλαγή σε ένα παραδοσιακό NMT μοντέλο, γίνετε η προθήκη ενός διακριτικού και όλα τα υπόλοιπα κομμάτια του συστήματος παραμένουν ίδια(ο κωδικοποιητής, ο αποκωδικοποιητής , το σύστημα προσοχής και το κοινό wordpiece όπως περιγράφεται στο Wu et al., (2016). Μερικά Πλεονεκτήματα είναι τα εξής:

#### *Απλότητα*

Δεδομένου ότι δεν γίνονται αλλαγές στην αρχιτεκτονική του μοντέλου, η μετάβαση σε περισσότερες γλώσσες είναι ασήμαντη, τυχόν νέα δεδομένα απλά προστίθενται, πιθανώς με υπέρ ή υπό-δειγματοληψία έτσι ώστε όλες οι γλώσσες να αντιπροσωπεύονται σωστά και απλά, οπότε χρησιμοποιείται ένα νέο διακριτικό(token) εάν έχει αλλάξει η γλώσσα-στόχος. Από την στιγμή που δεν γίνονται αλλαγές στη διαδικασία εκπαίδευσης, οι μικρές παρτίδες για την εκπαίδευση είναι απλώς δείγματα από το γενικό σύνολο όλων των γλωσσών που είναι προς εκπαίδευση όπως είναι και για χρήση σε μια απλή μετάφραση ενός ζευγαριού. Δεδομένου ότι δεν λαμβάνονται εκ των προτέρων αποφάσεις σχετικά με τον τρόπο κατανομής των παραμέτρων για διαφορετικά ζευγάρια γλωσσών, το σύστημα προσαρμόζεται αυτόματα για να χρησιμοποιήσει τον συνολικό αριθμό των παραμέτρων αποτελεσματικά για να ελαχιστοποιήσει την παγκόσμια απώλεια.

#### *Βελτίωση για γλώσσες με λίγα δεδομένα:*

Σε ένα πολύγλωσσο μοντέλο NMT, όλοι οι παράμετροι μοιράζονται έμμεσα από όλα τα ζεύγη γλωσσών που μοντελοποιούνται. Αυτό αναγκάζει το μοντέλο να γενικευθεί πέρα από τα

όρια της γλώσσας κατά τη διάρκεια της εκπαίδευσης. Έχει παρατηρηθεί ότι όταν τα ζεύγη γλωσσών με πολύ λίγα διαθέσιμα δεδομένα τα οποία συνδυάζονται με ζεύγη γλωσσών με άφθονα δεδομένα μέσα σε ένα μοντέλο, η ποιότητα τους (των ζευγαριών με λίγους πόρους) βελτιώνεται σημαντικά.

### *Zero-shot translation*

Ένα εκπληκτικό όφελος στην μοντελοποίηση διαφόρων ζευγών σε ένα μόνο μοντέλο είναι ότι το μοντέλο μπορεί να μάθει να μεταφράζει μεταξύ ζευγών γλωσσών που δεν έχει «δει» ποτέ ξανά αυτόν τον συνδυασμό κατά τη διάρκεια της εκπαίδευσης (zero-shot translation).

Για παράδειγμα, ένα πολύγλωσσο μοντέλο NMT το οποίο έχει εκπαιδευτεί με Πορτογαλικά → Αγγλικά και Αγγλικά → Ισπανικά παρόλο που δεν έχει δεχθεί καθόλου δεδομένα από Πορτογαλικά → Ισπανικά μπορεί να δημιουργήσει ικανοποιητικές μεταφράσεις.

Η αρχιτεκτονική του πολύγλωσσου μοντέλου είναι ίδια με το νευρωνικό δίκτυο μετάφρασης (GNMT) της Google (Wu et al., 2016) (με την προαιρετική προσθήκη των άμεσων συνδέσεων μεταξύ των επιπέδων του κωδικοποιητή και του αποκωδικοποιητή που χρησιμοποιείτε για ορισμένα από τα πειράματα).

Για να μπορέσει να γίνει η χρήση των πολύγλωσσων δεδομένων μέσα σε ένα ενιαίο σύστημα, προτείνετε μία απλή τροποποίηση στα δεδομένα εισόδου, που είναι η εισαγωγή ενός διακριτικού στην αρχή της πρότασης ως προς ένδειξη της γλώσσας-στόχου που πρέπει να μεταφράσει το μοντέλο.

Για παράδειγμα, το ακόλουθο ζεύγος ( Αγγλικά → Ισπανικά ) προτάσεων:

Πώς είσαι; → ¿Cómo estás?

Θα τροποποιηθεί σε:

<2es> Πώς είσαι; → ¿Cómo estás?

για να υποδείξει ότι τα ισπανικά είναι η γλώσσα στόχος. Σημειώνετε ότι δεν καθορίζετε η γλώσσα προέλευσης - το μοντέλο θα το μάθει αυτόματα.

