Συστήματα κ' Τεχνολογίες Γνώσης –
Εργασίες στην Επεξεργασία Φυσικής Γλώσσας

1. Διορθωτής Λέξεων

Αντικείμενο – Στόχος

Σκοπός της άσκησης είναι ο σχεδιασμός και η υλοποίηση συστήματος **διορθωτή λέξεων** βασισμένου στην **prolog** το οποίο:

- διαβάζει κείμενο από το πληκτρολόγιο ή από αρχείο
- ελέγχει την ορθότητα κάθε λέξης, και
- σε περίπτωση ορθογραφικού λάθους την αντικαθιστά με τη διορθωμένη λέξη.

Τα λάθη

Τα λάθη που θα μπορεί να εντοπίσει το σύστημα είναι τα εξής:

- **α.** ένα γράμμα είναι εσφαλμένο (διαφορετικό από το ορθό). *Π.χ. "πεδιοχή" αντί "περιοχή"*
- **β.** αναγραμματισμός (αντιμετάθεση 2 γειτονικών γραμμάτων). *Π.χ. "ονόμταος" αντί "ονόματος".*
- **γ.** παράλειψη ενός γράμματος (σε οποιοδήποτε σημείο της λέξης: αρχή, μέση ή τέλος). *Π.χ. "χουν" αντί "έχουν", "ελέχου" αντί "ελέγχου", "δέλεα" αντί "δέλεαρ".*

Λεξικό

Το σύστημα διαθέτει λεξικό με (τουλάχιστον 100) νόμιμες λέξεις .

Συμβάσεις – Υποδείξεις

- Οι λέξεις που ελέγχονται αποτελούνται μόνο από πεζά ελληνικά στοιχεία.
- *(Σ' αυτή την εφαρμογή δεν θεωρούνται λέξεις: α1, άνω-κάτω, Ε.Υ.Δ.Α.Π., κ.λπ., π.χ., Π.Χ. ...)*
- Οι λέξεις που ελέγχονται μπορεί να ανήκουν σε οποιαδήποτε κλιτική μορφή (γένος, αριθμό, πτώση, πρόσωπο, χρόνο, έγκλιση). Γιαυτό, το λεξικό θα πρέπει να περιέχει όλες τις μορφές της κάθε λέξης (πτώσεις, πρόσωπα κ.λπ.).
- Οι λέξεις που δεν αναγνωρίζονται (δεν περιέχονται στο λεξικό) αφήνονται ως έχουν.
- Δεν ελέγχεται η συντακτική δομή της πρότασης (θεωρώντας την ορθή).
- Σε περίπτωση αμφισημίας (όπου μια λέξη επιδέχεται διόρθωση με δύο ή περισσότερες λέξεις) η διόρθωση γίνεται με την πρώτη εκδοχή που θα συναντήσει το πρόγραμμά σας.

Φορμαλισμός – Υπολογιστικό περιβάλλον

Χρησιμοποιείστε φορμαλισμό **Ενοποιητικής Γραμματικής** της **Prolog** .

(Δείτε το GULP 3.1 στο «εκπαιδευτικό υλικό» στην ιστοσελίδα http://glotta.ntua.gr/courses/NLP-Seminar/index.html*: Επέκταση της Prolog για Ενοποιητικές Γραμματικές (Unification-Based Grammar)*

Επέκταση του συστήματος

- Διόρθωση λέξεων με **κεφαλαία** στοιχεία: Το πρώτο ή όλα τα γράμματα της λέξης μπορεί να είναι κεφαλαία.
- Σε περίπτωση **αμφισημίας** (όπου μια λέξη επιδέχεται διόρθωση με δύο ή περισσότερες λέξεις) θα προτείνονται όλες οι διορθωμένες λέξεις, με κάποιο φιλικό τρόπο. *Π.χ.:*
  - ***α.*** *Για την εσφαλμένη ορθογραφικά λέξη: "αρτή" προτεινόμενες διορθωμένες είναι "αστή", "αυτή", "αορτή".*
  - ***β.*** *Για την εσφαλμένη λέξη "πόος" προτείνονται οι λέξεις"πόνος", "πόρος" και "πόθος", "πόλος".*
- Συνεργασία με **μορφολογικό επεξεργαστή**, οπότε το λεξικό θα περιέχει μόνο λήμματα.

