ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΕΛΟΠΟΝΝΗΣΟΥ

ΥΠΟΛΟΓΙΣΤΙΚΗ ΓΛΩΣΣΟΛΟΓΙΑ

5η διάλεξη

Π. ΓΑΚΗΣ

ΥΠΟΛΟΓΙΣΤΙΚΗ ΓΛΩΣΣΟΛΟΓΙΑ

- Τι είναι η Υπολογιστική Γλωσσολογία
- Βασική έρευνα
- Πεδία εφαρμογής
- Εφαρμογές
- Συστήματα ερωταποκρίσεων στα ελληνικά
- **Αυτόματη Μετάφραση**
- **Υπολογιστική λεξικογραφία**

Αυτόματη επεξεργασία των φυσικών γλωσσών: ορισμός

- Αυτόματη επεξεργασία των φυσικών γλωσσών = **NLP: Natural Language Processing**
- Είναι μια επιστήμη που συνδυάζει την Πληροφορική, τη Γλωσσολογία και την Τεχνητή Νοημοσύνη.
- Φυσική γλώσσα = ανθρώπινη γλώσσα
- Απώτερος στόχος = κατανόηση της ανθρώπινης γλώσσας από τον υπολογιστή

Πεδία εφαρμογής

- Αυτόματη μετάφραση – MT: Machine Translation
- Ανάκτηση πληροφορίας – IR: Information Retrieval
- Εξαγωγή πληροφορίας – IE: Information Extraction
- Εξόρυξη δεδομένων – DM: Data Mining
- Αναγνώριση μερών του λόγου – Part-of-speech (POS) tagging
- Συντακτική ανάλυση – Parsing
- Αναγνώριση ονοματικών οντοτήτων – NER: Named Entities Recognition
- Μηχανική μάθηση – ML: Machine Learning
- Αυτόματη αναπαραγωγή κειμένων – NLG: Natural Language Generation
- Αναγνώριση προφορικού λόγου – Speech Recognition (Text-to-speech and Speech-to-text systems)
- …

Ανάκτηση και Εξαγωγή Πληροφορίας

- Δεδομένης μιας συλλογής κειμένων:
  - Η **ανάκτηση** πληροφορίας **(information retrieval)** επιλέγει ένα υποσύνολο κειμένων που (πιθανότατα) έχουν σχέση με ένα θέμα ή μία ερώτηση του χρήστη
  - Η **εξαγωγή** πληροφορίας **(information extraction)** θεωρεί ότι όλα τα κείμενα είναι σχετικά με το θέμα και εξάγει σχετική πληροφορία από τα κείμενα
- Η ανάκτηση και η εξαγωγή πληροφορίας είναι συμπληρωματικές τεχνολογίες

Διαφορά IR - IE

- Η διαφορά είναι ότι στο information retrieval έχουμε μια συλλογή και ψάχνουμε να βρούμε ποια κείμενα έχουν σχέση με τις λέξεις της ερώτησης.
- Στο information extraction έχουμε ένα κείμενο και προσπαθούμε να καταλάβουμε τι λέει.

Συστήματα Εξαγωγής Πληροφορίας

- Το κάθε σύστημα δίνει στη συντακτική επεξεργασία τη δική του βαρύτητα
  - Μερικά συστήματα δεν έχουν ξεχωριστή φάση συντακτικής επεξεργασίας
  - Άλλα συστήματα προσπαθούν να παράγουν ένα πλήρες συντακτικό δέντρο ανάλυσης
  - Τα περισσότερα συστήματα κάνουν κάτι ενδιάμεσο και παράγουν κομμάτια του δέντρου ανάλυσης (μερική συντακτική ανάλυση - partial parsing)

(Partial Parsing)

- Τα συστήματα που βασίζονται σε partial parsing κατασκευάζουν δομές για τις οποίες μπορούν να είναι σίγουρα είτε από συντακτικής είτε από σημασιολογικής άποψης
  - Για παράδειγμα, απλές ονοματικές φράσεις (άρθρο επίθετο ουσιαστικό) και ρηματικές φράσεις (ρήμα και βοηθητικά ρήματα) μπορούν να ανιχνευτούν εύκολα
  - Μεγαλύτερες δομές (πιο σύνθετες ονοματικές – ρηματικές φράσεις) μπορούν να κατασκευαστούν αν υπάρχει αρκετή σημασιολογική πληροφορία

- Στην εξαγωγή πληροφορίας
  - Δεν απαιτείται πλήρης κατανόηση του κειμένου
  - Δεν απαιτείται πλήρης συντακτική-σημασιολογική ανάλυση
  - Υπάρχει περιορισμός θέματος, ύφους
  - Τα συστήματα είναι αξιόπιστα και μπορούν να χειριστούν ακόμα και λάθος προτάσεις
  - Τα συστήματα μπορούν να αξιολογηθούν εύκολα

Ονοματικές οντότητες

- Αναγνωρίζονται διάφοροι τύποι κύριων ονομάτων και άλλες ειδικές μορφές οντοτήτων
  - (π.χ. ημερομηνίες, ποσά, διευθύνσεις)
- Οι ονοματικές οντότητες (named-entities) εμφανίζονται συχνά σε πολλά είδη κειμένων
- Η αναγνώριση και η ταξινόμησή τους διευκολύνει την περαιτέρω επεξεργασία

Αναγνώριση ονοματικών οντοτήτων

Τρομοκρατικές Ενέργειες

- Δίνεται μία συλλογή κειμένων γύρω από ειδήσεις τρομοκρατικών ενεργειών
- Για κάθε κείμενο καθόρισε:
  - Τον τύπο της ενέργειας
  - Την ημερομηνία
  - Την τοποθεσία
  - κτλ.
- Συμπλήρωσε μία βάση δεδομένων με αυτά τα στοιχεία (templates)

Είσοδος:

- 19 March - A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb - allegedly detonated by urban guerrilla commandos - blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).

Ονοματικές Οντότητες

Έξοδος:

- Incident type	bombing
- Date	March 19
- Location	El Salvador: San Salvador (city)
- Perpetrator	urban guerilla commandos
- Physical target	power tower
- Human target	-
- Effect on physical target	destroyed
- Effect on human target	no injury or death
- Instrument	bomb

Δυσκολίες

- Κανένα λεξικό δεν μπορεί να συμπεριλάβει όλα τα υπάρχοντα κύρια ονόματα, διευθύνσεις κτλ.
- Νέα κύρια ονόματα δημιουργούνται συνεχώς
- Η ίδια ονοματική οντότητα μπορεί να αναφέρεται με διάφορες παραλλαγές (Coca Cola - Coke)
- Συνήθως δημιουργούνται ακρωνύμια- συντομεύσεις για τα κύρια ονόματα
- Τα ακρωνύμια-συντομεύσεις δεν είναι πάντα κύρια ονόματα (π.χ. λ.χ. κτλ.)
- Η χρήση κεφαλαίων γραμμάτων δεν είναι πάντα κανόνας

Ασάφεια Ονομάτων - Οντοτήτων

- Ακόμα και αν αναγνωριστεί ένα όνομα-οντότητα συχνά είναι δύσκολο να ταξινομηθεί σωστά
  - Άνθρωπος ή εταιρεία: *Ford*, *Philip Morris*
  - Άνθρωπος ή τοποθεσία: *Jordan*, *JFK*
  - Άνθρωπος ή μήνας: *April*, *June*
  - Ακρωνύμιο ή οργανισμός: MRI (Magnetic Resonance Imaging, Mental Research Institute)
  - ...

Προσεγγίσεις

- Χειρονακτικοί κανόνες
  - Συνήθως αποδίδουν καλύτερα για εξειδικευμένες εφαρμογές
  - Δύσκολο να κατασκευαστούν
  - Domain-specific (Εξαρτώμενοι από την θεματική περιοχή)
- Μηχανική μάθηση
  - Μπορεί να προσαρμοστεί εύκολα σε νέα πεδία
  - Χρειάζεται domain-specific δεδομένα εκπαίδευσης
- Η αναγνώριση ονοματικών οντοτήτων είναι από τις πιο πετυχημένες εργασίες επεξεργασίας φυσικής γλώσσας (ακρίβεια ~95%)

Αυτόματη Μετάφραση

http://translate.google.com/

Μετάφραση – Η/Υ - web

- Πλεονεκτήματα:

1. Τυπογραφικά
2. Ηλεκτρονικά λεξικά (δίγλωσσα)
3. Βιβλιογραφικές πηγές (διαδικτυακά)
4. Ορολογική αναζήτηση
5. Συνεργασία με άλλους μεταφραστές διαδικτυακά

Ρόλος Διαδικτύου στη Μετάφραση

1. **Αναζήτηση πληροφορίας και έρευνα τεκμηρίωσης**: Μηχανές και μετα-μηχανές αναζήτησης, διαδικτυακοί κατάλογοι, έργα αναφοράς, διαδικτυακά λεξικά και γλωσσάρια, εργαλεία αναζήτησης σε λεξικά, βάσεις δεδομένων, σώματα κειμένων…
2. **Επεξεργασία κειμένων προς μετάφραση**: Συστήματα μηχανικής μετάφρασης, μεταφραστικές μνήμες, εργαλεία στοίχισης κειμένων, εξαγωγής ορολογίας…
3. **Επικοινωνία και συνεργασία**: Κοινότητες μεταφραστών, φόρα, ιστολόγια, κοινωνικά δίκτυα…
4. **Ενημέρωση και επιμόρφωση**: διαδικτυακά σεμινάρια (Webinars), εκπαιδευτικά βίντεο (tutorials), εγχειρίδια, συχνές ερωτήσεις άλλων χρηστών

Γιατί ο Η/Υ στη μετάφραση;

- Η αύξηση της ζήτησης
- Το είδος των κειμένων
- Η τυποποίηση της ορολογίας
- Η πίεση του χρόνου
- Η σχετικοποίηση της έννοιας της ποιότητας

Μεταφραστική τεχνολογία

- Η **μεταφραστική τεχνολογία** ή μεταφραστική (translation technology) δηλώνει το σύνολο των εργαλείων που επιτρέπουν την αυτοματοποίηση, άμεση ή έμμεση, ολική ή μερική, της διαδικασίας της μετάφρασης.
- Τα πεδία εφαρμογής των εργαλείων αυτών διευρύνονται συνεχώς προς πάσα κατεύθυνση και αφορούν όλα τα σύγχρονα μέσα πληροφορίας και επικοινωνίας: **την τηλεόραση** (συμβατική ή διαδικτυακή) **με τα εργαλεία υποτιτλισμού** και **μεταγλώττισης**, το **διαδίκτυο** με τα **εργαλεία επιχώριας προσαρμογής** (localization) προϊόντων και υπηρεσιών που προσφέρονται μέσω δικτυακών τόπων, την **τηλεφωνία** (συμβατική ή κινητή) με τα **εργαλεία αναγνώρισης** και **σύνθεσης φωνής** (φωνητική μετάφραση), τη **βιντεοσυνομιλία** με την ταυτόχρονη **διερμηνεία** ή ακόμη και την πολυμεσική **μετάφραση** (ήχος/κείμενο).
- Υπάρχουν λοιπόν ποικίλα συστήματα «μετάφρασης» όχι μόνο γραπτού λόγου αλλά και μεταφοράς από τον γραπτό λόγο στον προφορικό και αντίστροφα (text to text, text to speech, speech to text, speech to speech)

Μεταφραστική διαδικασία

- με κριτήριο τον βαθμό αυτοματισμού της μεταφραστικής διαδικασίας και την ανάγκη ή όχι ανθρώπινης παρέμβασης (βλ. Lehberger & Bourbeau 1988):

1. **πλήρως αυτόματη μετάφραση υψηλής ποιότητας** (fully automatic high quality translation – FAHQT)
2. **μηχανική μετάφραση με ανθρώπινη υποστήριξη** (human assisted machine translation – HAMT): Εδώ η ανθρώπινη υποστήριξη νοείται είτε ως προεπεξεργασία (αγγλ. pre-editing,) του προς μετάφραση κειμένου, είτε ως επιμέλεια (αγγλ. post-editing) του μεταφρασμένου κειμένου.
3. **ανθρώπινη μετάφραση με μηχανική/ηλεκτρονική υποστήριξη** (machine assisted human translation – MAHT

Μηχανική μετάφραση (ΜΜ)

- Με τον όρο **αυτόματη ή μηχανική μετάφραση** (αγγλ. machine translation) (εφεξής ΜΜ). **ορίζεται η αυτοματοποιημένη διαδικασία κατά την οποία μεταφέρεται ο γραπτός λόγος από μια γλώσσα-πηγή σε μια γλώσσα-στόχο**
- Η μετάφραση παράγεται **αυτόματα** από ένα **υπολογιστικό σύστημα** το οποίο διαχειρίζεται εξολοκλήρου τη μεταφραστική διαδικασία, χωρίς ανθρώπινη παρέμβαση (Σταύρου και Τζεβελέκου 2000, Σοφιανόπουλος 2009).
- Στη ΜΜ ο υπολογιστής δεν προσφέρει μόνο **μεταφράσεις λεξικών μονάδων**, αλλά παράγει **προτάσεις** ή **κείμενα**.
- Ο άνθρωπος χρήστης εισάγει ένα κείμενο πηγή στο σύστημα, πατάει το κουμπί για να ενεργοποιήσει τη λειτουργία της μετάφρασης και λαμβάνει στην έξοδο ένα κείμενο στόχο, χωρίς καθόλου να παρέμβει κατά τη διάρκεια της διαδικασίας
- Εδώ και δεκαετίες, η ΜΜ αποτελεί ένα από τα **δυσκολότερα επιστημονικά προβλήματα** της **υπολογιστικής γλωσσολογίας**, συγκεντρώνοντας το ενδιαφέρον επιστημόνων από διάφορους χώρους, όπως της πληροφορικής, της τεχνητής νοημοσύνης, της θεωρητικής και τυπικής γλωσσολογίας, της ψυχογλωσσολογίας, της λογικής και της φιλοσοφίας της γλώσσας.