### *Πολλές (γλώσσες-πηγής) σε μία (γλώσσα-στόχο)*

Σε αυτό το παράδειγμα διερευνάτε όταν υπάρχουν πολλές γλώσσες πηγής και μια μοναδική γλώσσα-στόχος , δηλαδή, ο πιο απλός τρόπος συνδυασμού ζευγαριών γλωσσών. Δεδομένου ότι υπάρχει μόνο μια μεμονωμένη γλώσσα στόχος δεν είναι απαραίτητο να χρησιμοποιηθεί ένα διακριτικό.

Στο πρώτο σετ πειραμάτων χρησιμοποιείται το σύνολοο δεδομένων WMT, όπου τα Γερμανικά → Αγγλικά και Γαλλικά → Αγγλικά συνδυάζονται για να εκπαιδευτεί ένα πολύγλωσσο μοντέλο. Η βάση είναι δύο ζεύγη γλωσσών: Τα ζευγάρια Γερμανικά → Αγγλικά και Γαλλικά → Αγγλικά εκπαιδεύτηκαν ανεξάρτητα. Αυτά τα πειράματα εκτελέστηκαν μία φορά με υπερ-δειγματοληψία και μια φορά χωρίς.

Στο δεύτερο σετ πειραμάτων το οποίο είναι σε δεδομένα παραγωγής συνδυάζονται τα ζεύγη Ιαπωνικά → Αγγλικά και Κορεάτικα → Αγγλικά χωρίς υπερ-δειγματοληψία. Το σημείο αναφοράς είναι δύο μοντέλα ξεχωριστών ζευγών γλωσσών εκπαιδεύτηκαν ξεχωριστά.

Τέλος, στο τρίτο σετ πειραμάτων το οποίο είναι και αυτό σε δεδομένα παραγωγής και συνδυάζεται τα ζεύγη Ισπανικά → Αγγλικά και Πορτογαλικά → Αγγλικά χωρίς υπερ-δειγματοληψία. Το σημείο αναφοράς είναι δύο μοντέλα ξεχωριστών ζευγών γλωσσών εκπαιδεύτηκαν ξεχωριστά.

Όλα τα πολυγλωσσικά και μονά ζεύγη μοντέλων έχουν τον ίδιο συνολικό αριθμό παραμέτρων με τα βασικά μοντέλα NMT που έχουν εκπαιδευτεί σε ένα μόνο ζευγάρι γλωσσών (χρησιμοποιώντας 1024nodes, 8 layers LSTM και ένα κοινόχρηστο λεξιλόγιο μοντέλου 32k, συνολικά 255M παραμέτρους ανά μοντέλο). Ένα αρνητικό αυτής της ίσης επιλογής των παραμέτρων είναι ότι είναι πιθανώς άδικο για τα πολύγλωσσα μοντέλα καθώς ο διαθέσιμος αριθμός των παραμέτρων για κάθε ζεύγος γλωσσών μειώνεται κατά έναν αριθμό N σε σύγκριση

δηλαδή με τα μοντέλα για μονά ζεύγη, εάν το N είναι ο αριθμός των ζευγών γλωσσών που συνδυάζονται στο πολύγλωσσο μοντέλο.

Το πολύγλωσσο μοντέλο πρέπει επίσης να χειριστεί το συνδυασμό των λεξιλογίων όλων των μονών μοντέλων. Γίνετε επιλογή για να διατηρηθεί ο αριθμός παραμέτρων σταθερός για όλα τα μοντέλα για να απλοποιήσουμε τον πειραματισμό.

*Πίνακας 3: Αποτελέσματα αξιολόγησης BLEU*

| Model | Single | Multi | Diff |
|---|---|---|---|
| WMT De→En | 30.43 | 30.59 | +0.16 |
| WMT Fr→En | 35.50 | 35.73 | +0.23 |
| WMT De→En* | 30.43 | 30.54 | +0.11 |
| WMT Fr→En* | 35.50 | 36.77 | +1.27 |
| Prod Ja→En | 23.41 | 23.87 | +0.46 |
| Prod Ko→En | 25.42 | 25.47 | +0.05 |
| Prod Es→En | 38.00 | 38.73 | +0.73 |
| Prod Pt→En | 44.40 | 45.19 | +0.79 |

Για τα πειράματα WMT, λαμβάνουμε ένα μέγιστο κέρδος +1,27 BLEU για Γαλλικά→Αγγλικά. Σημειώνουμε ότι τα αποτελέσματα και στα δύο σετ δοκιμών WMT είναι καλύτερα από οποιαδήποτε άλλα δημοσιευμένα αποτελέσματα από τα πιο σύγχρονα μοντέλα.

### *Μία (γλώσσα-πηγή) σε πολλές (γλώσσες-στόχος)*

Εδώ διερευνούμε την εφαρμογή της μεθόδου που είδαμε έως τώρα όταν υπάρχει μια γλώσσα πηγής και πολλές γλώσσες-στόχους. Εδώ πρέπει να μπει στην είσοδο ένα πρόσθετο διακριτικό για τον καθορισμό της γλώσσας-στόχου. Εκτελούνται τρία σετ πειραμάτων παρόμοια με τα προηγούμενα.