Σχόλιο - ερωτήματα

- Ποια σημεία αδυναμίας του συστήματός σας διαπιστώνετε;
- Μπορείτε να καταγράψετε και άλλα πιθανά σφάλματα; Ποιες θα μπορούσαν τότε να είναι οι θεμιτές ορθογραφικές διορθώσεις / υποδείξεις προς τον χρήστη;

Συστήματα κ' Τεχνολογίες Γνώσης – Εργασίες στην Επεξεργασία Φυσικής Γλώσσας

2. Διορθωτής Εκφράσεων

Αντικείμενο – Στόχος

Σκοπός της άσκησης είναι ο σχεδιασμός και η υλοποίηση συστήματος **διορθωτή εκφράσεων** βασισμένου στην **prolog** το οποίο:

- διαβάζει κείμενο από το πληκτρολόγιο ή από αρχείο
- ελέγχει την ορθότητα κάθε έκφρασης, και
- σε περίπτωση συντακτικού λάθους τυπώνει κατάλληλο μήνυμα και προτείνει τη διορθωμένη έκφραση (αν αυτό είναι εφικτό).

Τα λάθη

Τα λάθη που θα μπορεί να εντοπίζει το σύστημα, σύμφωνα με τη γραμματική που θα του δοθεί, μπορεί να είναι τα εξής:

- **α.** μια λέξη ανήκει σε άλλη γραμματική κατηγορία. Π.χ.:
  - *"Η γάτα ποντίκι"*
    - *Εδώ το σύστημα θα εντοπίζει το λάθος και θα προτείνει κάποιο ρήμα στη θέση της λέξης "ποντίκι" (π.χ. "νιαουρίζει"), με βάση τους συντακτικούς κανόνες :*

      S --> np vp, np - > d n, vp -> v.
- **β.** παράλειψη μιας λέξης. *Π.χ.*
  - *"Η γάτα το ποντίκι"*
    - *Εδώ το σύστημα θα εντοπίζει το λάθος και θα προτείνει κάποιο ρήμα (π.χ. "κυνηγά") με βάση τους συντακτικούς κανόνες:*

      S --> np vp, np - > d n, vp -> v np.
  - *"γάτα κυνηγά το ποντίκι"*
    - *Εδώ το σύστημα θα εντοπίζει το λάθος και θα προτείνει ένα άρθρο (π.χ. "η") με βάση τους (πιο πάνω) συντακτικούς κανόνες.*

Συντακτικοί Κανόνες

Θα πρέπει να περιέχονται τουλάχιστον οι κανόνες:

```
S --> np vp,
np - > d n,
vp -> v np.
vp -> v.
```

Συμβάσεις – Υποδείξεις

- Δώστε ένα λεξικό που να περιέχει τουλάχιστον τις λέξεις
  - ο, η, το, τα
  - ποντίκι, ποντίκια, γάτα, γάτες, σκύλος, σκύλοι, Κώστας, Μαρία
  - κυνηγά, κυνηγούν, νιαουρίζει, νιαουρίζουν

- Οι κανόνες θα πρέπει να εμπλουτιστούν με τα κατάλληλα χαρακτηριστικά ώστε να είναι ορθή η σύνταξη (με έλεγχο της συμφωνίας).
- Οι εκφράσεις που περιέχουν λέξεις εκτός λεξικού αφήνονται ως έχουν.
- Σε περίπτωση αμφισημίας (όπου μια έκφραση επιδέχεται διόρθωση με δύο ή περισσότερες εκδοχές) η διόρθωση γίνεται με την πρώτη εκδοχή που θα συναντήσει το πρόγραμμά σας.

Φορμαλισμός – Υπολογιστικό περιβάλλον

Χρησιμοποιείστε φορμαλισμό **Ενοποιητικής Γραμματικής** της **Prolog** .
(Δείτε το GULP 3.1 στο «εκπαιδευτικό υλικό» στην ιστοσελίδα http://glotta.ntua.gr/courses/NLP-Seminar/index.html*: Επέκταση της Prolog για Ενοποιητικές Γραμματικές (Unification-Based Grammar)*

Επέκταση του συστήματος

- Σε περίπτωση **αμφισημίας** (όπου μια έκφραση επιδέχεται διόρθωση με δύο ή περισσότερες εκδοχές) θα προτείνονται όλες οι εκδοχές, με κάποιο φιλικό τρόπο.