Παραδείγματα συστημάτων μηχανικής μετάφρασης (ΜΜ

- διατίθενται δωρεάν στο διαδίκτυο και, μεταξύ των προσφερόμενων γλωσσών, περιλαμβάνουν τα ελληνικά:

1. **Google** https://translate.google.com/
2. **Bing** http://www.bing.com/translator/
3. **Babelfish** http://www.babelfish.com/
4. **Wordlingo** http://www.worldlingo.com/
5. **Systran** http://www.systranet.com/translate
6. **SDL** http://www.freetranslation.com/
7. **Microsoft** http://www.windowslivetranslator.com/

Πόροι ΜΜ

Χρησιμοποιούμε τα γλωσσολογικά δεδομένα για να αναλύσουμε τα κείμενα (parsing), δηλαδή:

- μορφολογική ανάλυση των κειμένων, με την εφαρμογή ηλεκτρονικών λεξικών (electronic dictionaries)
- συντακτική και σημασιολογική ανάλυση των κειμένων, με την εφαρμογή γραμματικών (local grammars)

Σενάριο χρήσης: μετάφραση ελεύθερου κειμένου (13/06/2015)

Η μηχανική μετάφραση ενός ποιήματος στα αγγλικά, που κέρδισε σε διαγωνισμό παιδικού τραγουδιού στ
Μ. Βρετανία, αναδεικνύει ξεκάθαρα τα όρια και τα προβλήματα της ΜΜ. Χρησιμοποιούμε το σύστημα SD
FreeTranslation:

GB 2005 **by Ian McMillan** Grumpy Rooney fights off the ball Shaky Tony smiles like a doll All the Queen's grandsons and their media tribe are glimpses of GB 2005	GB 2005 από τον Ian McMillan Κατσούφη Rooney καταπολεμά την μπάλα Κουνημένο Τόνι χαμόγελα σαν κούκλα όλες οι Queen's τα εγγόνια και τα media φυλή κλεφτές ματιές του GB 2005 Humpty Dumpty και λίγο Bo Peep είχαν εύκολη- μειώθηκαν, έχασαν τα πρόβατα
Humpty Dumpty and Little Bo Peep they had it easy; they fell, they lost sheep The nursery rhyme's over, it's hard to survive in the mosaic that's GB 2005	του φυτωρίου της προσαρμοζόμαστε, είναι δύσκολο να επιβιώσει στο μωσαϊκό που GB 2005

Πηγή κειμένου: The Guardian 23/09/2005, http://www.theguardian.com/books/2005/sep/23/poetry.features
Μετάφραση 1: SDL FreeTranslation [13/06/2015]

Παρατηρήσεις

- Το σύστημα μπερδεύει τα ρήματα με ουσιαστικά (smiles),
- αφήνει αμετάφραστες του κοινού λεξιλογίου (Queen, media),
- αντιγράφει τη σειρά των όρων από τη μία γλώσσα στην άλλη (όλες οι Queen's τα εγγόνια και τα media φυλή),
- παραλείπει ρήματα και άρθρα (στο μωσαϊκό που GB 2005),
- αποτυγχάνει να αναγνωρίσει τα ακρωνύμια και τα κύρια ονόματα (GB, Humpty Dumpty), καθώς και τα όρια, τη δομή και την πτώση των ονοματικών συνόλων (λίγο Bo Peep, Κατσούφη Rooney, Κουνημένο Τόνι).
- Η διάκριση της διάρθρωσης του κειμένου (τίτλος, στροφές) εξαλείφεται.
- Η στίξη άλλοτε αντιγράφεται πιστά (κόμματα) και άλλοτε προσαρμόζεται (άνω τελεία που γίνεται παύλα).

Δοκιμή χρησιμοποιώντας το σύστημα Babelfish (13/06/2015)

Original

GB 2005 by Ian McMillan Grumpy Rooney fights off the ball Shaky Tony smiles like a doll All the Queen\'s grandsons and their media tribe are glimpses of GB 2005 Humpty Dumpty and Little Bo Peep they h

---

Translation

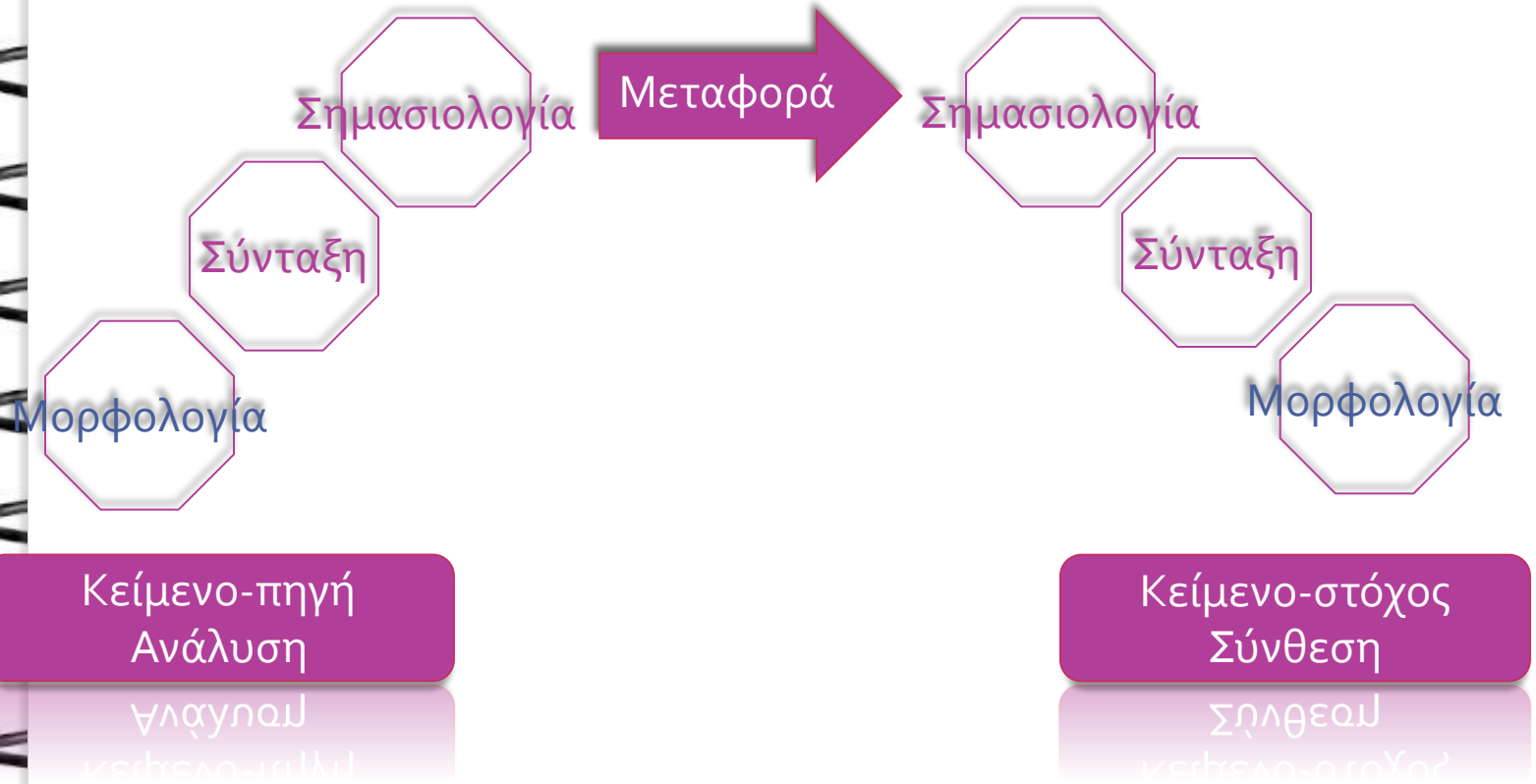
2005 GB από Ian McMillan γκρινιάρης Ρούνεϊ καταπολεμά την μπάλα Επισφαλής Tony χαμόγελα σαν μια κούκλα Το Queen\ εγγόνια και τους φυλή μέσα ενημέρωσης είναι αναλαμπές GB 2005 Humpty Dumpty και μικρή Bo Peep αυτοί h

Παρατηρήσεις

- Το σύστημα αποδεικνύεται ότι έχει συγκεκριμένο όριο λέξεων εισόδου, με αποτέλεσμα να κόβει το πρωτότυπο κείμενο (κάτι που δυστυχώς δεν αναφέρεται προειδοποιητικά στον χρήστη).
- Επιπλέον, το σύστημα δε λαμβάνει υπόψη τη μορφοποίηση του κειμένου εισόδου και ουσιαστικά την καταστρέφει. (*Στην προκειμένη περίπτωση πρόκειται για ποίημα, κάτι που δεν είναι πλέον αναγνωρίσιμο στο κείμενο εξόδου*).

Αυτόματη Μετάφραση

- Ruled-based machine translation (RBMT) – SYSTRAN



Αυτόματη Μετάφραση

- Statistical Machine Translation (SMT)

Προβλήματα:

- Ευθυγράμμιση προτάσεων (Sentence alignement)
- Σύνθετες λέξεις και ιδιωματικές εκφράσεις
- Μορφολογικοί τύποι
- Σύνταξη

Παραδοσιακή Μετάφραση

- Μετατροπή ενός κειμένου από μία φυσική γλώσσα σε μία άλλη
- Διαδικασία:
  - Διάβασμα του κειμένου σε μία φυσική γλώσσα
  - Κατανόηση του κειμένου
  - Παραγωγή κειμένου σε άλλη φυσική γλώσσα
- Δεν είναι μία απλή διαδικασία
  - Χρησιμοποιούνται άνθρωποι-ειδικοί για μεταφράσεις καλής ποιότητας

Μηχανική Μετάφραση (Machine Translation)

- Από τις πρώτες εφαρμογές της υπολογιστικής γλωσσολογίας (1950)
- Τεράστιες εμπορικές εφαρμογές
  - Η ΕΕ ξοδεύει πάνω από 1 δις € σε κόστη μετάφρασης κάθε χρόνο
- Πολύ δύσκολο πρόβλημα ειδικά για μετάφραση:
  - εντελώς αυτοματοποιημένη
  - πραγματικού χρόνου
  - ανοιχτού λεξιλογίου

Ιστορική αναδρομή

- 1950's
  - Έντονη δραστηριότητα - Απλά συστήματα - Ανταγωνισμός Αμερικής-Ρωσίας
- 1966: ALPAC report
  - Αρνητική αναφορά για την πρόοδο της έρευνας
  - Περικοπή κονδυλίων
- 1966-1975
  - Συστήματα 2ης γενιάς: πιο περίπλοκα από γλωσσολογικής και υπολογιστικής άποψης
- 1975-1985
  - Οι πρώτες επιτυχίες: Météo - Systran
- 1985-σήμερα
  - Eurotra, Ιαπωνικά συστήματα, Επανέναρξη έρευνας στην Αμερική
  - Πρώτα εμπορικά συστήματα (για PCs)
  - Στατιστικές μέθοδοι
  - Αμφιλεγόμενα αποτελέσματα, κριτική

Εμπορικό αλλά και ερευνητικό ενδιαφέρον

- Συνδυάζει πολλές τεχνολογίες επεξεργασίας φυσικής γλώσσας:
  - Αναγνώριση μερών του λόγου
  - Συντακτική ανάλυση
  - Σύνθεση
  - Άρση αμφισημίας λέξης
  - Αναγνώριση ονομάτων –οντοτήτων
  - Επίλυση ασάφειας αναφορών
  - Κατανόηση φυσικής γλώσσας
  - Αναπαράσταση γνώσης του κόσμου

Σήμερα

- Υπάρχουν αρκετά αξιόπιστα εμπορικά συστήματα
  - Μερικά αρκετά φτηνά ($50)
  - Χωρίς ιδιαίτερες υπολογιστικές απαιτήσεις (PC)
- Ελεύθερη μετάφραση μέσω WWW
  - Μετάφραση ιστοσελίδων και email
  - Η χαμηλή ποιότητα μετάφρασης είναι αποδεκτή
  - Καλύπτεται ένα μικρό μέρος των φυσικών γλωσσών
- Έρευνα στη μετάφραση προφορικού λόγου
  - Verbmobil

Βασικές ανάγκες εξυπηρέτησης από ΜΜ

(1) **Παραγωγή**: Πρόκειται για την παραγωγή μεταφράσεων που κρίνονται «δημοσιεύσιμες». Αυτό δε σημαίνει απαραίτητα ότι τα κείμενα πρόκειται στην πραγματικότητα να δημοσιευτούν, αλλά ότι τα κείμενα θα πρέπει να είναι ικανοποιητικής ποιότητα

(2) **Κατανόηση**: Πρόκειται για μετάφραση κειμένων για απλή παρακολούθηση, άντληση ή «φιλτράρισμα» πληροφοριών ή μετάφραση κειμένων από περιστασιακούς χρήστες (π.χ. μη εξειδικευμένο κοινό), όπου το εξαγόμενο από το σύστημα κείμενο δεν χρειάζεται να υποβληθεί σε περαιτέρω επεξεργασία (γενική ιδέα για το τι θέλει να πει το κείμενο)

Βασικές ανάγκες εξυπηρέτησης από ΜΜ

(3) **Επικοινωνία**: Στο πλαίσιο της διαγλωσσικής και πολύγλωσσης επικοινωνίας σε προσωπικό ή επαγγελματικό επίπεδο, διά αλληλογραφίας, μέσω ηλεκτρονικού ταχυδρομείου ή τηλεφώνου, η ποιότητα της μετάφρασης (ή/και η πιστότητα στο πρωτότυπο) μπορεί και πάλι να μην θεωρείται τόσο σημαντική, αρκεί οι άνθρωποι να καταφέρνουν στοιχειωδώς να παίρνουν τις πληροφορίες που θέλουν, να κατανοούν το μήνυμα που λαμβάνουν ή να μεταδίδουν αυτό που θέλουν να πουν

(4) **Ανεύρεση λεξικών ισοδυναμιών**: Η χρήση της ΜΜ ως λεξικού για την άντληση πληροφοριών από μια βάση γλωσσικών δεδομένων σε μια ξένη γλώσσα, την οποία ο χρήστης δεν κατανοεί αρκετά καλά, γεγονός που δείχνει την κυρίως πλέον χρήση των μεταφραστικών μηχανών για την αναζήτηση στο διαδίκτυο και την πρόσβαση σε ιστοσελίδες.