Φαίνεται ότι τα πολυγλωσσικά μοντέλα είναι συγκρίσιμα και σε ορισμένες περιπτώσεις υπερτερούν, από τις βάσεις που υπάρχουν σαν παραδείγματα, αλλά όχι πάντα. Λαμβάνετε ένα μεγάλο κέρδος +0,9 BLEU για Αγγλικά→ Ισπανικά. Σε αντίθεση με το προηγούμενο σύνολο

αποτελεσμάτων, υπάρχουν μικρότερα σημαντικά κέρδη σε αυτήν τη ρύθμιση. Αυτό οφείλεται ίσως στο γεγονός ότι ο αποκωδικοποιητής έχει δυσκολότερη δουλεία στο να μεταφράσει σε πολλές γλώσσες-στόχους που μπορεί ακόμη και να έχουν διαφορετική δομή, τα οποία προέρχονται από ένα συνδυαστικό λεξιλόγιο.

*Πίνακας 4: Αποτελέσματα 2 αξιολόγησης BLEU*

| Model | Single | Multi | Diff |
|---|---|---|---|
| WMT En→De | 24.67 | 24.97 | +0.30 |
| WMT En→Fr | 38.95 | 36.84 | -2.11 |
| WMT En→De* | 24.67 | 22.61 | -2.06 |
| WMT En→Fr* | 38.95 | 38.16 | -0.79 |
| Prod En→Ja | 23.66 | 23.73 | +0.07 |
| Prod En→Ko | 19.75 | 19.58 | -0.17 |
| Prod En→Es | 34.50 | 35.40 | +0.90 |
| Prod En→Pt | 38.40 | 38.63 | +0.23 |

Παρατηρείται ότι η υπερ-δειγματοληψία βοηθά στο μικρότερο ζεύγος γλωσσών (Αγγλικά→Γερμανικά) με κόστος την χαμηλότερη ποιότητα για το μεγαλύτερο ζεύγος γλωσσών (Αγγλικά→Γαλλικά). Το μοντέλο χωρίς υπερ-δειγματοληψία επιτυγχάνει καλύτερα αποτελέσματα στο μεγαλύτερο ζευγάρι γλωσσών σε σύγκριση με το μικρότερο όπως ήταν αναμενόμενο. Διαπιστώνουμε επίσης ότι αυτό το αποτέλεσμα είναι πιο εμφανές σε μικρότερα σύνολα δεδομένων (WMT) και πολύ λιγότερο στα πολύ μεγαλύτερα σύνολα δεδομένων παραγωγής.

### *Πολλές (γλώσσες-πηγής) σε πολλές (γλώσσες-πηγής)*

Σε αυτήν την ενότητα, αναφέρονται πειράματα όταν υπάρχουν πολλές γλώσσες-πηγή και πολλαπλές γλώσσες-στόχοι σε ένα μόνο μοντέλο, το οποίο είναι το πιο δύσκολο πρόβλημα. Δεδομένου ότι δίνονται πολλές γλώσσες-στόχοι, πρέπει να μπει στην είσοδο ένα διακριτικό με τη γλώσσα-στόχο. Τα αποτελέσματα παρουσιάζονται στον πίνακα.

*Πίνακας 5: Αποτελέσματα 3 αξιολόγησης BLEU*

| Model | Single | Multi | Diff |
|---|---|---|---|
| WMT En→De | 24.67 | 24.49 | -0.18 |
| WMT En→Fr | 38.95 | 36.23 | -2.72 |
| WMT De→En | 30.43 | 29.84 | -0.59 |
| WMT Fr→En | 35.50 | 34.89 | -0.61 |
| WMT En→De* | 24.67 | 21.92 | -2.75 |
| WMT En→Fr* | 38.95 | 37.45 | -1.50 |
| WMT De→En* | 30.43 | 29.22 | -1.21 |
| WMT Fr→En* | 35.50 | 35.93 | +0.43 |
| Prod En→Ja | 23.66 | 23.12 | -0.54 |
| Prod En→Ko | 19.75 | 19.73 | -0.02 |
| Prod Ja→En | 23.41 | 22.86 | -0.55 |
| Prod Ko→En | 25.42 | 24.76 | -0.66 |
| Prod En→Es | 34.50 | 34.69 | +0.19 |
| Prod En→Pt | 38.40 | 37.25 | -1.15 |
| Prod Es→En | 38.00 | 37.65 | -0.35 |
| Prod Pt→En | 44.40 | 44.02 | -0.38 |

Παρατηρείται ότι τα πολυγλωσσικά μοντέλα παραγωγής με τα ίδιο μέγεθος μοντέλου και ίδιο μέγεθος λεξιλογίου με τα μονά μοντέλα είναι αρκετά κοντά στη βάση, δηλαδή ο μέσος όρος της σχετικής απώλειας της βαθμολογίας BLEU σε όλα τα πειράματα είναι μόνο περίπου 2,5%.

