



ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

Κάτια Κερμανίδου

kerman@ionio.gr

ΜΟΡΦΟΛΟΓΙΑ

Μερικές εικόνες είναι δανεισμένες από τους Sun & Chen στο University of Cambridge

Μορφολογία

❖ Μελετά την εσωτερική δομή μιας λέξης

Μερικές καλασαβάνες θυμοκάλησαν επιπατωμικά μέσα στο σιωτοπό με τα βαλάματα.



Μερικ-ές καλασαβάν-ες θυμοκάλησ-αν επι-πατωμ-ικά μέσα στο σιωτοπ-ό με τα βαλάματ-α.

Μόρφημα (Morpheme): το μικρότερο συστατικό μιας λέξης που φέρει μορφολογική πληροφορία.

Οι λέξεις σχηματίζονται από μορφήματα.

- ❑ Πρόθεμα (prefix): *ξε-πλένω*
- ❑ Ένθεμα (infix): *δι-έ-σχισα*
- ❑ Κατάληξη ή επίθεμα (suffix): *κατεργαρ-άκος*

Ρίζα (Root): ο πυρήνας της λέξης στον οποίο επικολλούνται τα μορφήματα και μένει αναλλοίωτος κατά τις μορφολογικές κλίσεις της λέξης (-πλεν- , κατεργαρ-)

Λήμμα (Lemma): η μορφή της λέξης με την οποία την ψάχνω σε ένα λεξικό (ξεπλένω, κατεργάρης)

Μορφολογία: Κλίση και Παραγωγή

❖ Κλίση (Inflection)

Παράγει διαφορετικές μορφές της ίδιας λέξης, αλλάζει η γραμματική κατηγορία της λέξης

παίζω, παίζεις, παίζει, παίζουμε, έπαιξε...

❖ Παραγωγή (Derivation)

Παράγει διαφορετικές λέξεις, αλλάζει το νόημα της λέξης, πολλές φορές και το μέρος του λόγου της

πλένω, ξεπλένω, προπλένω

establish (V)

establish+ment (N)

establish+ment+ary (Adj)

establish+ment+ari+an (N)

establish+ment+ari+an+ism (N)

dis+establish+ment+ari+an+ism (N)

anti+dis+establish+ment+ari+an+ism (N)

Μορφολογία: Stemming

❖ Stem: αυτό που μένει αν από μια λέξη αφαιρέσουμε όλα τα κλιτικά μορφήματα (inflectional affixes - συνήθως επιθέματα), διατηρούμε όμως τα παραγωγικά.

- marsupials -> marsupial

- trying -> try

Στις περισσότερες εφαρμογές ΓΤ, ειδικά για τα Αγγλικά, είναι αρκετό ως μορφολογική επεξεργασία ενός κειμένου.

Το πιο γνωστό εργαλείο είναι ο stemmer του Porter (1980) <http://www.tartarus.org/~martin/PorterStemmer/>

❖ Βασίζεται σε έναν αριθμό απλών κανόνων

- ❖ ATIONAL -> ATE (relational -> relate)

- ❖ ING -> ε (knowing -> know)

❖ Οι stemmers δεν έχουν απόλυτη ακρίβεια

- ❖ organization -> organ policy -> police

- ❖ Περιπτώσεις που ο Porter stemmer δεν καλύπτει:

- ❖ matrices -> matrix explain -> explanation

Oxford Dictionary - Word of the Year

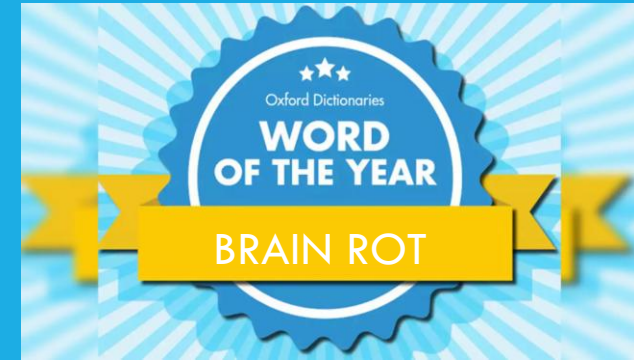
Emojis, Compounds και Multi-word Expressions



2015



2019



2024

Σύνθετες Λέξεις (Compounds): Σχηματίζονται από περισσότερες από μια ρίζες
Earthquake, πιασγύρισμα

Multi-word expressions: Συνδυασμοί δυο ή περισσότερων λέξεων που παρουσιάζουν ιδιαίτερη συντακτική και σημασιολογική συμπεριφορά

Fixed		(Syntactically) flexible
<i>by and large</i>		<i>put on the clothes</i> <i>put the clothes on</i>
Non-compositional	Semi-compositional	Compositional
<i>kick the bucket</i>	<i>spill the beans</i> (reveal the secret)	<i>strong tea</i>

Εργασίες στο Μορφολογικό Επίπεδο

Tokenization: Η Αναγνώριση των λεκτικών συμβόλων σε ένα κείμενο.

Λεκτικά Σύμβολα (Token): Λέξεις, αριθμοί, σύμβολα, σημεία στίξης, ακρωνύμια, συντομογραφίες, Urls, emojis, ξενικές λέξεις, ...

Όχι προφανής διαδικασία (πχ ό,τι χωρίζεται με το κενό)

To *New York* είναι ένα token; Το *we're*;

Truecasing: μετατροπή όλων των χαρακτήρων σε πεζά

Μπορεί να χρησιμοποιηθεί στην Ανάκτηση Πληροφορίας, αν οι χρήστες χρησιμοποιούν πεζά ή στην Μηχ. Μετάφραση. Όχι σε εργασίες που τα κεφαλαία/πεζά παίζουν ρόλο (Αναγνώριση ονομάτων)

Κανονικοποίηση (Standardization): όλες οι λέξεις σε συνεπες φορματ

(U.S.A., USA → USA uhuh, uh-huh-→ uhuh 4, for → for)

Stemming

Λημματοποίηση : η εύρεση του λήμματος μιας λέξης (saw → see)

Εργασίες στο Μορφολογικό Επίπεδο

Συνέχεια

Αναγνώριση μερών του λόγου: Απόδοση μορφολογικής ετικέτας στη λέξη βάσει των συμφραζομένων της (saw → see, ΡΗΜΑ, ΑΟΡΙΣΤΟΣ)

Σύνθεση (Generation): Δημιουργία λέξης βάσει του λήμματος και της μορφολογικής πληροφορίας που θέλω να έχει (see, ΡΗΜΑ, ΑΟΡΙΣΤΟΣ → saw)