Σχόλιο - ερωτήματα

- Ποια σημεία αδυναμίας του συστήματός σας διαπιστώνετε;
- Μπορείτε να καταγράψετε και άλλα πιθανά σφάλαματα εκφράσεων; Ποιες θα μπορούσαν τότε να είναι οι θεμιτές διορθώσεις / υποδείξεις προς τον χρήστη;

Συστήματα κ' Τεχνολογίες Γνώσης –
Εργασίες στην Επεξεργασία Φυσικής Γλώσσας

3. Απόδοση Σημασίας Εκφράσεων

Αντικείμενο – Στόχος

Σκοπός της άσκησης είναι να σχεδιαστεί Συντακτικός και Σημασιολογικός Αναλυτής ο οποίος να δέχεται στην είσοδο μία πρόταση και να δίνει ως αποτέλεσμα το συντακτικό της δέντρο προσθέτοντας τη σημασία της.

Οι εκφράσεις

Παραδείγματα εκφράσεων και αντίστοιχων σημασιών:

έκφραση	*υποκείμενο*	*αντικείμενο*	*χρόνος*
Ο κυβερνήτης παραιτήθηκε	ένα	0	παρελθόν
Οι δικαστές συνεδριάζουν αύριο	πολλά	0	μέλλον
Ο δικαστής συνεδριάζει	ένα	0	παρόν
Οι δικαστές καταδίκασαν τους κλέφτες	πολλά	πολλά	παρελθόν
Ο δικαστής θα δικάσει αύριο	ένα	0	μέλλον
Οι δικαστές θα δικάσουν τους κλέφτες	πολλοί	πολλά	μέλλον

Συμβάσεις – Υποδείξεις

- Δώστε ένα λεξικό που να περιέχει τουλάχιστον τις λέξεις
  - ο, οι, τον, τους
  - κυβερνήτης, κυβερνήτες, δικαστής, δικαστές, κλέφτης, κλέφτες
  - συνεδριάζει, συνεδριάζουν, παραιτείται, παραιτήθηκε, (ενν. θα) παραιτηθεί, (ενν. θα) παραιτηθούν, (ενν. θα) δικάσει, (ενν. θα) δικάσουν
  - αύριο, σήμερα, χτες
  - θα

- Οι κανόνες θα πρέπει να εμπλουτιστούν με τα κατάλληλα χαρακτηριστικά ώστε να είναι ορθή η σύνταξη (με έλεγχο της συμφωνίας).
  - Π.χ. δεν θα πρέπει να δέχεται τις εκφράσεις του τύπου
    - οι δικαστές θα δικάσουν χτες *
    - οι δικαστές δίκασαν αύριο *
  - Προσοχή στις νόμιμες προτάσεις του τύπου:
    - οι δικαστές δικάζουν αύριο

Φορμαλισμός – Υπολογιστικό περιβάλλον

Χρησιμοποιείστε φορμαλισμό **Ενοποιητικής Γραμματικής** της **Prolog** .
(Δείτε το GULP 3.1 στο «εκπαιδευτικό υλικό» στην ιστοσελίδα http://glotta.ntua.gr/courses/NLP-Seminar/index.html*: Επέκταση της Prolog για Ενοποιητικές Γραμματικές (Unification-Based Grammar)*

Επέκταση του συστήματος

Ποια επιπλέον σημασιολογικά χαρακτηριστικά προτείνετε ώστε να καταγράφεται η βεβαιότητα / αβεβαιότητα ότι η πράξη (ενέργεια / δράση) του ρήματος έχει επιτελεστεί, αν συμπεριλάβουμε προτάσεις του τύπου:

- Οι δικαστές θα πρέπει να δίκασαν χτες
- Οι δικαστές μάλλον δίκασαν χτες

Συστήματα κ' Τεχνολογίες Γνώσης –
Εργασίες στην Επεξεργασία Φυσικής Γλώσσας

4. *Συντακτικός και Σημασιολογικός Αναλυτής Φυσικής Γλώσσας - "η γλώσσα των νηπίων"*

Αντικείμενο – στόχος

Σκοπός της άσκησης είναι να σχεδιαστεί Συντακτικός και Σημασιολογικός Αναλυτής ο οποίος να δέχεται στην είσοδο μία πρόταση του νηπίου και να δίνει ως αποτέλεσμα το πλήρες συντακτικό της δέντρο (προσθέτοντας τις λέξεις που λείπουν) και τη σημασία ή τις σημασίες τους (αν υπάρχει αμφισημία).

Ο σχεδιασμός της **γραμματικής** θα πρέπει να καλύπτει την έλλειψη των ρημάτων (π.χ. θέλω, έκανα, κάνω), ως υπονοούμενων και να αποδίδει τις **σημασίες** των ελλειπτικών προτάσεων των νηπίων.