Αν και υπάρχουν σημαντικές απώλειες στην ποιότητα από την εκπαίδευση πολλών γλωσσών από κοινού χρησιμοποιώντας ένα μοντέλο με τον ίδιο συνολικό αριθμό παραμέτρων όπως τα μοντέλα ζεύγους μιας γλώσσας, αυτά τα μοντέλα μειώνουν τη συνολική πολυπλοκότητα που υπάρχει κατά την εκπαίδευση και την παραγωγή.

### *Πειράματα μεγάλης κλίμακας*

Υπάρχουν τα αποτελέσματα από τον συνδυασμό από 12 ζεύγη γλωσσών που έχουν συνολικά 3B παραμέτρους (255M ανά μονό μοντέλο) σε ένα μόνο πολύγλωσσο μοντέλο. Εκπαιδεύτηκε μια ποσότητα από πολύγλωσσα μοντέλα, ξεκινώντας από το ίδιο μέγεθος με ένα μονό ζεύγος γλωσσών μοντέλου με 255M παραμέτρους (1024 nodes) έως και 650M παραμέτρους (1792 nodes). Φαίνεται ότι τα πολύγλωσσα μοντέλα είναι κατά μέσο όρο χειρότερα από τα μόνα μοντέλα (περίπου 5,6% έως 2,5% ανάλογα το μέγεθος, ωστόσο, μερικά γίνονται καλύτερα).Θα

πρέπει να σημειωθεί ότι το μεγαλύτερο πολυγλωσσικό μοντέλο που εκπαιδεύτηκε έχει πέντε φορές λιγότερες παραμέτρους από τα συνδυασμένα μονά μοντέλα.

Το πολύγλωσσο μοντέλο απαιτεί επίσης μόνο περίπου 1/12 του χρόνου εκπαίδευσης (ή υπολογιστικούς πόρους) για να συγκλίνουν σε σύγκριση με τα συνδυασμένα μεμονωμένα μοντέλα (ο συνολικός χρόνος εκπαίδευσης για όλα τα μοντέλα παίρνει σε κάποιες εβδομάδες). Ένα άλλο σημαντικό είναι ότι από τι στιγμή που εκπαιδεύεται λίγο περισσότερο από ένα τυπικό μονό μοντέλο, τα μεμονωμένα ζεύγη γλωσσών γίνονται μόνο με ένα πολύ μικρό μέρος δεδομένων π.χ. $1/12^{o}$ των δεδομένων σε σύγκριση με το μονά μοντέλα ζευγών γλωσσών αλλά παρόλα αυτά εξακολουθούν να παράγουν ικανοποιητικά αποτελέσματα. Εν ολίγοις, το πολύγλωσσο NMT μας επιτρέπει να ομαδοποιούμε διάφορες γλώσσες με μία μικρή απώλεια ποιότητας ενώ έχουμε όλα τα οφέλη όπως αποτελεσματικότητας στην εκπαίδευση, μικρότερο αριθμό μοντέλων και ευκολότερη παραγωγή.

*Πίνακας 6: Αποτελέσματα 4 αξιολόγησης BLEU*

| Model | Single | Multi | Diff |
|---|---|---|---|
| WMT En→De | 24.67 | 24.49 | -0.18 |
| WMT En→Fr | 38.95 | 36.23 | -2.72 |
| WMT De→En | 30.43 | 29.84 | -0.59 |
| WMT Fr→En | 35.50 | 34.89 | -0.61 |
| WMT En→De* | 24.67 | 21.92 | -2.75 |
| WMT En→Fr* | 38.95 | 37.45 | -1.50 |
| WMT De→En* | 30.43 | 29.22 | -1.21 |
| WMT Fr→En* | 35.50 | 35.93 | +0.43 |
| Prod En→Ja | 23.66 | 23.12 | -0.54 |
| Prod En→Ko | 19.75 | 19.73 | -0.02 |
| Prod Ja→En | 23.41 | 22.86 | -0.55 |
| Prod Ko→En | 25.42 | 24.76 | -0.66 |
| Prod En→Es | 34.50 | 34.69 | +0.19 |
| Prod En→Pt | 38.40 | 37.25 | -1.15 |
| Prod Es→En | 38.00 | 37.65 | -0.35 |
| Prod Pt→En | 44.40 | 44.02 | -0.38 |

### *Zero-Shot Translation*

Η πιο απλή προσέγγιση μετάφρασης μεταξύ γλωσσών όπου δεν υπάρχουν καθόλου ή μόνο κάποια λίγα παράλληλα δεδομένα είναι η χρήση μίας σαφής γεφύρωσης, δηλαδή να γίνετε μετάφραση σε μια ενδιάμεση γλώσσα αρχικά και μετά να γίνετε μετάφραση στην επιθυμητή

γλώσσα-στόχο. Η ενδιάμεση γλώσσα είναι συνήθως Αγγλικά π.χ. xx → Αγγλικά και Αγγλικά → yy διότι τα δεδομένα είναι πιο εύκολα διαθέσιμα. Τα δύο πιθανά μειονεκτήματα αυτής της προσέγγισης είναι: α) ο συνολικός χρόνος μετάφρασης διπλασιάζεται, β) η πιθανή απώλεια ποιότητας μεταφράζοντας προς / από την ενδιάμεση γλώσσα.