Χρησιμότητα Μηχανικής Μετάφρασης

- Σε εργασίες όπου μία πρόχειρη μετάφραση είναι επαρκής
  - Μετάφραση ιστοσελίδων
  - Διαγλωσσική ανάκτηση πληροφορίας
- Σε εργασίες όπου μπορεί να γίνει διόρθωση της αυτόματης μετάφρασης από κάποιον άνθρωπο-ειδικό
  - Human-assisted machine translation
- Σε εργασίες όπου επεξεργάζονται υπογλώσσες
  - Δελτία καιρού
  - Εγχειρίδια συσκευών

Babel Fish (http://babelfish.altavista.com/)

Babel Fish: Greek ➡ English

- Καλώς ήρθατε στο Τμήμα Πληροφορικής του Ιονίου Πανεπιστημίου. Το Τμήμα Πληροφορικής δημιουργήθηκε στο πλαίσιο του ΕΠΕΑΕΚ και λειτουργεί από το ακαδημαϊκό έτος 2004-05. Το Τμήμα δέχεται φοιτητές/τριες από το 2ο και 4ο επιστημονικό πεδίο και έχει ως αντικείμενο τόσο τη θεωρητική όσο και την εφαρμοσμένη Πληροφορική.

- Well you came in the Department of Information technology of Ionian University. The Department of Information technology was created in the frame of SPECIAL TRAINING PROGRAM and functions from the academic year 2004-05. the Department accepts students/trjes from the 2nd and 4th scientific field and has as object so much the theoretical what applied Information technology.

Babel Fish: English ➡ Greek

- Welcome to the Department of Informatics of the Ionian University. The Department of Informatics was founded by the Ministry of National Education and Religious Affairs in 2004 and its scope covers Theoretical as well as Applied Informatics.

- Υποδοχή στο τμήμα πληροφορικής του ιόνιου πανεπιστημίου. Το τμήμα πληροφορικής ιδρύθηκε από το Υπουργείο εθνικής παιδείας και θρησκευτικών υποθέσεων το 2004 και το πεδίο του καλύπτει τη θεωρητική καθώς επίσης και εφαρμοσμένη πληροφορική.

Προκλήσεις στην Αυτόματη Μετάφραση

Οι φυσικές γλώσσες διαφέρουν σε πολλά μεταξύ τους

- Μορφολογικές διαφορές
  - the → ο, η, το, τα, του, της, των, ...
- Αντωνυμίες
  - Σε πολλές γλώσσες (μορφολογικά πλούσιες) η αντωνυμία-υποκείμενο στην πρόταση εννοείται και τα μορφολογικά της χαρακτηριστικά καθορίζονται από την μορφολογία του ρήματος
  - Η κατάληξη του ρήματος στα Ισπανικά δείχνει ποιά αντωνυμία εννοείται
    - -o = I
    - -as = you
    - -a = he/she/it !!! (Ποιο θα επιλεγεί;)
    - -amos = we
    - -an = they

Προκλήσεις στην αυτόματη Μετάφραση

- Συντακτικές διαφορές (διάταξη των όρων)
  - language use → χρήση γλώσσας
    english(N1 N2) → greek(N2 N1)
  - the new house → la casa nueva
    english(DT J N) → spanish(DT N J)
  - IBM bought Lotus → IBM Lotus bought
    english(SUBJ V OBJ) → japanese(SUBJ OBJ V)
- Διαφορές στην έκφραση
  - Αγγλικά: *I am hungry* (είμαι πεινασμένος)
  - Γερμανικά: *Ich habe Hunger* (έχω πείνα)
  - Ελληνικά: *Πεινάω*

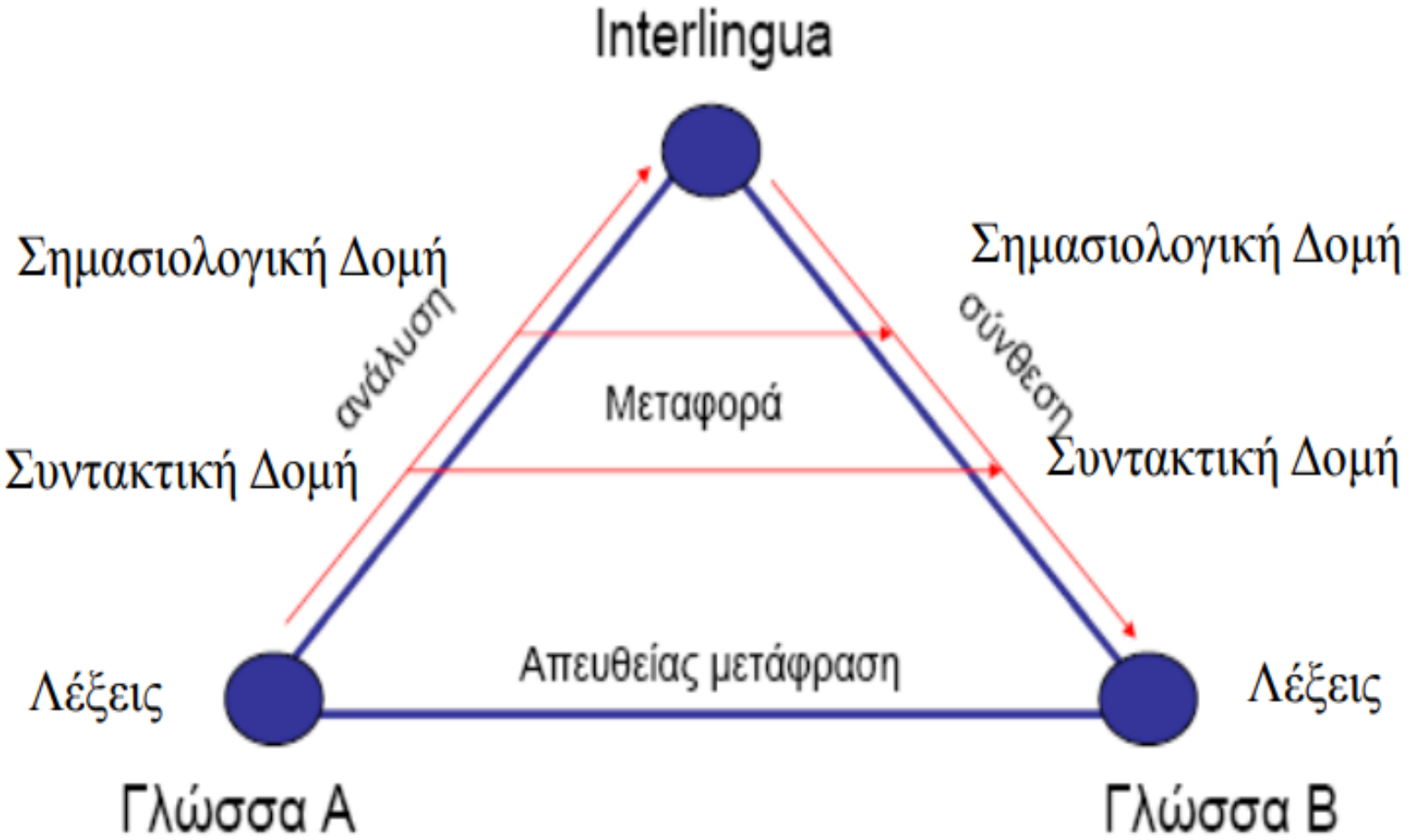
Προκλήσεις στην Αυτόματη Μετάφραση

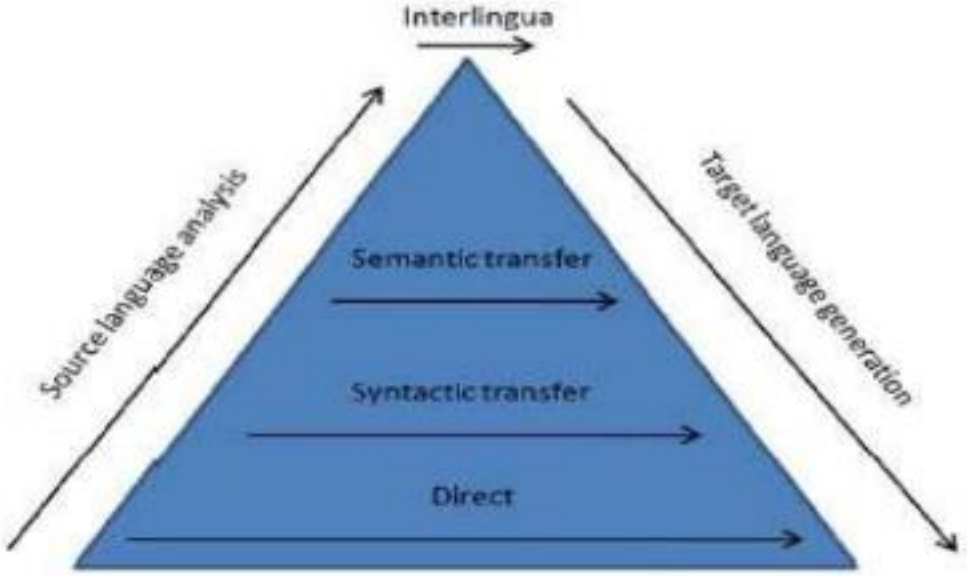
- Ο χρόνος των ρημάτων
  - I have been playing the piano for three years
  - Παίζω πιάνο τρία χρόνια
- Ιδιωματισμοί
  - He kicked the bucket → Πέθανε
  - She has always been a lame duck → Πάντα ήταν άχρηστη/βαρετή/ανίκανη

Κλασσικά Μοντέλα Μετάφρασης

- Interlingua
  - Για τη μετάφραση από μία γλώσσα Α σε μία γλώσσα Β χρησιμοποιείται ως ενδιάμεσο μία ουδέτερη γλώσσα (interlingua - αναπαράσταση νοήματος)
- Transfer (μεταφορά)
  - Για τη μετάφραση από μία γλώσσα Α σε μία γλώσσα Β ορίζεται μία διαδικασία ανάλυσης, μεταφοράς και σύνθεσης
- Direct (word-for-word) translation (απευθείας μετάφραση)
  - Για τη μετάφραση από μία γλώσσα Α σε μία γλώσσα Β γίνεται απευθείας μεταφορά από την μία στην άλλη

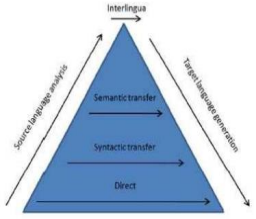
1. Τρίγωνο Vauquois





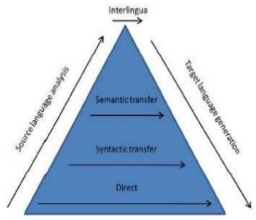
Η αριστερή πλευρά του τριγώνου αντιστοιχεί στη διαδικασία ανάλυσης του κειμένου στη γλώσσα-πηγή **ξεκινώντας από τη βάση του τριγώνου** και καταλήγοντας στην **κορυφή**, ενώ η δεξιά πλευρά αντιστοιχεί στη διαδικασία παραγωγής του κειμένου στη γλώσσα-στόχο, η οποία ξεκινά από την κορυφή της πυραμίδας και καταλήγει στο κάτω δεξιό άκρο.

