Parts of Speech

חלקי-הדיבור

- מקובל למנות 9~ קבוצות מילים המכונות "חלקי- דיבר":
- שם עצם (noun),שם תואר (adjective),
כינוי (pronoun),שם מספר (numeral),פועל (verb),
תואר הפועל (adverb),מלת יחס (preposition),
מלת חיבור (conjunction),מלת קריאה (interjection).
- אך זו רק חלוקה אחת

למה זה טוב?

- בסיס לניתוח parsing
- מציאת ביטויים term identification/chunking
- יצירת קול (TTS) – אופן הביטוי של המילה:
  - רכבת/רכבת
  - CONtent/conTENT, OBJect/objECT, DIScount/disCOUNT
- בעברית – ניקוד אוטומטי
- רכיב בתרגום
- רכיב בחיפוש: book about flight book a flight

איך מגדירים חלקי דיבר?

- באופן מסורתי, ההגדרה של חלקי הדיבר מבוססת על תכונות מורפולוגיות של המילה או על המילים שמופיעות לידן בסמיכות distributional properties.
- באופן עקרוני, יש למילים מאותו חלק דיבר דמיון סמנטי, כלומר, הן מתארות איברים מאותן קבוצות למשל
  - שמות עצם –nouns אנשים, מקומות, דברים – thought, table, sister
  - שמות תואר – adjectives תכונות, כמויות big, lazy
  - לואי פעולה – adverbs – מתארים אופן, מקום, זמן, איכות quickly
  - פעלים – אירועים, התרחשויות או מצבי קיום – eat, is, write
  - ויש גם מילות יחס, מילות איחוי ועוד...

The yinkish dripner blorked quastofically into the nindin with the pidibs.

The yinkish dripner blorked quastofically into the nindin with the pidibs.

- yinkish -adj
- dripner -noun
- blorked -verb
- nindin -noun
- pidibs -noun
- quastofically -adverb

The yinkish dripner blorked quastofically into the nindin with the pidibs.

- yinkish -adj nindin -noun
- dripner -noun pidibs -noun
- blorked -verb quastofically -adverb

- We determine the P.O.S of a word by the affixes that are attached to it and by the syntactic context (where in the sentence) it appears in.

Open class vs. Closed class types

- Closed class:
  - small group, does not (usually) grow
    - “function words”, determiners, prepositions, pronouns,...
- Open class:
  - large group, and grows larger
    - verbs, nouns, adjectives
    - productive group: “to google”, “to fax”, “googling”

שמות עצם

- Nouns
  - take *-s, 's, -ness, -ment, -er*, affixes
  - Occur with determiners (*a,the,this,some…*)
  - can be a subject of a sentence.
- Semantically: can be concrete – *chair, train*, or abstract – *relationship*.
- או שמות פעולה, למשל: אכילה, לאכול, *eating*

Types of Nouns

- Proper Nouns:
  - David, Israel, Microsoft
  - Aren't preceded by articles
  - Capitalized (In English)
- Common Nouns:
  - Count Nouns:
    - allow grammatical enumeration (book, books)
    - can be counted (one apple, 50 thoughts)
  - Mass Nouns: snow, salt, communism, …

Verbs

- מילים המתייחסות לפעולות או תהליכים

Main verbs – *draw, provide, differ* –

Auxiliaries (usually considered closed class) –

- מערכת הטיה מורפולוגית

eat, eats, eating, eaten –

Adjectives

- מבחינה סמנטית, קבוצה הכוללת ביטויים המתארים תכונות או איכויות, משהו כמו פרדיקט חד-מקומי.
- שפות רבות כוללות:
  - צבעים (yellow, green)
  - גילאים (young, old)
  - וערכים. (good, bad)
- יש שפות בלי שמות תואר.

Adverbs

- *קבוצה מעורבת למדי...*

- *Unfortunately,* John walked *home extremely slowly yesterday*
- Directional: sideways, downhill
- Locative: home, here
- Degree: extremely, somewhat
- Manner: slowly, delicately
- Temporal: yesterday, Monday

Closed class

- Prepositions – on, under, over, near, by, at, from, to, with
- Determiners – a, an, the
- Pronouns – it, she I
- Conjunctions – and, but, or, as, if, when
- Auxiliary verbs – can, may, should, are
- Particles – up, down, on, off, in , at, by
- Numerals – one, two , second, third

Prepositions and particles

- Prepositions
  - ...on top, by then, with him
  - מילות יחס המופיעות לפני שם עצם
  - מצינות יחסי זמן/מקום, אבל לא רק.
- Particles
  - go **on**, look **up**, turn **down**
  - מופיעים אחרי פועל, ובפעלים טרנזיטיביים, גם אחרי המושא
    - *The horse went off its truck/throw off sleep*
    - *The horse went its track off/throw sleep off**

Articles (determiners)

- a, an, the
- noun phrase מופיעים בתחילה צירוף שמני
- גם: **this** chapter, **that** page
- שכיחים מאוד בטקסטים

Conjunctions

- מאחים שני phrases , צירופים , משפטים, וכו.
- Or, and, but מאחים צירופים מאותו סטטוס
- Subordinating conjunctionsמשמשים לאיחוי צירופים מקוננים

- *I thought that you might like some milk.*
  - *I thought – main clause*
  - *That you might… - subordinating clause*

ויש עוד...

Tagsets

Tagset

The set of possible tags for parts of speech. (size is changing in applications, languages...)

A tagset should include the information that is needed for the next steps in the process, and that people can annotate well

Brown corpus – 87 tags

Penn Treebank – 45 tags

Large: 146-tag C7 tagset of used to tag the British National Corpus BNC.

“Universal” - 12 tags

Tagsets

Tagset

The set of possible tags for parts of speech. (size is changing in applications, languages...)

A tagset should include the information that is needed for the next steps in the process, and that people can annotate well

Brown corpus – 87 tags

Penn Treebank – 45 tags

Large: 146-tag C7 tagset of used to tag the British National Corpus BNC.

“Universal” - ~12 tags

Penn Tagset

- **Noun** (person, place or thing)
  - Singular (**NN**): dog, fork
  - Plural (**NNS**): dogs, forks
  - Proper (**NNP**, **NNPS**): John, Springfields
  - Personal pronoun (**PRP**): I, you, he, she, it
  - Wh-pronoun (**WP**): who, what
- **Verb** (actions and processes)
  - Base, infinitive (**VB**): eat
  - Past tense (**VBD**): ate
  - Gerund (**VBG**): eating
  - Past participle (**VBN**): eaten
  - Non 3rd person singular present tense (**VBP**): eat
  - 3rd person singular present tense: (**VBZ**): eats
  - Modal (**MD**): should, can
  - To (**TO**): to (to eat)

Penn Tagset (cont.)

- **Adjective** (modify nouns)
  - Basic (**JJ**): red, tall
  - Comparative (**JJR**): redder, taller
  - Superlative (**JJS**): reddest, tallest
- **Adverb** (modify verbs)
  - Basic (**RB**): quickly
  - Comparative (**RBR**): quicker
  - Superlative (**RBS**): quickest
- **Preposition (IN)**: on, in, by, to, with
- **Determiner**:
  - Basic (**DT**) a, an, the
  - WH-determiner (WDT): which, that
- **Coordinating** Conjunction (**CC**): and, but, or,
- **Particle** (**RP**): off (took off), up (put up)

Universal tagset

- Can describe over 22 languages with the same set of tags.
  - Why do we want to do that?
    - Language transfer, ease of use for developers
- The tags:
  - Noun, Verb, Adv, Adj, Pron, Det, Adp, Num, Conj, Prt, Punc, X

Universal tagset

- Can describe over 22 languages with the same set of tags.
  - Why do we want to do that?
    - Language transfer, ease of use for developers
- The tags:
  - Noun, Verb, Adv, Adj, Pron, Det, **Adp**, Num, Conj, Prt, Punc, X

preposition or postposition

Universal tagset

- Can describe over 22 languages with the same set of tags.
  - Why do we want to do that?
    - Language transfer, ease of use for developers
- The tags:
  - Noun, Verb, Adv, Adj, Pron, Det, Adp, Num, Conj, Prt, Punc, **X**

"other"

Part-Of-Speech Tagging

- תיוג הוא התהליך של השמת חלקי דיבר או סימון לקסיקלי אחר לכל מילה בקורפוס. (tokenization)
- תיוג מתבצע בדרך כלל גם על סימני פיסוק
- הקלט הוא רצף מילים ו-tagset מהסוג שראינו.
- הפלט הוא התיוג הטוב ביותר עבור כל אחת מן המילים.
- והבעייה המרכזית, היא – ambiguity:
  - אשה נעלה נעלה נעלה נעלה את הדלת בפני בעלה

“Around” can be a preposition, particle, or adverb

I bought it at the shop around/IN the corner.

I never got around/RP to getting a car.

A new Prius costs around/RB $25K.

“like” can be a verb or a preposition:

time flies like an arrow

fruit flies like a banana

(“flies” can be a verb or a noun...)

The Distribution of Tags

- Tags follow all the usual frequency-based distributional behavior.
- Most word types have only one part of speech.
- Of the rest, most have two. Things go pretty much as we'd expect from there on.
- Of course, as usual, the most frequently occurring word types tend to have multiple tags.
- (As we'll see later in the semester, they also tend to have more meanings).
- Therefore while its easy to determine the correct tag for most word types, it isn't necessarily so easy to tag most texts.

Word Types in the Brown Corpus

Unambiguous : 1 tag	35340
Ambiguous: > 1 tag	4100
2 tags	3760
3 tags	264
4 tags	61
5 tags	12
6 tags	2
7 tags	1 (“still”)

State of the Art

- A dumb tagger that simply assigns the most common tag to each word achieves ~90%
- Best approaches give ~96/97%
- This still means that there will be on average one tagging error per sentence
- Life is much more difficult if we do not have a lexicon and/or training corpus or if we use a tagger across domains and genres.