Ένα ενδιαφέρον πλεονέκτημα της προσέγγισής αυτής είναι ότι επιτρέπει την έμμεση γεφύρωση (zero-shot translation) μεταξύ ενός ζευγαριού γλωσσών για το οποίο δεν έχουν δοθεί σαφή η εμφανή παράλληλα δεδομένα εκπαίδευσης χωρίς καμία τροποποίηση του μοντέλου. Προφανώς, το μοντέλο θα είναι σε θέση να κάνει zero-shot translation μεταξύ των γλωσσών που έχει δει ξεχωριστά ως γλώσσες-πηγής και γλώσσες-στόχου κατά τη διάρκεια της εκπαίδευσης κάποια άλλη στιγμή και όχι για εντελώς νέα γλώσσες που δεν έχουν εμφανιστεί ξανά.

Παρακάτω υπάρχουν 2 πολυγλωσσικά μοντέλα, το ένα έχει εκπαιδευτεί με δύο διαφορετικά ζευγάρια γλωσσών, δηλαδή: Πορτογαλικά → Αγγλικά και Αγγλικά → Ισπανικά (Μοντέλο 1), το δεύτερο μοντέλο έχει παραδείγματα από 4 διαφορετικά ζευγάρια γλωσσών, δηλαδή: Αγγλικά ←→ Πορτογαλικά και Αγγλικά ←→ Ισπανικά (Μοντέλο 2). Όπως και στα προηγούμενα πολυγλωσσικά μοντέλα και τα δύο από αυτά τα μοντέλα μπορούν να έχουν αποτελέσματα κοντά στην βάση με τα μόνα μοντέλα. Επίσης φαίνεται ότι και τα 2 μοντέλα μπορούν να έχουν καλά αποτελέσματα για Πορτογαλικά → Ισπανικά παρόλο που δεν υπάρχει αυτό το ζευγάρι στην εκπαίδευση.

Στον παρακάτω πίνακα (7) φαίνονται τα αποτελέσματα από το πρώτο πραγματικό zero-shot translation για την μετάφραση από Πορτογαλικά → Ισπανικά.

*Πίνακας 7: Αποτελέσματα 5 αξιολόγησης BLEU*

| | Model | Zero-shot | BLEU |
|---|---|---|---|
| (a) | PBMT bridged | no | 28.99 |
| (b) | NMT bridged | no | 30.91 |
| (c) | NMT Pt→Es | no | 31.50 |
| (d) | Model 1 (Pt→En, En→Es) | yes | 21.62 |
| (e) | Model 2 (En↔{Es, Pt}) | yes | 24.75 |
| (f) | Model 2 + incremental training | no | 31.77 |

Η σειρά (a) και (b) είναι η απόδοση από μετάφραση με βάση φράσεων και NMT με σαφή γεφύρωση (Πορτογαλικά → Αγγλικά, και Αγγλικά → Ισπανικά). Μπορούμε να δούμε ότι η απόδοση από ένα νευρωνικό δίκτυο είναι καλύτερη από ένα PBMT σύστημα σχεδόν κατά 2 BLEU μονάδες. Στο (c) έχουμε ένα μονό μοντέλο NMT με παράλληλες προτάσεις για Πορτογαλικά → Ισπανικά.

Μία ενδιαφέρουσα παρατήρηση είναι και το Μοντέλο 1 και το Μοντέλο 2 μπορούνε να εκτελέσουνε zero-shot μετάφραση με αξιόλογα αποτελέσματα στην ποιότητα (παραδείγματα (d) και (e)) σε σύγκριση με τις πρώτες προσδοκίες ότι δεν θα δουλέψει καθόλου.

Το Μοντέλο 2 ξεπερνάει το Μοντέλο 1 σε απόδοση κατά 3 BLEU μονάδες παρόλο που το 2$^{ο}$ εκπαιδεύτηκε με 4 ζευγάρια γλωσσών σε αντίθεση με το 1$^{ο}$ που εκπαιδεύτηκε με 2 και επίσης ότι και τα 2 Μοντέλα έχουν το ίδιο αριθμό παραμέτρων. Γίνετε υπόθεση ότι αυτό είναι μια απρόσμενη συνέπεια η οποία είναι εφικτή μόνο για τον λόγο ότι υπάρχει μια κοινή αρχιτεκτονική που δίνει την δυνατότητα στο μοντέλο να σχηματίσει μία ενδιάμεση γλώσσα μεταξύ όλων αυτών των γλωσσών.

Στο τελευταίο παράδειγμα σταδιακά εκπαιδεύεται παραπάνω το Μοντέλο 2 με ένα μικρό ποσοστό από Πορτογαλικά → Ισπανικά παράλληλων δεδομένων (με λιγότερα δεδομένα από το παράδειγμα (c) και επιτυγχάνετε καλύτερη ποιότητα μετάφρασης με τον μισό χρόνο αποκωδικοποίησης σε σύγκριση με γεφύρωση του παραδείγματος (b).