Η πυραμίδα (τρίγωνο) τοι Vauquois



- Όσον αφορά την ανάλυση της γλώσσας, σύμφωνα με το τρίγωνο του Vauquois υπάρχουν **τρία διαφορετικά επίπεδα ανάλυσης**: η μορφολογία, η σύνταξη και η σημασιολογία.
- Ξεκινώντας από το κάτω μέρος του τριγώνου, **το πρώτο επίπεδο που συναντάμε είναι το μορφολογικό**, όπου η ανάλυση περιορίζεται στο επίπεδο της λέξης.
- Αμέσως παραπάνω, βρίσκεται η **ανάλυση σε επίπεδο δομών**, η οποία λαμβάνει υπόψη πληροφορίες που ξεπερνούν τα όρια της λέξης.
- Το τρίτο επίπεδο αντιστοιχεί στη σημασιολογική ανάλυση, η οποία προκύπτει μετά την ολοκλήρωση της αναγνώρισης της συντακτικής θέσης των συστατικών των προτάσεων του κειμένου στη γλώσσα-πηγή.

Η πυραμίδα (τρίγωνο) του Vauquois

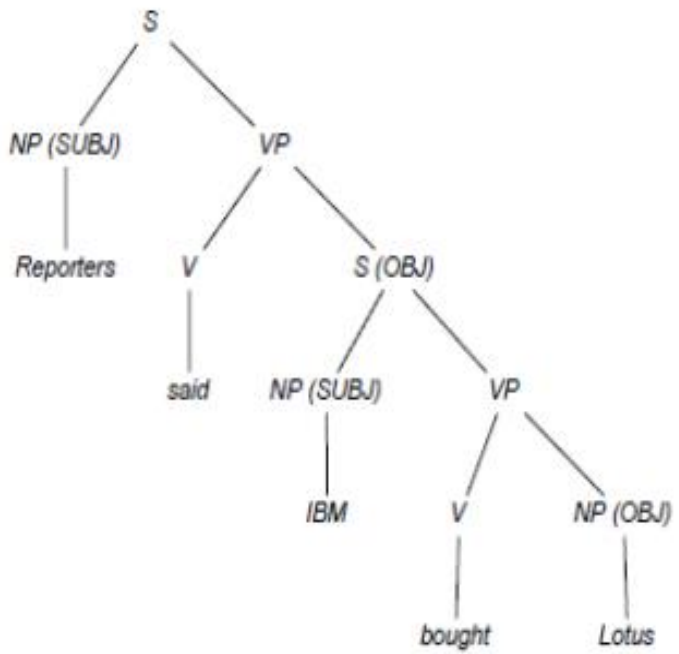


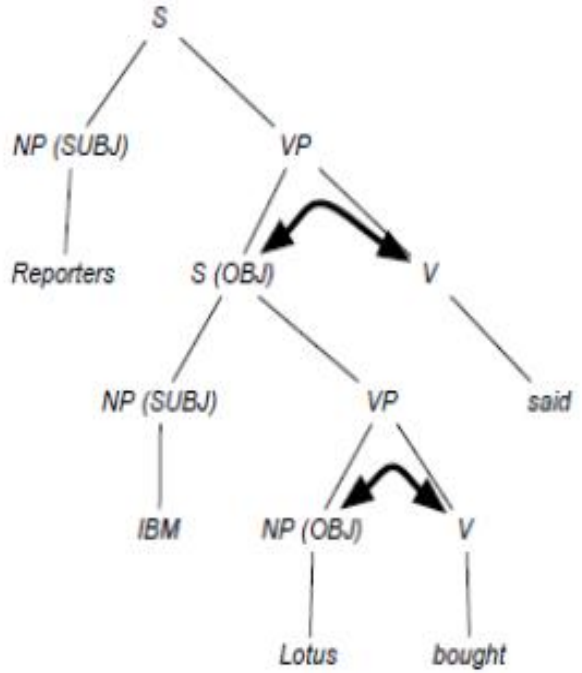
- Η αντιστοίχιση των γλωσσικών δεδομένων από τη μια γλώσσα στην άλλη μπορεί να πραγματοποιηθεί σε κάθε επίπεδο. Αν η μετάφραση γίνει λέξη προς λέξη (αγγλ. direct/literal translation) χωρίς να προηγηθεί συντακτική ανάλυση του κειμένου στη γλώσσα-πηγή, τότε η αντιστοίχιση γίνεται στο κατώτατο επίπεδο της γλώσσας (βέλος Direct).
- Αν για τη μετάφραση λαμβάνονται υπόψη συντακτικοί κανόνες των δύο υπό μελέτη γλωσσών, τότε η αντιστοίχιση γίνεται μετά την ολοκλήρωση της συντακτικής ανάλυσης του κειμένου στη γλώσσα-πηγή (βέλος Syntactic transfer). Σε αυτή την περίπτωση, έχει ήδη ολοκληρωθεί η μορφολογική ανάλυση.
- Ομοίως, το βέλος Semantic transfer αντιστοιχεί στη μετάφραση που έχει γίνει, ενώ έχει ήδη προηγηθεί σημασιολογική ανάλυση του κειμένου στη γλώσσα-πηγή.
- Στην **κορυφή του τριγώνου** υπάρχει ο όρος **interlingua** ο οποίος αναφέρεται σε ένα θεωρητικό κωδικοποιημένο γλωσσικό σύστημα που είναι κοινό για όλες τις φυσικές γλώσσες και το οποίο χρησιμεύει ως **ενδιάμεση τεχνητή γλώσσα** για τη μεταφορά του μηνύματος από μια φυσική γλώσσα σε μια άλλη.

1. Μοντέλο Interlingua

- Μειονεκτήματα:
  - Απαιτεί πολύ προσεκτικό σχεδιασμό της interlingua
  - Απαιτεί για κάθε γλώσσα την δυνατότητα μετάβασης από και προς την interlingua
  - Το μεγαλύτερο πρόβλημα είναι ότι πρέπει να αποσαφηνίσουμε εντελώς το νόημα σε κάθε περίπτωση
- Η interlingua μπορεί να είναι είτε μία τεχνητή γλώσσα (αναπαράσταση νοήματος) είτε μία τρίτη φυσική γλώσσα

2. Μοντέλο Μεταφοράς





- Η πρόταση-πηγή αναλύεται συντακτικά
- Πραγματοποιούνται οι απαραίτητες αναδιατάξεις των όρων
- Μεταφράζονται οι λέξεις

Μοντέλο Μεταφοράς

- Πραγματοποιείται ανάλυση του κειμένου-εισόδου
- Εφαρμόζονται κανόνες μετασχηματισμού της γλωσσολογικής δομής της εισόδου στην γλωσσολογική δομή της εξόδου
- Από την συντακτική δομή της εξόδου, παράγεται η πρόταση εξόδου (σύνθεση)
- Η διαδικασία της σύνθεσης μπορεί να παράγει πολλές εναλλακτικές εξόδους από τις οποίες επιλέγεται η καλύτερη
- Πλεονέκτημα: Αντιμετωπίζει το πρόβλημα της διάταξης
- Μειονέκτημα: Πρέπει να κατασκευαστούν κανόνες συντακτικών μετασχηματισμών για κάθε ζεύγος γλωσσών

3. Απευθείας Μετάφραση

- Πλεονεκτήματα
  - Συμπεριλαμβάνει μόνο μορφολογική ανάλυση
  - Γίνεται απλή μεταφορά των λέξεων από τη μία γλώσσα στην άλλη με χρήση μεγάλου δίγλωσσου λεξικού
- Μειονεκτήματα
  - Η πρόταση-μετάφραση απαιτεί αναδιάταξη των όρων
    - Σειρά ουσιαστικών-επιθέτων
    - Πρόσθεση/αφαίρεση άρθρων, προθέσεων
    - Μορφολογική σύνθεση

4. Συνδυασμός Μοντέλων

- Το Systran αποτελείται από 3 συστατικά:
  - Ανάλυση
    - Μορφολογική ανάλυση και αναγνώριση ΜΤΛ
    - Ανίχνευση ονοματικών και προθετικών φράσεων
    - Ρηχή συντακτική ανάλυση
  - Μεταφορά
    - Μετάφραση ιδιωματισμών
    - Άρση Αμφισημίας Λέξεων
    - Προσάρτηση προθετικών φράσεων
  - Σύνθεση
    - Χρήση δίγλωσσου λεξικού για την μετάφραση των λέξεων
    - Αναδιάταξη λέξεων
    - Μορφολογική σύνθεση

5. Στοχαστικές Μέθοδοι

- Σε αντίθεση με τις ορθολογιστικές προσεγγίσεις βάσει κανόνων, οι στοχαστικές μέθοδοι μετάφρασης στηρίζονται στα δεδομένα (κείμενα)
- Είναι η πιο ελπιδοφόρα προσέγγιση αφού υπάρχει πλέον
  - Μεγάλη αποθηκευτική ικανότητα
  - Μεγάλη επεξεργαστική ισχύς
  - Τεράστιες ποσότητες διαθέσιμων δεδομένων
- Δύο βασικές προσεγγίσεις
  - Στατιστική μηχανική μετάφραση
  - Μετάφραση βάσει παραδειγμάτων

Στατιστική Μηχανική μετάφραση

- Πλεονεκτήματα
  - Αντιμετωπίζει την αμφισημία
  - Αντιμετωπίζει ιδιωματισμούς
  - Απαιτεί ελάχιστη ανθρώπινη παρέμβαση
    - Υλοποίηση χωρίς μεγάλο οικονομικό και χρονικό κόστος
  - Μπορεί να υλοποιηθεί για οποιοδήποτε ζεύγος γλωσσών που διαθέτει δεδομένα εκπαίδευσης
- Μειονεκτήματα
  - Δεν αντιμετωπίζει ρητά συντακτικές δομές

Στατιστική Μηχανική Μετάφραση

On voit Jon à la télévision

	good English? P(E)	good match to French? P(F|E)
Jon appeared in TV.		✓
Appeared on Jon TV.		✓
In Jon appeared TV.		
Jon is happy today.	✓	
Jon appeared on TV.	✓	✓
TV appeared on Jon.	✓	
TV in Jon appeared.		
Jon was not happy.	✓	

Ευθυγράμμιση - Alignment

- Παράλληλα κείμενα
  - Τα ίδια κείμενα γραμμένα στις δύο γλώσσες
- Επιπλέον, τα κείμενα πρέπει να είναι ευθυγραμμισμένα (aligned)
  - Σε ποια πρόταση (ή προτάσεις) μιας γλώσσας αντιστοιχεί μια πρόταση της άλλης γλώσσας
  - Σε ποια λέξη/φράση μιας γλώσσας αντιστοιχεί μια λέξη/φράση της άλλης γλώσσας

Στοίχιση κειμένων (alignment)

- Όσον αφορά τη στοίχιση κειμένων, οι δυσκολίες που αντιμετωπίζει το σύστημα είναι δύο:

α) η κατάτμηση των δύο κειμένων και

β) η στοίχιση των τμημάτων κειμένου.

Στοίχιση κειμένων (alignment)

- Υπάρχουν δύο βασικές τεχνικές (Briel 2011):

1. Η πρώτη τεχνική αφορά την «**οπτική**» **στοίχιση κειμένων**. Το σύστημα εμφανίζει έναν δίστηλο πίνακα με το πρωτότυπο κείμενο και το μετάφρασμα, όπου έχει επιχειρήσει, άλλοτε με περισσότερη και άλλοτε με λιγότερη επιτυχία, να αντιστοιχίσει τα τμήματα κειμένου πηγής και στόχου. Στη συνέχεια, ο χρήστης μπορεί να επέμβει και να διορθώσει τη στοίχιση, συγχωνεύοντας ή διασπώντας ανάλογα τα τμήματα. Αυτό μπορεί να είναι **κοπιαστικό** και **χρονοβόρο**, ωστόσο με αυτό τον τρόπο το αποτέλεσμα είναι **τέλειο**, αφού διαμορφώνεται και ελέγχεται από τον χρήστη. Αυτή είναι συνήθως η καλύτερη επιλογή όταν ο όγκος των κειμένων δεν είναι τεράστιος και όταν απαιτείται, κατά το δυνατόν, μεγαλύτερη ακρίβεια στο αποτέλεσμα.
2. Η δεύτερη τεχνική αφορά την **«αυτόματη» στοίχιση κειμένων.**. Αν φυσικά χρειάζεται να στοιχιστεί ένα σώμα κειμένων 300.000 λέξεων, που θα αποτελέσουν σώμα αναφοράς, τότε δεν έχει μεγάλη σημασία αν μερικές προτάσεις βρεθούν χωρίς αντιστοιχία.

Παράλληλο Σώμα Κειμένων (Parallel Corpus)

what is more , the relevant cost dynamic is completely under control.	im übrigen ist die diesbezügliche kostenentwicklung völlig unter kontrolle .
sooner or later we will have to be sufficiently progressive in terms of own resources as a basis for this fair tax system .	früher oder später müssen wir die notwendige progressivität der eigenmittel als grundlage dieses gerechten steuersystems zur sprache bringen .
we plan to submit the first accession partnership in the autumn of this year .	wir planen , die erste beitrittspartnerschaft im herbst dieses jahres vorzulegen .
it is a question of equality and solidarity .	hier geht es um gleichberechtigung und solidarität .
the recommendation for the year 1999 has been formulated at a time of favourable developments and optimistic prospects for the european economy .	die empfehlung für das jahr 1999 wurde vor dem hintergrund günstiger entwicklungen und einer für den kurs der europäischen wirtschaft positiven perspektive abgegeben .
that does not , however , detract from the deep appreciation which we have for this report .	im übrigen tut das unserer hohen wertschätzung für den vorliegenden bericht keinen abbruch .