Λεξικό – Συντακτικό

Θεωρούμε το **λεξιλόγιο** (λεξικό) των νηπίων:
μαμά, μπαμπά, νινί, ντα, άτα, μαμ, τσίσα, κακά, νάνι

και το **συντακτικό** {με μορφή παραδειγματικών φράσεων, μαζί με τις σημασίες τους}:

- πρόταση: νινί άτα
  σημασία: { (ενν. εγώ θέλω να) πάω βόλτα} | {(ενν. εγώ) πήγα βόλτα} | {(ενν. εγώ) πάω βόλτα}
- πρόταση: τσίσα
  σημασία: { (ενν. εγώ θέλω να) κάνω τσίσα} | {(ενν. εγώ) έκανα τσίσα} | {(ενν. εγώ) κάνω τσίσα}
- πρόταση: μαμ
  σημασία: { (ενν. εγώ Θέλω να) φάω} | {(ενν. εγώ) έφαγα} | {(ενν. εγώ) τρώω}
- πρόταση: νάνι
  σημασία: { (ενν. εγώ θέλω να) κοιμηθώ} | {(ενν. εγώ) κοιμάμαι} | {(ενν. εγώ) κοιμήθηκα}
- πρόταση: μπαμπά ντα
  σημασία: {ο μπαμπάς δέρνει} | {ο μπαμπάς με έδειρε} | {ο μπαμπάς θα με δείρει}

Συμβάσεις – Υποδείξεις

Η παραδοχή που κάνουμε για τους συντακτικούς κανόνες και για τις αντίστοιχες σημασίες τους (με αμφισημία) είναι:

- Δεν χρησιμοποιούνται άρθρα (περισσότητα ... ελληνικής γλώσσας)
- Τα έμψυχα ουσιαστικά (μαμά, μπαμπά, νινί) τα οποία απαντώνται ως πρώτη λέξη της φράσης υποδηλώνουν το υποκείμενο. Αν δεν υπάρχει έμψυχο ουσιαστικό

(ελλειπτικός λόγος) υπονοείται ως υποκείμενο το βρέφος που μιλά. Π.χ. "νάνι" ==> (ενν. εγώ θέλω) νάνι ==> {εγώ θέλω να κοιμηθώ}.

- Τα ουσιαστικά τα οποία απαντώνται ως δεύτερη λέξη (ντα, άτα, τσίσα κ.λπ.) υποδηλώνουν ενέργεια, δράση, κατάσταση και θεωρούνται ότι αποδίδουν κάποιο ρήμα:
  Π.χ. "νάνι" ==> (ενν. εγώ θέλω) νάνι ==> {εγώ θέλω να κοιμηθώ}
- Ο χρόνος του ρήματος εισάγει σημασιολογική αμφισημία.
  Π.χ. "νάνι" ==> {(εγώ) κοιμάμαι} ή {(εγώ) κοιμήθηκα}, ή {(εγώ θέλω να) κοιμηθώ}

Επέκταση της Γραμματικής

Στις προτάσεις προσθέτουμε και ένα αντικείμενο, ως τρίτη λέξη. Π.χ.:

- πρόταση: μπαμπά ντα νινί
  σημασία: { ο μπαμπάς έδειρε/δέρνει/θα δείρει το νινί}

Με αυτή την προσθήκη (συνθετότερη σύνταξη) ο ελλειπτικός λόγος περιορίζεται.

Φορμαλισμός – Υπολογιστικό περιβάλλον

Χρησιμοποιείστε φορμαλισμό **Ενοποιητικής Γραμματικής** της **Prolog** .
(Δείτε το GULP 3.1 στο «εκπαιδευτικό υλικό» στην ιστοσελίδα http://glotta.ntua.gr/courses/NLP-Seminar/index.html*: Επέκταση της Prolog για Ενοποιητικές Γραμματικές (Unification-Based Grammar)*

Συστήματα κ' Τεχνολογίες Γνώσης –
Εργασίες στην Επεξεργασία Φυσικής Γλώσσας

5. Μηχανική Μετάφραση Αριθμητικών

Αντικείμενο - στόχος

Θέλουμε να μεταφράζουμε μηχανικά την ολογραφική μορφή των αριθμητικών από μια φυσική γλώσσα σε μια άλλη.