Το αποτέλεσμα χρησιμοποιώντας παράλληλα δεδομένα κατά την εκπαίδευση για την βελτίωση του δεν μπορεί να χαρακτηριστεί σαν ένα zero-shot translation. Γενικά όμως αυτό αποδεικνύει ότι αυτή η τεχνική, μια έμμεση γεφύρωση χρησιμοποιώντας zero-shot translation με πολύγλωσσικά μοντέλα μπορούνε να εξυπηρετήσουνε μια καλή αρχή ώστε μετέπειτα να εκπαιδευτούνε περαιτέρω με παράλληλα δεδομένα για καλά αποτελέσματα.

Αυτό το αποτέλεσμα είναι σημαντικό ειδικά για όχι-Αγγλικές γλώσσες, διότι μπορεί να έχουν πολύ μικρές πηγές, όπως επίσης και τα ζευγαριών γλωσσών αυτών των γλωσσών, όπου στην προκειμένη περίπτωση μπορεί να είναι ευκολότερο να βρεθούνε παράλληλες πηγές με Αγγλικά αλλά πολύ πιο δύσκολο να βρεθούνε πηγές για ζευγάρια γλωσσών όπου ούτε η γλώσσα-στόχος ή η γλώσσα-πηγή είναι Αγγλικά.

Από τη στιγμή που τα Πορτογαλικά και τα Ισπανικά είναι από την ίδια οικογένεια, μια ενδιαφέρουσα ερώτηση είναι τι θα γίνει εάν εφαρμοστεί το zero-shot translation σε δύο γλώσσες που δεν έχουν σχέση μεταξύ τους. Στο παρακάτω πίνακα (8) βλέπουμε τα αποτελέσματα από zero-shot translation με έμμεση και άμεση γεφύρωση μεταξύ Ισπανικών και Ιαπωνικών χρησιμοποιώντας το μεγαλύτερο Μοντέλο που είδαμε προηγούμενος (12 γλώσσες). Όπως και αναμενόμενο η zero-shot translation δεν δίνει καλά αποτελέσματα με μεγάλη πτώση στην ποιότητα (έως και 50% σε BLEU σκορ). Παρόλο την πτώση στην ποιότητα αυτό μας δείχνει ότι η προσέγγιση με zero-shot translation είναι εφικτή ακόμα και σε τελείως άσχετες γλώσσες μεταξύ τους.

*Πίνακας 8: Αποτελέσματα 6 αξιολόγησης BLEU*

| Model | BLEU |
|---|---|
| NMT Es→Ja explicitly bridged | 18.00 |
| NMT Es→Ja implicitly bridged | 9.14 |

### *Η επίδραση των παράλληλων δεδομένων*

Υπάρχουνε δύο τρόποι αξιοποίησης των διαθέσιμων παράλληλων δεδομένων για τη βελτίωση της ποιότητας zero-shot translation, παρόμοια με αυτά των Firat και λοιποί, (2016c). Για την πολύγλωσση αρχιτεκτονική που χρησιμοποιείτε όπως παραπάνω λαμβάνουμε υπόψη τις παρακάτω μεθόδους:

Εκπαιδεύοντας σταδιακά το πολύγλωσσο μοντέλο μαζί με τα πρόσθετα παράλληλα δεδομένα για τις μεθόδους zero-shot.

Εκπαιδεύοντας ένα νέο πολύγλωσσο μοντέλο με όλα τα διαθέσιμα παράλληλα δεδομένα να αναμιγνύονται εξίσου.

Εν ολίγοις, η κοινόχρηστη αρχιτεκτονική διαμορφώνει τα ζεύγη γλωσσών zero-shot αρκετά καλά και ως εκ τούτου μας επιτρέπει να βελτιώσουμε εύκολα την ποιότητά τους με μια μικρή ποσότητα επιπλέον παράλληλων δεδομένων.

## 4.6 Αυτοματοποιημένη αξιολόγηση BLEU και NIST, μη αυτοματοποιημένη ARPA

Η μετρική αξιολόγησης BLEU (Bi Lingual Evaluation Understudy) προτάθηκε το 2002 από την IBM. Από τότε αποτελεί το επίσημο μετρικό των συστημάτων ΜΜ και κυρίως εκείνων που μεταφράζουν χρησιμοποιώντας στατιστικές μεθόδους. Αυτό συμβαίνει διότι από έρευνες προκύπτει πως τα αποτελέσματά του συσχετίζονται κατά πολύ με τα αποτελέσματα της ανθρώπινης αξιολόγησης. Ο αλγόριθμος λειτουργεί ανεξάρτητου γλώσσας, αλλά δε λαμβάνει υπόψη τη γραμματική ορθότητα των προτάσεων.