Παραδείγματα Παράλληλων Κειμένων

- Πρακτικά Καναδικής Βουλής
- Επίσημη Εφημερίδα Ευρωπαϊκής Ένωσης
- Αναφορές Ηνωμένων Εθνών
- Εγχειρίδια χρήσης συσκευών
- Νομοθεσία Hong-Kong, Macao
- ...

Αξιολόγηση ΜΜ

- Τα συστήματα ΜΜ μπορούν να έχουν καλύτερα αποτελέσματα όταν μεταφράζουν **σύντομα** και **τυποποιημένα κείμενα**.
- Επιπλέον, η ΜΜ είναι αρκετά πιο αποτελεσματική στην **τεχνική** και **επιστημονική μετάφραση**, παρά στην οικονομική, νομική ή λογοτεχνική μετάφραση.
- Αντίθετα, τα συστήματα ΜΜ **αδυνατούν** να αποδώσουν **ικανοποιητικά αποτελέσματα** σε κείμενα με δημιουργικό, αισθητικό ή καλλιτεχνικό χαρακτήρα (λογοτεχνικά, ποιητικά, διαφημιστικά, χιουμοριστικά), καθώς δεν μπορούν να αποδώσουν τις λεπτές νοηματικές αποχρώσεις ούτε τις ιδιαιτερότητες στο ύφος (π.χ. ειρωνεία, χιούμορ) ή στο επίπεδο λόγου.

Κριτική και Συμπεράσματα

- Είναι αξιοσημείωτο ότι ορισμένα από τα βασικότερα προβλήματα της ΜΜ παραμένουν άλυτα:

1. επίλυση αμφισημιών,
2. λανθασμένη επιλογή λέξεων γλώσσας-στόχου,
3. επιλογή γένους αντωνυμιών και άρθρων,
4. διατήρηση συντακτικών σχημάτων γλώσσας - πηγής,
5. προβλήματα συμφωνίας όρων της πρότασης,
6. προβλήματα με προτάσεις που περιέχουν δευτερεύουσες προτάσεις κ.ο.κ.

Κριτική και Συμπεράσματα

- Είναι πράγματι **περίεργο** πώς μετά από 50 χρόνια έρευνας στον χώρο της ΜΜ υπάρχουν εμπορικά συστήματα τα οποία εξακολουθούν να παράγουν λανθασμένη μορφολογία, λανθασμένη συμφωνία όρων πρότασης ή να τοποθετούν τα ρήματα στην αρχή ή στο τέλος της πρότασης

Κριτική και Συμπεράσματα

- Γεγονός όμως είναι ότι σήμερα ποικίλοι παράγοντες έχουν οδηγήσει τη ΜΜ σε μια δεύτερη περίοδο ακμής, οι οποίοι συνοψίζονται ως εξής (Μπαμπούρης & Τριανταφυλλοπούλου 2012):

1. Ο ανεπαρκής αριθμός μεταφραστών για τις ανάγκες της παγκόσμιας αγοράς.
2. Η ανάγκη διάσωσης γλωσσών απειλούμενων με εξαφάνιση.
3. Ζητήματα εθνικής κυριαρχίας, σχετιζόμενα κυρίως με την κατασκοπία και τον πόλεμο ενάντια στη διεθνή τρομοκρατία.
4. Ζητήματα μεταναστευτικής πολιτικής, σχετιζόμενα κυρίως με τη διαχείριση των μεταναστών.
5. Η ανάγκη για ταυτόχρονη κυκλοφορία ενός νέου προϊόντος σε όλα τα μήκη και τα πλάτη της παγκοσμιοποιημένης αγοράς.
6. Η χρήση ελεγχόμενων γλωσσών στην τεχνική κειμενογραφία.
7. Η ανάγκη μείωσης του κόστους της μετάφρασης.
8. Η ανάγκη διαγλωσσικής αναζήτησης στον Παγκόσμιο Ιστό.
9. Η ανάγκη ενημέρωσης σε πραγματικό χρόνο

«Κάθε μήνα, χρησιμοποιούν το Translate 500 εκατ. χρήστες», υποστηρίζει ο Macduff Hughes, engineering director του Google Translate. Με το 80%-90% του διαδικτύου να αφορά μόλις 10 γλώσσες, προσθέτει, η μετάφραση αποτελεί πλέον κρίσιμο κομμάτι της διαδικασίας μάθησης για πολλούς ανθρώπους.

Οι γλώσσες οι οποίες ακόμη δεν υποστηρίζονται από το Google Translate είναι οι παρακάτω:

- Αμαρικά
- Ασαμέζικα
- Καντώνα
- Τσερόκι
- Κορσικάνος
- Ντζόνγκχα
- Φρίσιαν
- Χαβανέζικα

- Το Google Translate, όπως και τα άλλα εργαλεία αυτόματης μετάφρασης, **έχει τα όριά του**. Η υπηρεσία περιορίζει τον αριθμό των παραγράφων και το εύρος των τεχνικών όρων που μπορεί να μεταφραστεί και ενώ μπορεί να βοηθήσει τον αναγνώστη να κατανοήσει το γενικό περιεχόμενο ενός κειμένου ξένης γλώσσας, **δεν παραδίδει πάντα ακριβείς μεταφράσεις**. Ένα συγκεκριμένο γραμματικό
- Η **γνώση της υποτακτικής** είναι σχεδόν ανύπαρκτη.
- Επιπλέον, η "αγένεια" σε δεύτερο πρόσωπο συχνά επιλέγεται (ανεξάρτητα από το πλαίσιο και τα συναλλακτικά ήθη). Ο λόγος για αυτό μπορεί να είναι ότι αυτές οι αποχρώσεις απαιτούν κάποια ανθρώπινη σκέψη και συναίσθημα, κάτι που δεν μπορεί να επιτευχθεί σε καμία αυτόματη μηχανή μετάφρασης ή εργαλείο.
- Η εναλλακτική άποψη είναι ότι υπάρχουν απλοί γραμματικοί κανόνες που θα μπορούσαν εύκολα να ενσωματωθούν στο λογισμικό για να βελτιωθούν τα πράγματα.

Β΄ ΜΕΡΟΣ

Βασική έρευνα

- Δημιουργία **υπολογιστικών λεξικών**
- Εμπλουτισμός και διαχείριση υπολογιστικών λεξικών
- Συντακτικο-σημασιολογική περιγραφή των ιδιοτήτων των ρημάτων με συμπλήρωμα που δηλώνει τον τόπο.
- Δημιουργία **γραμματικών** για τη συντακτική ανάλυση, μερική ή πλήρη, δομών

Υπολογιστικά λεξικά

- Με τον όρο «υπολογιστικά λεξικά» εννοούμε τα λεξικά που έχουν μια συγκεκριμένη δομή, η πληροφορία είναι κωδικοποιημένη και μπορούν να χρησιμοποιηθούν σε αναλυτές κειμένων (parsers).
- Διαφορά ηλεκτρονικών λεξικών και λεξικών σε ηλεκτρονική μορφή.

Γραμματική κατηγορία

- **ΑΝΟΙΚΤΕΣ - ΚΛΕΙΣΤΕΣ**
- Οι βασικές γραμματικές κατηγορίες της ΝΕ είναι τα απλά και σύνθετα ρήματα, ουσιαστικά, επίθετα και επιρρήματα. Αυτές οι γραμματικές κατηγορίες ονομάζονται **ανοικτές,** διότι μπορούν να δεχτούν νέα μέλη (νεολογισμοί).
- Οι υπόλοιπες κατηγορίες (οι προσδιοριστές, οι αντωνυμίες, οι προθέσεις, οι σύνδεσμοι, τα επιφωνήματα, τα μόρια) θεωρούνται **κλειστές**

Κατάλογος των γραμματικών κατηγοριών

- **N** (ουσιαστικό)
- **ADV** (επίρρημα)
- **ADJ** (επίθετο)
- **DET** (προσδιοριστής)
- **INTJ** (επιφώνημα)
- **CONJ** (σύνδεσμος)
- **PREP** (πρόθεση)
- **V** (ρήμα)
- **PRO** (αντωνυμία)
- **PART** (μόριο)

Το ρήμα

- Στις γλώσσες που διαθέτουν **απαρέμφατο,** λ.χ. γαλλική, ο τύπος αυτός επιλέγεται αυθαίρετα ως λημματικός τύπος.
- Στη ΝΕ, που δεν διαθέτει απαρέμφατο, ως λημματικός τύπος επιλέγεται αυθαίρετα το 1ο πρόσωπο του ενικού της οριστικής του ενεστώτα στην ενεργητική φωνή
- Στη ΝΕ το ρήμα διαθέτει 6 πρόσωπα, 17 χρόνους και 5 εγκλίσεις
- Στο μορφολογικό επίπεδο, υπάρχουν 7 απλοί χρόνοι (ενεστώτας και αόριστος οριστικής, υποτακτικής και προστακτικής καθώς και ο παρατατικός της οριστικής), 6 χρόνοι με μόρια και απλούς τύπους (*θα γράφω, θα γράψω, να γράφω, να γράψω, θα έγραφα*) και 7 χρόνοι με **βοηθητικά ρήματα** και απλούς τύπους (με ή χωρίς μόρια) (*έχω αγοράσει, είχα αγοράσει, θα έχω αγοράσει, θα είχα αγοράσει, έχοντας αγοράσει*).

Το όνομα: Ουσιαστικά και επίθετα

- Για τα ουσιαστικά και τα επίθετα, πρέπει να λάβουμε υπόψη μας όλα τα χαρακτηριστικά της ελληνικής γλώσσας, δηλαδή:
- Το **γένος** με τις τρεις τιμές του

  αρσενικό

  θηλυκό

  ουδέτερο
- Ουσιαστικά με δύο τιμές γένους (*η ψήφος – ο ψήφος, ο πλούτος – τα πλούτη*)

Το όνομα: Ουσιαστικά και επίθετα

- Τους 2 **αριθμούς**
  ενικός
  πληθυντικός

- Τις 4 **πτώσεις** (4 για τον ενικό και 4 για τον πληθυντικό αριθμό)
  ονομαστική
  γενική
  αιτιατική
  κλητική

Το όνομα: Ουσιαστικά και επίθετα

- Στα αρχαία ελληνικά υπήρχε και η **δοτική.** Ωστόσο, στα νέα ελληνικά, η πτώση αυτή χρησιμοποιείται μόνο σε **παγιωμένες εκφράσεις** (*δόξα τω Θεώ*), σε απλά και σύνθετα επιρρήματα (*πράγματι, εν ανάγκη, εν πάση περιπτώσει*) και σε **ημιπαγιωμένους φρασεολογισμούς** (*ελλείψει* + γενική, π.χ. *ελλείψει χρημάτων*, *βάσει* + γενική, π.χ. *βάσει των προαναφερθέντων*).

Κλειστές κατηγορίες

- Τα επιρρήματα, οι προθέσεις, οι σύνδεσμοι και τα επιφωνήματα παραμένουν αμετάβλητα/άκλιτα και **δεν παρουσιάζουν μεγάλες δυσκολίες** κατά την κατασκευή ενός ηλεκτρονικού μορφολογικού λεξικού.
- Αντιθέτως, τα ρήματα, τα ουσιαστικά και τα επίθετα πρέπει να συνοδεύονται από μορφολογικές κλιτικές πληροφορίες (τιμή γένους, αριθμό και πτώση για τα ουσιαστικά και τα επίθετα, φωνή, έγκλιση, χρόνο και πρόσωπο για τα ρήματα).

Επιρρήματα

- Τα επιρρήματα παραμένουν αμετάβλητα και μπορούν να είναι απλά (μονολεκτικά) ή σύνθετα (πολυλεκτικά).

  αργά,.ADV

  καθ'εκάστη,.ADV

  κατά λέξη,.ADV

- Επιρρηματική λειτουργία μπορούν να αναλάβουν πολλά είδη ακολουθιών, διευρύνοντας με τον τρόπο αυτό τον κατάλογο των επιρρημάτων

Η Μαρία διάβασε το βιβλίο νυχτιάτικα/τη νύχτα/κατά τη νύχτα/μες στη μαύρη νύχτα/τη νύχτα (που + όπου) έφυγε ο αδελφός της.

(ΜΟΡΦΟΛΟΓΙΚΗ ΔΗΛΩΣΗ???)