> *Περιορισμός περιβάλλοντος: Περιοριζόμαστε στα απόλυτα αριθμητικά ουδετέρου γένους (π.χ. «δέκα τρία») και δεν εξετάζουμε τα άλλα γένη (π.χ. «δέκα τρεις») ούτε τα τακτικά αριθμητικά (π.χ. «δέκατο τρίτο»).*

Παραδείγματα αριθμητικών σε τέσσερεις γλώσσες:

> Ελληνικά: εκατό, ενενήντα τρία, εξακόσια δώδεκα, …
> Αγγλικά: a hundred, ninety three, six hundred twelve, …
> Γαλλικά: cent, quatre vingt treize, six cents douze, …
> Γερμανικά: ein Hundert, drei und neunzig, sechs Hundert zwoelf, …

Γλώσσα

Επιλέξτε γλωσσικό ζευγάρι: Την Ελληνική γλώσσα και μία από τις γλώσσες Αγγλική, Γαλλική, Ιταλική ή Γερμανική.

Σύστημα - Συντακτικός Αναλυτής (parser) Φυσικής Γλώσσας

Σχεδιάστε ένα σύστημα «μηχανικού μεταφραστή» (φυσικής γλώσσας), ο οποίος να δέχεται ένα αριθμητικό, από μηδέν (0) μέχρι και εννιακόσια ενενήντα εννέα (999), σε μία γλώσσα (Α) και να επιστρέφει το αριθμητικό αυτό μαζί με τη μετάφρασή του σε μια άλλη γλώσσα (Β). Η υλοποίησή σας θα πρέπει να λειτουργεί και για την ανάστροφη μετάφραση (από τη γλώσσα Β στη γλώσσα Α).

Περιγράφουμε μια γλώσσα (απόλυτων) αριθμητικών, όπου 'νόμιμες προτάσεις' είναι τα αριθμητικά που αντιστοιχούν στους μονοψήφιους, διψήφιους ή τριψήφιους αριθμούς σε ολογραφική μορφή (ολογράφως). Οι νόμιμες προτάσεις συγκροτούνται από μία ή περισσότερες λέξεις της φυσικής γλώσσας.

> *Π.χ. (στην ελληνική γλώσσα): μηδέν, ένα, δύο, τρία, ... δέκα, έντεκα, δώδεκα, ... δεκαπέντε, ... είκοσι ένα, ... εκατό, ... εκατόν πέντε, ... διακόσια δώδεκα. ... κ.λπ. μέχρι εννιακόσια ενενήντα εννέα.*

Κάθε τέτοια 'νόμιμη πρόταση' της φυσικής γλώσσας ελέγχεται ως προς την ορθότητά της από το σύστημα και δημιουργείται μια δομή (ενοποιητικής γραμματικής) η οποία αντλεί τις ιδιότητες και τις τιμές κάθε 'λέξης' από το 'λεξικό'.

Γιαυτό, αρχικά, καταγράφουμε όλες τις ιδιομορφίες που παρουσιάζει το ζευγάρι

γλωσσών που επιλέξαμε και οργανώνουμε το ‘λεξικό’ κατάλληλα ώστε να επιλέγεται για κάθε νόμιμη δομή το κατάλληλο λεκτικό, μέσω ιδιοτήτων και τιμών (paths: attribute-value pairs) των ‘φραστικών συστατικών’ του, εν προκειμένω των λέξεων.

Φορμαλισμός – Υπολογιστικό περιβάλλον

Χρησιμοποιείστε φορμαλισμό **DCG** και την **Ενοποιητική Γραμματική** της **Prolog**.
(Δείτε το GULP 3.1 στο «εκπαιδευτικό υλικό» στην ιστοσελίδα http://glotta.ntua.gr/courses/NLP-Seminar/index.html*: Επέκταση της Prolog για Ενοποιητικές Γραμματικές (Unification-Based Grammar)*

Υπόδειξη: Ιδιαιτερότητες Αριθμητικών (παραδειγματικά, όχι εξαντλητικά)

Αμφισημία:

Το απόλυτο αριθμητικό που αντιστοιχεί στο 100 γράφεται αλλιώς («εκατό») αν δεν έχει παρακολούθημα και αλλιώς («εκατόν ...») αν ακολουθείται από άλλον αριθμό (π.χ. «εκατόν πέντε»).

Στις διαφορετικές γλώσσες μία λέξη μπορεί να μεταφράζεται σε μία, δύο ή τρεις λέξεις και αντιστρόφως:

treize, quatorze, quinze ... => δεκατρία, δεκατέσσερα, δεκαπέντε
nineteen ... => δέκα εννέα
εκατό => a hundred, ein Hundert
ογδόντα => quatre vingt
ενενήντα => quatre vingt dix
εξακόσια => six hundred