Για την εκτίμηση της βαθμολογίας, η μεταφρασμένη πρόταση συγκρίνεται με 1-3 διαφορετικές μεταφράσεις από επαγγελματία μεταφραστή. Για τον λόγο αυτόν, είναι πολύ δύσκολο μια πρόταση να βαθμολογηθεί με άριστα, πράγμα που συμβαίνει πολύ συχνά κατά την ανθρώπινη αξιολόγηση. Για την εξαγωγή του τελικού αποτελέσματος ο αλγόριθμος λαμβάνει υπόψη του το μήκος της μεταφρασμένης πρότασης καθώς οι αυτόματα μεταφρασμένες προτάσεις συνήθως έχουν μικρότερο μήκος. Ένας άλλος αλγόριθμος που μοιάζει πολύ με τον BLEU, είναι ο NIST. Η ονομασία του προέρχεται από τα αρχικά του Εθνικού Ιδρύματος Προτύπων και Τεχνολογίας (National Institute of Standards and Technology) των ΗΠΑ.

Ο αλγόριθμος αυτός βασίζεται στην ίδια λογική με τον BLEU και διαφέρει σε τρία σημεία όσον αφορά την εκτίμηση της βαθμολογίας. Η σημαντικότερη διαφορά των δύο αλγορίθμων είναι πως ο NIST χρησιμοποιεί τον αριθμητικό μέσο και όχι τον γεωμετρικό μέσο για τον υπολογισμό της ακρίβειας των ngrams. Επίσης, αυτή η μετρική αξιολόγησης εστιάζει κυρίως στην επάρκεια των μεταφράσεων, αφού ενδιαφέρεται περισσότερο για τα unigrams. Παρόλα ταύτα, οι δύο μετρικές χρησιμοποιούνται συμπληρωματικά καθώς η μία συμπληρώνει την άλλη. Η εκτίμηση του τελικού αποτελέσματος για το σύστημα που αξιολογείται, προκύπτει από τον μέσο όρο της βαθμολογίας όλων των προτάσεων του κειμένου, αναγόμενο σε κλίμακα από 0-1.

Για την αξιολόγηση από ανθρώπους έχουν προταθεί κατά διαστήματα διάφορες μεθοδολογίες. Συνήθως ως αξιολογητές χρησιμοποιούνται ειδικοί γλωσσολόγοι της γλώσσας-στόχου, ενώ η αξιολόγηση γίνεται σε επίπεδο μεμονωμένων προτάσεων και όχι σε επίπεδο κειμένου. Μια από τις πιο γνωστές μεθοδολογίες είναι η ‘Advanced Research Projects Agency’ (ARPA), η οποία ξεκίνησε το 1991 και συνεχίζεται μέχρι και σήμερα.

Σε αυτή την αξιολόγηση οι αξιολογητές βαθμολογούν προτάσεις λαμβάνοντας υπόψη δύο παράγοντες: την ‘επάρκεια’ (adequacy) και την ‘ευφράδεια’ (fluency) της γλώσσας-στόχου. Στόχος της ‘επάρκειας’ είναι να αξιολογηθεί η ποσότητα της πληροφορίας δηλαδή πόσο επαρκής είναι η μετάφραση από τη γλώσσα-πηγή στη γλώσσα-στόχο ανεξάρτητα από τη γραμματική, δηλαδή αν το νόημα της πρότασης μεταφέρεται σωστά. Στόχος της ‘ευφράδειας’ είναι να αξιολογηθεί η ποιότητα της γλώσσας από πλευρά γραμματικής και συντακτικού και όχι νοήματος (Κανελλιάδου & Χατζηθεοδώρου, 2008).

## Συμπεράσματα

Η μετάφραση είναι άρρηκτα συνδεδεμένη με τον πολιτισμό, την επικοινωνία και την τεχνολογία. Προϋπάρχει σαν έννοια και σαν τεχνική πολύ πριν την εμφάνιση των υπολογιστών.

Η εξέλιξη της μετάφρασης είναι συνεχής και αδιάκοπη, η χρησιμότητά της παρατηρείται σε όλους τους τομείς είτε είναι τεχνολογία είτε όχι.

Η σημερινή εποχή κατακλύζεται από δεδομένα τα οποία η τεχνολογία είναι σε θέση να τα επεξεργάζεται και να τα αναλύει προς όφελος του πολίτη. Αυτή η αύξηση τω δεδομένων σε συνδυασμό με την ύπαρξη του διαδικτύου δημιούργησαν πρόσφορο έδαφος για την στατιστική μηχανική μετάφραση. Οι μηχανές εκπαιδεύονται και τροφοδοτούνται συνεχώς με δεδομένα.

Η στατιστική μηχανική μετάφραση είναι το επίκαιρα κύριο πρόβλημα του machine learning.

Υπάρχουν ακόμη άλυτα προβλήματα που προκύπτουν στη αντιστοίχιση των εννοιών ανά τις διάφορες γλώσσες και της αδυναμίας της τεχνολογίας και του υπολογιστικού συστήματος να αντιληφθεί έννοιες που για τον άνθρωπο είναι δεδομένες, όπως ύφος γραφής, ειρωνεία στον λόγο, μεταφορική χρήση της γλώσσας κ.α.

Η εργασία του Toral και Way το 2018 στο πανεπιστήμιου του Cornel υποδεικνύει την βελτίωση του συστήματος μετάφρασης με την πάροδο των ετών και επισημαίνει πως η αυτόματη μετάφραση βρίσκεται ακόμη πίσω σε σχέση με την μετάφραση του ανθρώπου.