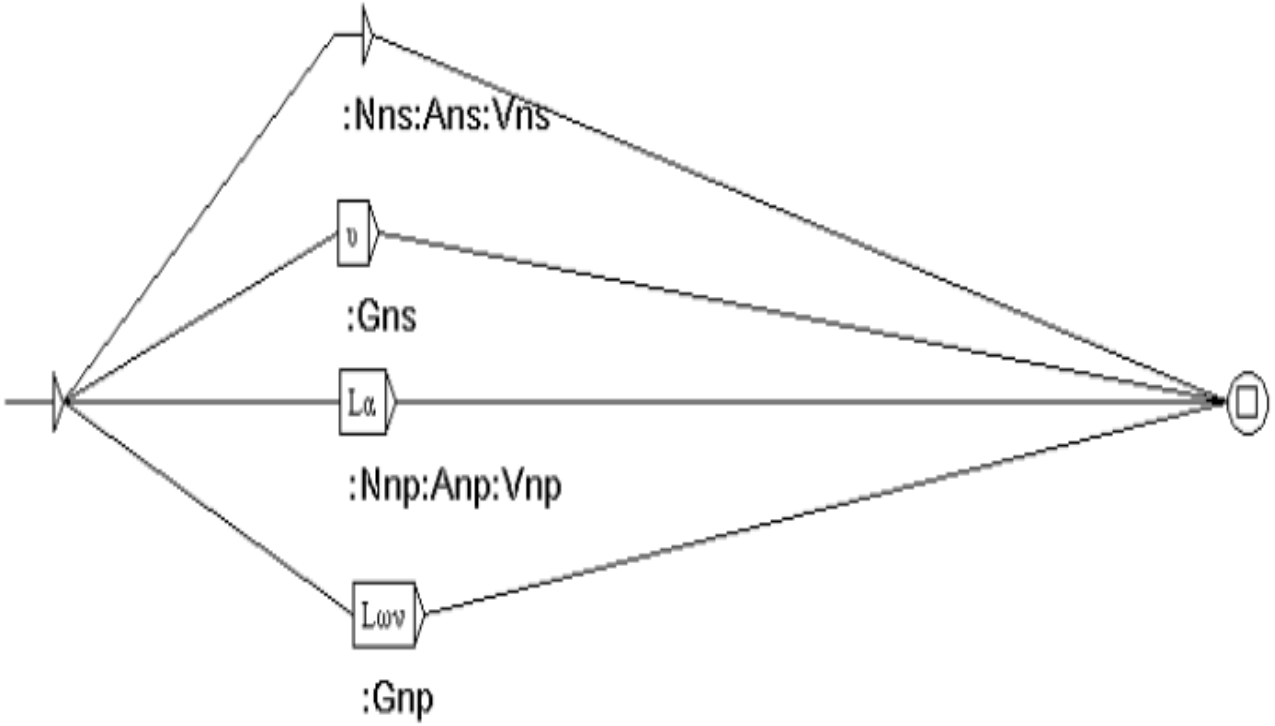
Οι προσδιοριστές και οι αντωνυμίες

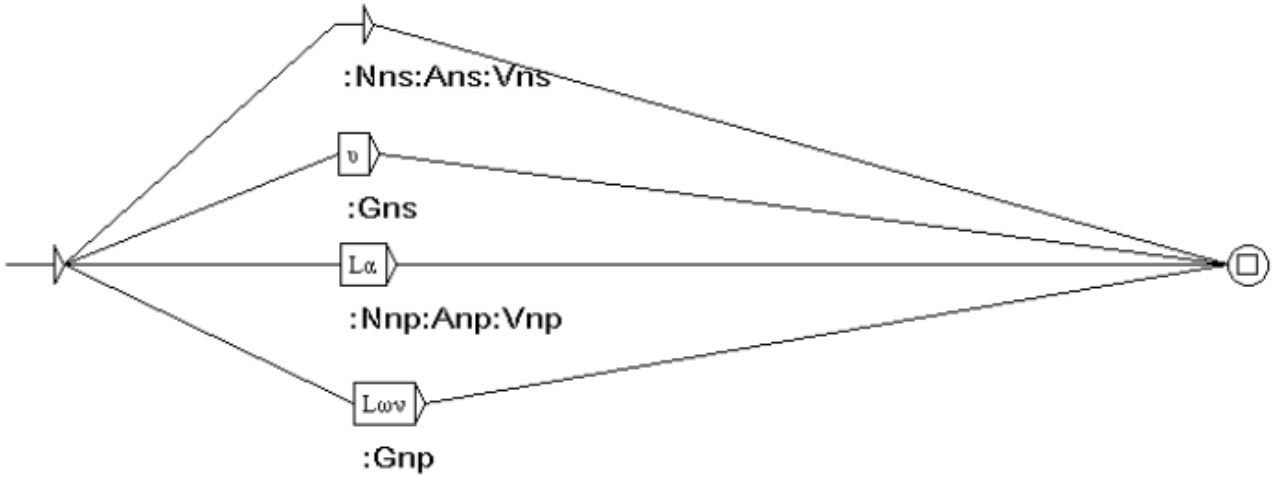
- Οι προσδιοριστές και οι αντωνυμίες της ΝΕ παρουσιάζουν κλιτικούς τύπους, όπως και τα ουσιαστικά, αλλά δεν έχουν μορφολογικά κοινό θέμα, π.χ. *εγώ - εμένα*. Γι' αυτόν τον λόγο, δεν τους αποδίδονται **κλιτικά διανύσματα,** και επομένως όλοι τους οι τύποι καταχωρίζονται στο λεξικό των κλιτικών τύπων χωρίς να προκύπτουν από το πρόγραμμα κλίσης

Ουσιαστικά

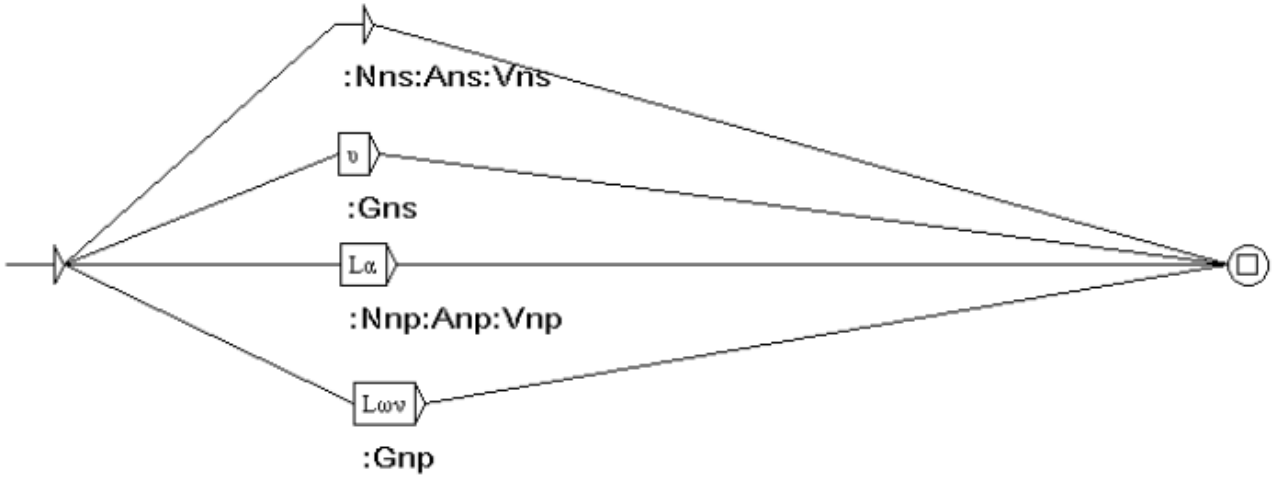
- Τα ουσιαστικά ταξινομήθηκαν σε **κλιτικά παραδείγματα** τα οποία χωρίστηκαν σε πέντε κατηγορίες: οι τρεις πρώτες αφορούν τις τρεις τιμές γένους (αρσενικό, θηλυκό, ουδέτερο), μια τέταρτη που περικλείει τα ουσιαστικά με διπλή τιμή γένους, δηλ. αρσενικό και θηλυκό (π.χ. *ο καθηγητής, η καθηγήτρια*), καθώς και μία ακόμη που περιλαμβάνει τα ουσιαστικά με αλλαγή γένους στον πληθυντικό αριθμό (π.χ. *ο πλούτος, τα πλούτη*).
- Επίσης τα δάνεια, συμμορφωμένα ή μη στο μορφολογικό σύστημα της ΝΕ, εντάσσονται στην κατάλληλη κατηγορία ανάλογα με την τιμή του γένους τους.

Κλιτικός γράφος για τα ουδέτερα ουσιαστικά σε -ο.





Στην προκειμένη περίπτωση, ο γράφος αποτελείται από τέσσερις διαφορετικές διαδρομές: η πρώτη διαδρομή δεν μεταβάλλει τον λημματικό τύπο παρά μόνο προσθέτει τις μορφολογικές πληροφορίες *Nns*, *Ans*, *Vns* για την ονομαστική, αιτιατική και κλητική του ουδετέρου στον ενικό αριθμό. Εδώ, καθώς πρόκειται για ουσιαστικό ουδετέρου γένους, οι τρεις αυτές πτώσεις του ενικού έχουν τον ίδιο κλιτικό τύπο. Η δεύτερη διαδρομή προσθέτει έναν επιπλέον χαρακτήρα (το -*υ*), προκειμένου να σχηματιστεί η γενική ενικού *λεωφορείου*



- Η τρίτη διαδρομή αποβάλλει ένα γράμμα μέσω του χαρακτήρα *L* (σε περίπτωση που πρέπει να απαλειφθούν περισσότερα γράμματα, τοποθετούμε και το αντίστοιχο πλήθος γραμμάτων L) και έπειτα προσθέτει την κατάληξη *-α* και τις μορφολογικές πληροφορίες *Nnp*, *Anp*, *Vnp* για την ονομαστική, αιτιατική, κλητική του πληθυντικού αριθμού αντίστοιχα. Τέλος, η τέταρτη διαδρομή του γράφου απαλείφει ένα γράμμα μέσω του χαρακτήρα *L* και προσθέτει την κατάληξη *-ων* και την πληροφορία *Gnp* για τη γενική ουδετέρου στον πληθυντικό αριθμό.

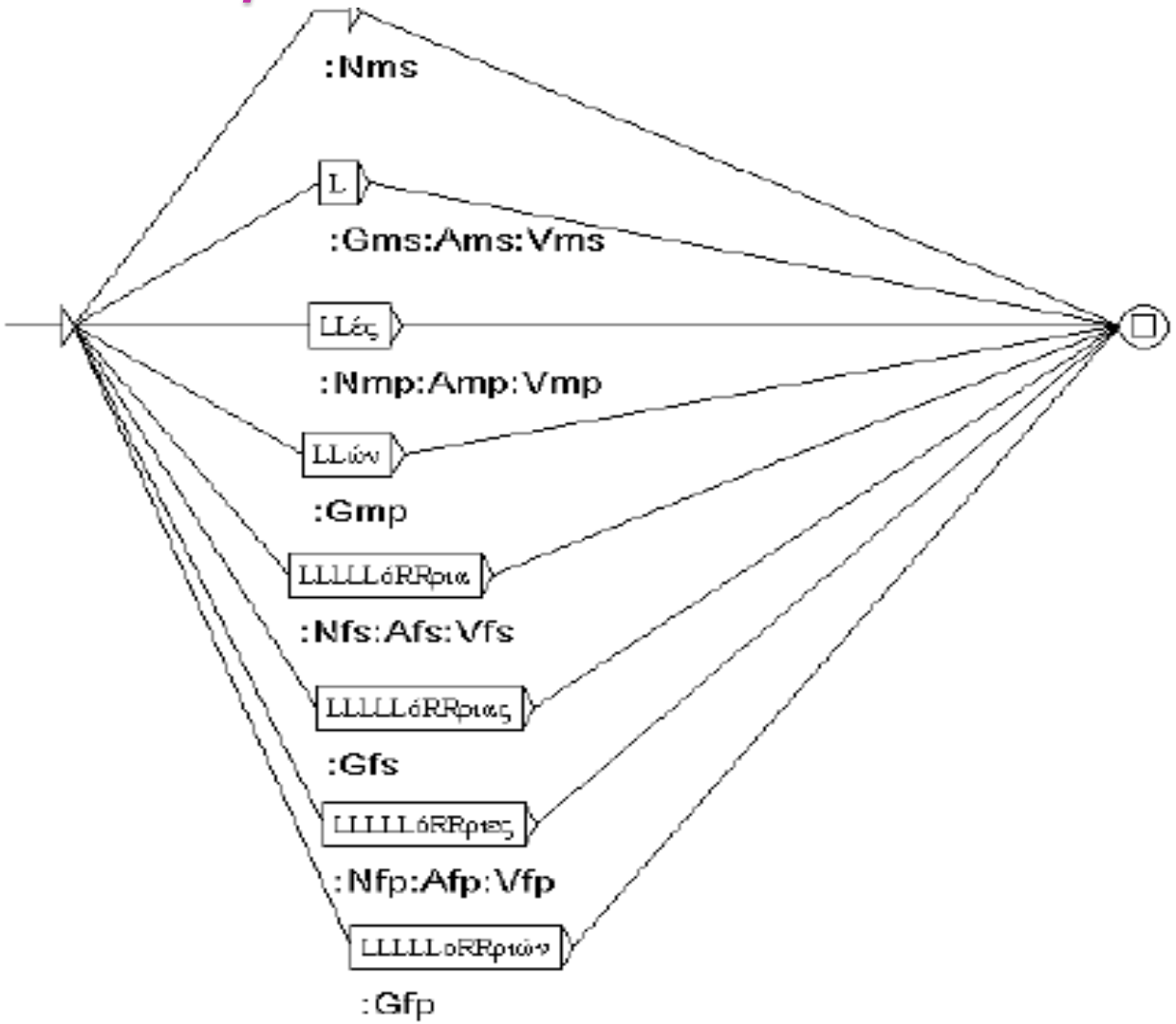
Κλιτική Παραγωγή

- Οι κλιτικοί τύποι που προκύπτουν από το παραπάνω πεπερασμένο αυτόματο είναι οι εξής:

  λεωφορείο,λεωφορείο.N:Vns
  λεωφορείο,λεωφορείο.N:Ans
  λεωφορείο,λεωφορείο.N:Nns
  λεωφορείου,λεωφορείο.N:Gns
  λεωφορεία,λεωφορείο.N:Vnp
  λεωφορεία,λεωφορείο.N:Anp
  λεωφορεία,λεωφορείο.N:Nnp
  λεωφορείων,λεωφορείο.N:Gnp

- Παρατηρούμε ότι στα αποτελέσματα που εξάγονται από το σύστημα αρχικά εμφανίζεται ο κλιτικός τύπος της εκάστοτε λεξικής μονάδας, στη συνέχεια ο λημματικός τύπος και έπονται οι μορφολογικές πληροφορίες

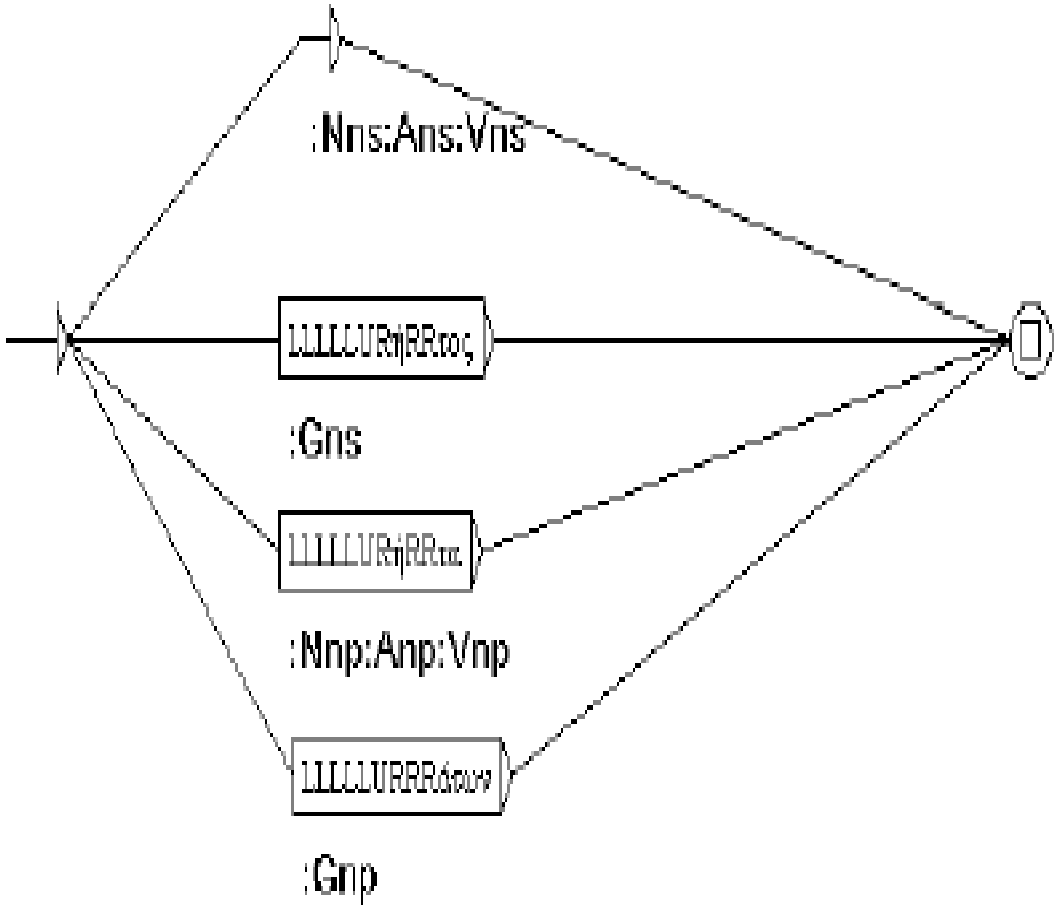
Κλιτικός Γράφος (*εθελοντής & εθελόντρια*



Κλιτικοί γράφοι

- Ο μεγάλος αριθμός των κλιτικών γράφων κυρίως για τα ουσιαστικά (περίπου 400) οφείλεται σε μεγάλο ποσοστό στον καταβιβασμό του τόνου στη γενική και στην αιτιατική πτώση του ενικού και του πληθυντικού αριθμού.
- Δεδομένου ότι για τον ηλεκτρονικό υπολογιστή **ένα άτονο γράμμα και ένα γράμμα με τόνο θεωρούνται δύο διαφορετικοί χαρακτήρες**, τα κλιτικά διανύσματα πρέπει να είναι με τέτοιο τρόπο τυποποιημένα, ώστε το σύστημα να «γνωρίζει» με σαφήνεια πού θα πρέπει να αποβάλει έναν άτονο χαρακτήρα και να τον αντικαταστήσει με έναν χαρακτήρα με τόνο και το αντίθετο.
- Για παράδειγμα, κατά την κλίση του ουσιαστικού *μάθημα* πρέπει να προβλέψουμε στη γενική του ενικού απαλοιφή του τόνου από την πρώτη συλλαβή, αντικατάσταση του άτονου γράμματος *-η* με το *-ή* και προσθήκη του κλιτικού διανύσματος *-τος*

Κλιτικός γράφος για τα ουδέτερα σε -μα.



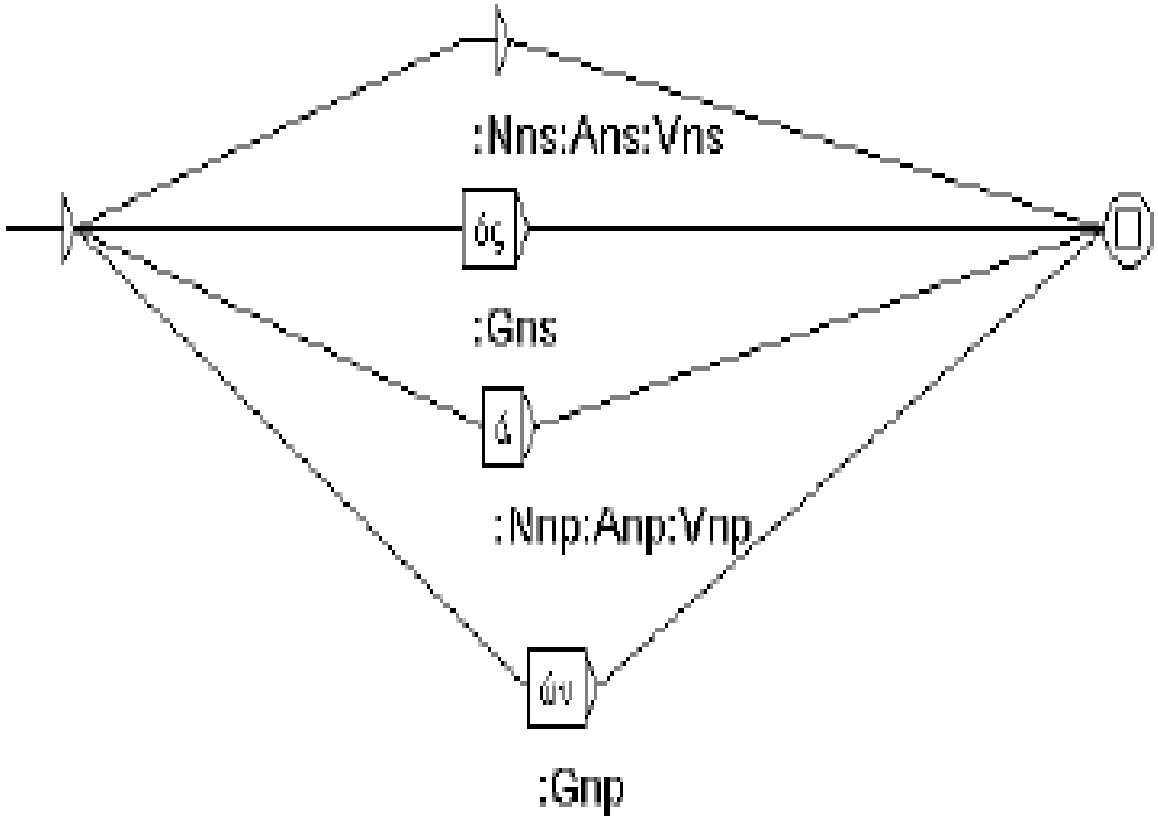
- Μια άλλη περίπτωση αφορά την εμφάνιση τόνου σε ονόματα τα οποία δεν τονίζονται στον λημματικό τύπο, αλλά εμφανίζουν τόνο κατά την κλίση τους. Λόγου χάρη:

  μυς → μυός

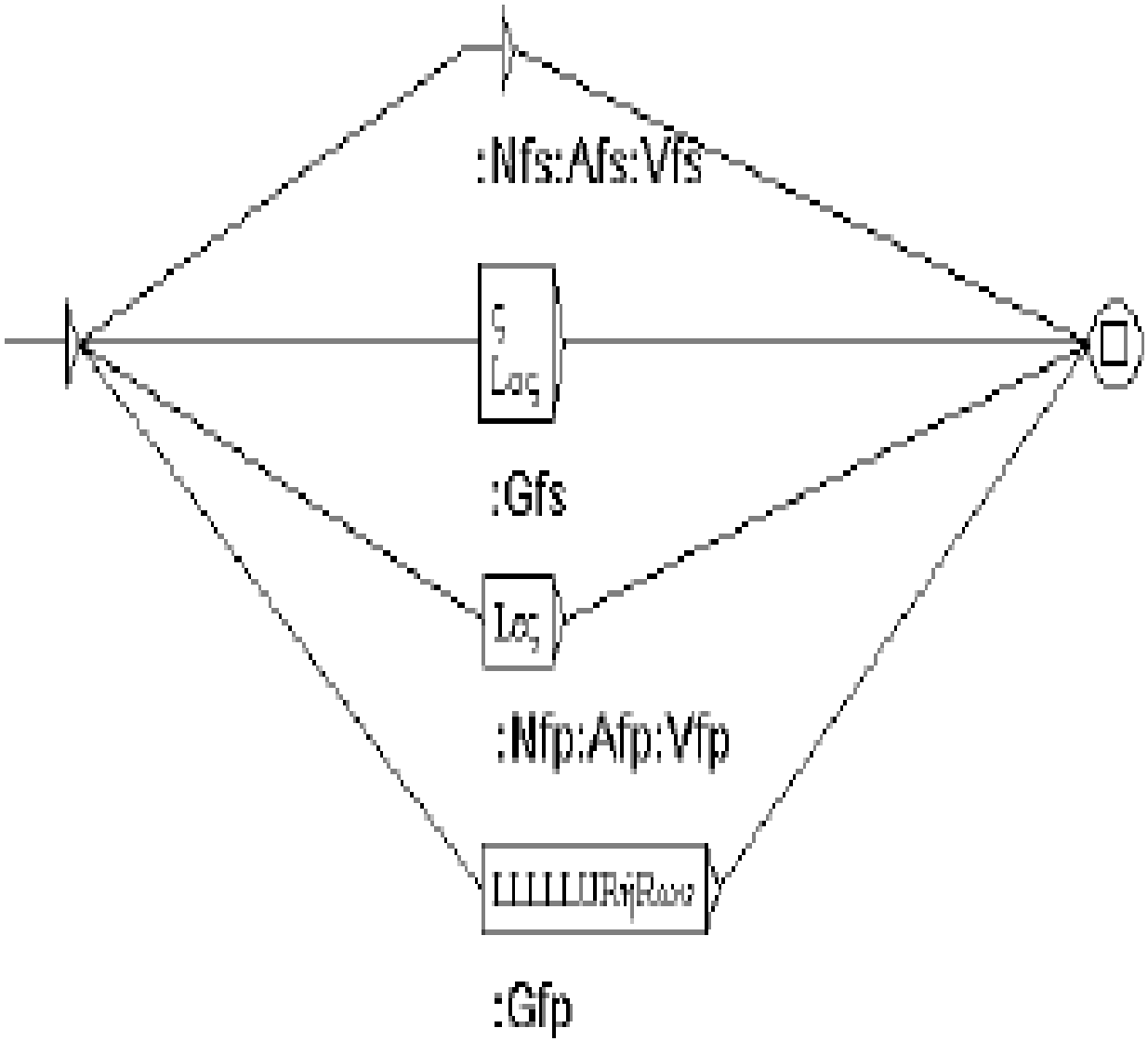
  φως → φωτός

  πυρ → πυρός

Κλιτικός γράφος για το ουσιαστικό πυρ



- Κατά την κλίση των λεξικών μονάδων πρέπει να αντιμετωπίσουμε επίσης τις περιπτώσεις παραλλαγών, όπως λ.χ. τη συνύπαρξη των **λόγιων** και **μη λόγιων τύπων,** π.χ. *μειονότητας* και *μειονότητος*. Για τις περιπτώσεις αυτές, προβλέπουμε και τους δύο τύπους μέσα στον κλιτικό γράφο

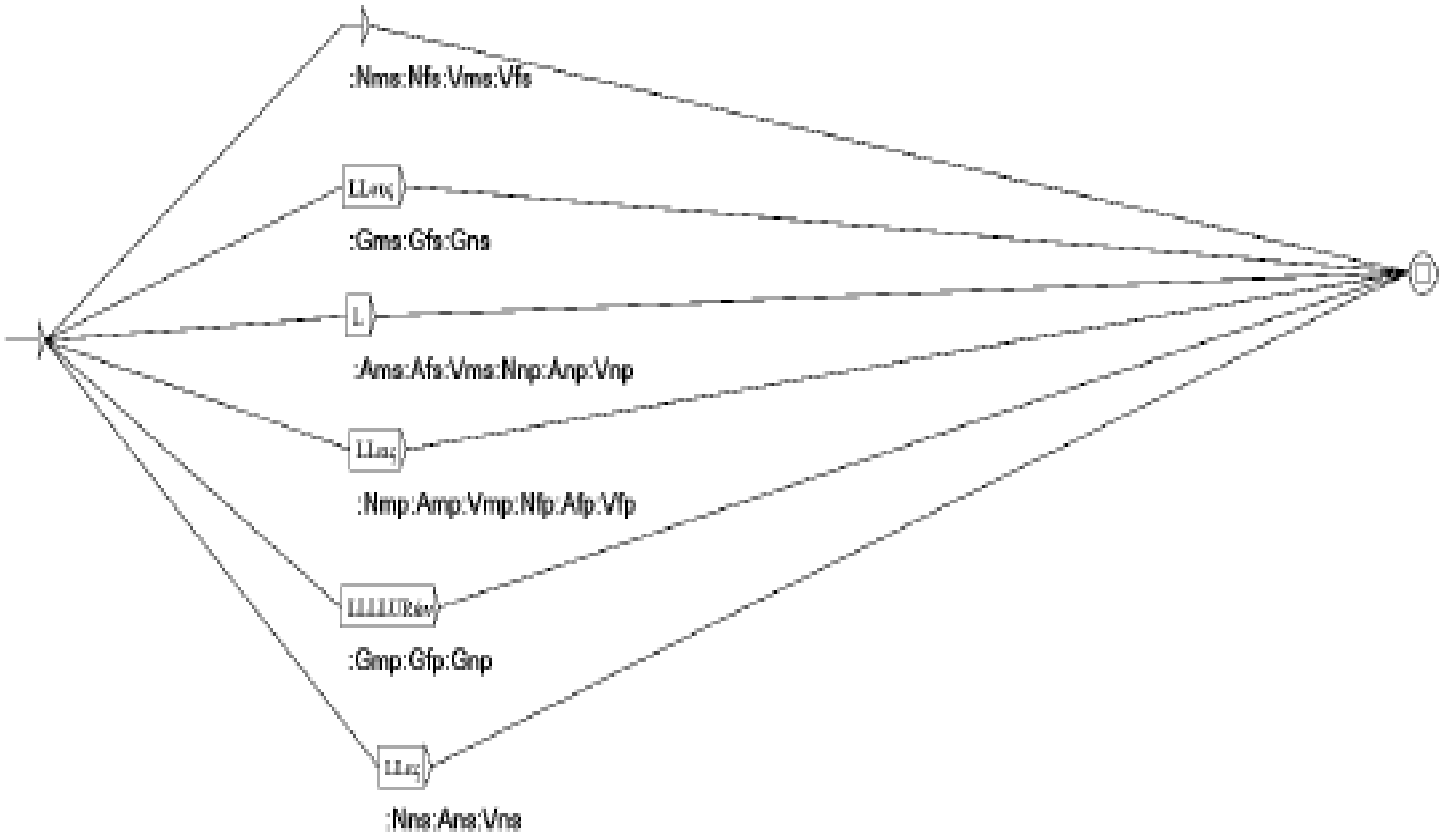

:Nfs:Afs:Vfs
:Gfs
:Nfp:Afp:Vfp
:Gfp

Επίθετα

- Όπως για τα ουσιαστικά, έτσι και για τα επίθετα κάθε κωδικός έχει το αντιπροσωπευτικό του κλιτικό παράδειγμα που συμπεριλαμβάνει όλες τις τιμές γένους του επιθέτου.
- Οι **παρατηρήσεις που προηγήθηκαν** αναφορικά με τον καταβιβασμό του τόνου και τις παραλλαγές, **αφορούν και τα επίθετα**, και πραγματοποιήθηκε ανάλογη επεξεργασία και γι' αυτή τη γραμματική κατηγορία.
- Ακολουθούν δύο παραδείγματα πεπερασμένων αυτομάτων για την κλίση των επιθέτων της ΝΕ, το πρώτο για τα επίθετα με τρεις τιμές γένους (π.χ. *οικονομικός, -ή, -ό*) και το δεύτερο για τα τριγενή και δικατάληκτα επίθετα (π.χ. *ο, η αγχώδης, το αγχώδες*):

Κλιτικός γράφος για τα επίθετα σε -ός, -ή, -ό.

Κλιτικός γράφος για τα επίθετα σε -ης, -ης, -ες.



Επίθετα

- Υπάρχει και μια κατηγορία επιθέτων που δεν παρουσιάζουν αλλαγές κατά την κλίση. Πρόκειται για δάνεια επίθετα που παραμένουν άκλιτα, επειδή δεν έχουν προσαρμοστεί στο μορφολογικό σύστημα της ΝΕ, όπως για παράδειγμα το επίθετο *ποπ*. Ωστόσο, και τα επίθετα αυτά συνοδεύονται από τις αντίστοιχες μορφολογικές πληροφορίες

Άκλιτα Επίθετα

- *ποπ,ποπ.A:Vnp*
- *ποπ,ποπ.A:Anp*
- *ποπ,ποπ.A:Gnp*
- *ποπ,ποπ.A:Nnp*
- *ποπ,ποπ.A:Vfp*
- *ποπ,ποπ.A:Afp*
- *ποπ,ποπ.A:Gfp*
- *ποπ,ποπ.A:Nfp*
- *ποπ,ποπ.A:Vmp*
- *ποπ,ποπ.A:Amp*
- *ποπ,ποπ.A:Gmp*
- *ποπ,ποπ.A:Nmp*
- *ποπ,ποπ.A:Vns*
- *ποπ,ποπ.A:Ans*
- *ποπ,ποπ.A:Gns*
- *ποπ,ποπ.A:Nns*
- *ποπ,ποπ.A:Vfs*
- *ποπ,ποπ.A:Afs*
- *ποπ,ποπ.A:Gfs*
- *ποπ,ποπ.A:Nfs*
- *ποπ,ποπ.A:Vms*
- *ποπ,ποπ.A:Ams*
- *ποπ,ποπ.A:Gms*
- *ποπ,ποπ.A:Nms*

Ρήματα

- Όπως και για τα ουσιαστικά και τα επίθετα, για την κατασκευή του λεξικού των κλιτικών ρηματικών τύπων απαιτείται:

1. η παραγωγή όλων των μορφολογικών ποικιλιών των ρημάτων,
2. η γραμματική ταυτοποίηση (έγκλιση, χρόνος, πρόσωπο και αριθμός) κάθε κλιτικού τύπου,
3. ο συνδυασμός κάθε κλιτικού τύπου με τις γραμματικές του πληροφορίες και τον αντίστοιχο λημματικό τύπο.

Ρήμα

- Η αυτόματη παραγωγή των ρηματικών κλιτικών τύπων προϋποθέτει:

1. τη **συστηματική τυποποίηση** των γλωσσικών δεδομένων,
2. τη **δημιουργία ενός προγράμματος επεξεργασίας αυτών** των δεδομένων για την κατασκευή του ηλεκτρονικού λεξικού των κλιτικών τύπων.

Ρήμα

- Σχηματίζεται με μια σειρά καταλήξεων που αντιστοιχούν σε ένα συγκεκριμένο χρόνο και έγκλιση. Με βάση τη σειρά αυτή, το πρόγραμμα υπολογίζει όλα τα πρόσωπα του χρόνου αυτού.
- Στο τέλος, κάθε λημματικός τύπος εμφανίζεται με τις μορφολογικές πληροφορίες του, αφού πρώτα ομαδοποιηθούν οι όμοιοι κλιτικοί τύποι.

Κλιτικοί τύποι του ρήματος *λερώνω*

λερώνω,λερώνω.V:W:P1s:D1s:T1s
λερώνοντας,λερώνω.V:G
λερωμένος,λερώνω.V:Kms
λερωμένη,λερώνω.V:Kfs
λερωμένο,λερώνω.V:Kns
λερωμένοι,λερώνω.V:Kmp
λερωμένες,λερώνω.V:Kfp
λερωμένα,λερώνω.V:Knp
λερώνεις,λερώνω.V:P2s:D2s:T2s
λερώνει,λερώνω.V:P3s:D3s:T3s
λερώνουμε,λερώνω.V:P1p:D1p:T1p:Y1p
λερώνετε,λερώνω.V:P2p:D2p:T2p:Y2p
λερώνουν,λερώνω.V:P3p:D3p:T3p

Attributes

V=Ρήμα, W=Λημματικός τύπος ρήματος, P1s=Ενεστώτας 1° πρόσωπο ενικού αριθμού, D1s=Εξακολουθητικός μέλλοντας 1° πρόσωπο ενικού αριθμού, T1s=Υποτακτική ενεστώτα 1° πρόσωπο ενικού αριθμού, V:G=Ενεργητική μετοχή, V-Kms=Παθητική μετοχή αρσενικού γένους ενικού αριθμού, V:Kfs=Παθητική μετοχή θηλυκού γένους ενικού αριθμού, V:Kns=Παθητική μετοχή ουδετέρου ενικού αριθμού, V:Kmp=Παθητική μετοχή αρσενικού γένους πληθυντικού αριθμού, V:Kfp=Παθητική μετοχή θηλυκού γένους πληθυντικού αριθμού, V:Knp=Παθητική μετοχή ουδετέρου πληθυντικού αριθμού, P2s=Ενεστώτας 2° πρόσωπο ενικού αριθμού, D2s=Εξακολουθητικός μέλλοντας 2° πρόσωπο ενικού αριθμού, T2s=Υποτακτική ενεστώτα 2° πρόσωπο ενικού αριθμού, P3s=Ενεστώτας 3° πρόσωπο ενικού αριθμού, D3s=Εξακολουθητικός μέλλοντας 3° πρόσωπο ενικού αριθμού, T3s=Υποτακτική ενεστώτα 3° πρόσωπο ενικού αριθμού, P1p=Ενεστώτας 1° πρόσωπο πληθυντικού αριθμού, D1p=Εξακολουθητικός μέλλοντας 1° πρόσωπο πληθυντικού αριθμού, T1p=Υποτακτική ενεστώτα 1° πρόσωπο πληθυντικού αριθμού, Y1p=Προστακτική ενεστώτα 1° πρόσωπο πληθυντικού αριθμού, P2p=Ενεστώτας 2° πρόσωπο πληθυντικού αριθμού, D2p=Εξακολουθητικός μέλλοντας 2° πρόσωπο πληθυντικού αριθμού, T2p=Υποτακτική ενεστώτα 2° πρόσωπο πληθυντικού αριθμού, Y2p=Προστακτική ενεστώτα 2° πρόσωπο πληθυντικού αριθμού, P3p=Ενεστώτας 3° πρόσωπο πληθυντικού αριθμού, D3p=Εξακολουθητικός μέλλοντας 3° πρόσωπο πληθυντικού αριθμού, T3p=Υποτακτική ενεστώτα 2° πρόσωπο πληθυντικού αριθμού.

http://www.domeczek.pl/~natko/teaching/wk2011s/greek-tagset.txt

TAG	DESCRIPTION
ABBR	abbreviation
AdXxBa	adverb base form
AdXxCp	adverb comparative form
AdXxSu	adverb superlative form
AjBaFeNvNv	adjective base feminine
AjBaFePlAc	adjective base feminine plural accusative
AjBaFePlGe	adjective base feminine plural genitive
AjBaFePlNm	adjective base feminine plural nominative
AjBaFeSgAc	adjective base feminine singular accusative
AjBaFeSgDa	adjective base feminine singular dative
AjBaFeSgGe	adjective base feminine singular genitive
AjBaFeSgNm	adjective base feminine singular nominative
AjBaMaNvNv	adjective base masculine
AjBaMaPlAc	adjective base masculine plural accusative
AjBaMaPlGe	adjective base masculine plural genitive
AjBaMaPlNm	adjective base masculine plural nominative
AjBaMaPlVo	adjective base masculine plural vocative
AjBaMaSgAc	adjective base masculine singular accusative
AjBaMaSgGe	adjective base masculine singular genitive
AjBaMaSgNm	adjective base masculine singular nominative
AjBaMaSgVo	adjective base masculine singular vocative
AjBaNeNvNv	adjective base neutral
AjBaNePlAc	adjective base neutral plural accusative
AjBaNePlDa	adjective base neutral plural dative
AjBaNePlGe	adjective base neutral plural genitive
AjBaNePlNm	adjective base neutral plural nominative
AjBaNeSgAc	adjective base neutral singular accusative
AjBaNeSgDa	adjective base neutral singular dative
AjBaNeSgGe	adjective base neutral singular genitive

Ευχαριστώ για την προσοχή σας