## Βιβλιογραφία

### Ελληνική


- Ευθυμίου, Ε. (1992). Ορολογία και αυτόματη μετάφραση, Διήμερο συνέδριο¨Τυποποίηση Ορολογίας¨, Αθήνα
- Κανελλιάδου, Π. &Χατζηθεοδώρου, Κ. (2008). Η Αυτοματοποιημένη και μη-αυτοματοποιημένη αξιολόγηση συστήματος Στατιστικής Μηχανικής Μετάφρασης για το γλωσσικό ζεύγος Ελληνικά – Ιταλικά
- Κεντρώτης, Γ. (2000). Θεωρία και πράξη της μετάφρασης, Εκδόσεις Δίαυλος
- Μαλαγάρδη, Ι.Δ. (1995). Συγκριτική ανάλυση να και για να δομών της νέας Ελληνικής με αντίστοιχες δομές της Γερμανικής και εφαρμογή στη μηχανική μετάφραση (Διδακτορική διατριβή). Πανεπιστήμιο Αθηνών Φιλοσοφική σχολή.
- Σοφιανόπουλος, Σ. (2009). Μοντελοποίηση γλώσσας για συστήματα μηχανικής μετάφρασης με μονόγλωσσο σώμα κειμένων (Διδακτορική διατριβή). Εθνικό Μετσόβιο Πολυτεχνείο Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Αθήνα.


### Ξενόγλωσση


- Arnold, D. Balkan, L. Meijer, S. Humphreys, R. L. Sadler, L. (1994). Machine Translation: An Introductory Guide
- Banerjee, S. &Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Ann Arbor*, Michigan.
- Carl, M., Schmidt, P. and Schütz, J. (2005): Reversible Template-based Shake & Bake Generation. Proceedings of the Example-Based Machine Translation Workshop held in conjunction with the 10th Machine Translation Summit, Phuket, Thailand, pp. 17-26.
- Cho, K. Gulcehre, C. Merrienboer, B. Schwenk, H. Bougare, F. Bahdanau, D. &Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation
- Gavrilidou, M. Koutsombogera, M.Patrikakos, A.&Piperidis, S. (2012). Η ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ ΓΙΑ ΤΑ ΕΛΛΗΝΙΚΑ. *The Greek Language in the Digital Age, 17–36.*
- Geer, D. (2005). Statistical Translation Gains Respect, *IEEE Computer*, pp. 18 – 21
- Gil, J.R.B. & Pym, A. (2008). Technology and translation. University Tarragona, Spain

- Graham, Y. Haddow, B. Koehn, P. (2019). Translationese in Machine Translation Evaluation
- Hatim, B. & Mason, I. (1997).The Translator as Communicator. London: Routledge.
- Hochreiter, S., and Schmidhuber, J. Long short-term memory. Neural computation 9, 8 (1997), 1735–1780.
- Hutchins, J. (1978). Machine translation and Machine-Aided Translation. Journal of Documentation, 34, 2, June 978, 119-159
- Hutchins, W.J. & Somers, H.L. (1992). An introduction to machine translation, *London Academic Press*
- Johnson, M. et al. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation." Transactions of the Association for Computational Linguistics 5 (2017): 339-351.
- Kenny, D & Way, A. (2001). Teaching Machine Translation & Translation Technology: *A Constrastive Study,* University of Dublin
- Koehn, P. (2010). Statistical machine translation, Univerity of Edinburgh
- Koehn, P. Federico, M. Cowan, B. Zens, R. Dyer, C Bojar, O. (2007). Moses: Open source toolkit for statistical machine translation
- Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation
- Labropoulou, P., Mantzari, E., Gavrilidou, M.. 1996. Lexicon-Morphosyntactic Specifications: Language Specific Instantiation (Greek), PP-PAROLE MLAP 63-386 report Lee, H.A., Park. J.C. and Kim, G.C. (1999): Lexical Selection with a Target Language Monolingual Corpus and an MRD. TMI Proceedings of 8th International Conference on Theoretical and Methodological Issues in Machine Translation, Chester, UK, pp. 150-160.
- Lewis, T. (2015). Guidance on technology for translators,
- Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism.
- Marchisio, K. Guo, J. Lai, C.I. Koehn, P. (2019). Controlling the Reading Level of Machine Translation Output, Johns Hopkins University
- Sennrich, R. and Haddow, B., "Linguistic Input Features Improve Neural Machine Translation", 2016.
- Thames and Hudson. (1960). An introduction to machine translation
- Vogel, S. (2000). Statisticla methods for machine translation, *Research Gate*

- Wahlster, W. (2000). Verbmobil*:* Foundations of Speech-to-Speech Translation. *Artificial Intelligence.*
- Wu, Y. & Pan, Q. (2013). On the Development of Translation Technology, *Theory and Practice in Language Studies*, Vol. 3, No. 12, pp. 2240-2244
- Wu, Y. et al. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation."