How Many Different Words?

10,000 sentences from the Europarl corpus

<table>
<thead>
<tr>
<th>Language</th>
<th>Different words</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>16k</td>
</tr>
<tr>
<td>French</td>
<td>22k</td>
</tr>
<tr>
<td>Dutch</td>
<td>24k</td>
</tr>
<tr>
<td>Italian</td>
<td>25k</td>
</tr>
<tr>
<td>Portuguese</td>
<td>26k</td>
</tr>
<tr>
<td>Spanish</td>
<td>26k</td>
</tr>
<tr>
<td>Danish</td>
<td>29k</td>
</tr>
<tr>
<td>Swedish</td>
<td>30k</td>
</tr>
<tr>
<td>German</td>
<td>32k</td>
</tr>
<tr>
<td>Greek</td>
<td>33k</td>
</tr>
<tr>
<td>Finnish</td>
<td>55k</td>
</tr>
</tbody>
</table>

Why the difference? Morphology.

Interlude: types and tokens

The word word is ambiguous.

- **Word type:** “10k sentences from English Europarl have 16k different words” (unique strings, lexical items)
- **Word token:** “English Europarl has 54m words” (possibly repeated instances)

A cat and a brown dog chased a black dog: 10 tokens, 7 types.
What is morphology?
The study of wordforms and word formation.

- Structured relationships between words:
  - play, played, replay, player
  - played, walked, jumped

- Units of meaning (*morphemes*) and their ordering (*morphotactics*):
  - de+salin+ate+ion but not ate+salin+ion+de

Why does morphology matter?

- Information retrieval: return pages with related forms.
- Language modelling: make predictions about unseen words
- Machine translation and language understanding: signals differences in meaning (might be expressed using word order in other languages).

Example (Russian):

zhenshina devochka dala knigu
woman+NOM girl+DAT gave book+ACC
‘the woman gave the girl a book’

vs.

zhenshine devochka dala knigu
woman+DAT girl+NOM gave book+ACC
‘the girl gave the woman a book’

Morphemes: Stems and Affixes

- Two types of morphemes
  - stems: small, cat, walk
  - affixes: +ed, un+
- Four types of affixes
  - suffix
  - prefix
  - infix
  - circumfix
Stems vs. Lemmas

- Lemma: the canonical form or dictionary form of a set of words
  - fly, flies, flew and flying all have the lemma fly.
  - walk, walks, walked and walking all have the lemma walk.
  - walker, walkers have the lemma walker.

- Stem: definitions can vary, but often: the part of the word that is common to all its variants
  - stem of produce, production is produc.
  - stem of walk, walks, walked, walking, walker, walkers is walk.
  - Do fly, flies, flew, flying have a common stem fl?
    Or maybe only fly and flying share a stem: fly.
    Decision may depend on application.

Suffix

- Plural of nouns
  cat+s
- Comparative and superlative of adjectives
  small+er
- Formation of adverbs
  great+ly
- Verb tenses
  walk+ed
- All inflectional morphology in English uses suffixes

Prefix

- In English: these typically change the meaning
- Adjectives
  un+friendly
  dis+interested
- Verbs
  re+consider
- Some language use prefixing much more widely
Infix

- In English: inserting profanity for emphasis
  - abso+bloody+lutely
  - unbe+fucking+lievable
- Why not:
  - ab+bloody+solutely

Circumfix

- No example in English
- German past participle of verb:
  - ge+sag+t (German)

Not that easy...

- Affixes are not always simply attached
- In writing, some letters may be changed/added/removed
  - walk+ed
  - frame+d
  - emit+ted
  - carr(–y)+ied
- In speaking, some sounds may be changed/added/removed
  - Compare the final sound: cats [s] vs dogs [z] vs foxes [ɔz]

Irregular Forms

- Some words have irregular forms:
  - is, was, been
  - eat, ate, eaten
  - go, went, gone
- Irregular forms tend to be the most frequent (and vice versa)
Inflectional Morphology

• In English, we inflect
  – nouns for count (plural: +s) and for possessive case (+'s)
  – verbs for tense (+ed, +ing) and a special 3rd person singular present form (+s)
  – adjectives in comparative (+er) and superlative (+est) forms.

• In German, we inflect
  – nouns for count and case
  – verbs for tense, person, and count
  – adjectives for count, case, gender, and definiteness
  – determiners for count, case and gender

Forms of the German the

<table>
<thead>
<tr>
<th>Case</th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>male</td>
<td>fem.</td>
</tr>
<tr>
<td>nominative (subject)</td>
<td>der</td>
<td>die</td>
</tr>
<tr>
<td>genitive (possessive)</td>
<td>des</td>
<td>der</td>
</tr>
<tr>
<td>dative (indirect object)</td>
<td>dem</td>
<td>der</td>
</tr>
<tr>
<td>accusative (direct object)</td>
<td>den</td>
<td>die</td>
</tr>
</tbody>
</table>

Not only many different forms, but each form is highly ambiguous

Inflectional vs. Derivational Morphology

• Inflectional morphology typically
  – does not change basic meaning or part of speech
  – expresses grammatical features or relations between words
  – applies to all words of the same part of speech

• Derivational morphology
  – may change the part of speech or meaning of a word
  – is not driven by syntactic relations outside the word
  – may be “picky”: drama+(t)ize but not traged(-y)+ize
  – applies closer to the stem; whereas inflection occurs at word edges: govern+ment+s, centr+al+ize+d

Derivational Morphology

• Changing the part of speech, e.g. noun to verb
  
  word → wordify

• Is it a real word?

• Consulting Google (a few years ago):
  – 8,840 hits: e.g., wordify mugs, tshirts and magnets

• Google now returns 25k hits. (Why?)
Derivational Morphology

- Changing the verb back to a noun
  
  \[
  \text{wordify} \rightarrow \text{wordification} \quad (2k \text{ hits on Google})
  \]

- A person/thing who engages in \textit{wordification}
  
  \[
  \text{wordification} \rightarrow \text{wordificator} \quad (\text{was } 8 \text{ hits, now } 3k: \text{ another app!})
  \]

- A person/thing who \text{wordifies}
  
  \[
  \text{wordify} \rightarrow \text{wordifier} \quad (700 \text{ hits on Google})
  \]

- What is the difference between a \textit{wordifier} and a \textit{wordificator}?

Compounds

- Creating new words by merging multiple words
  
  \[
  \text{home work} \rightarrow \text{homework}
  \]
  
  \[
  \text{web site} \rightarrow \text{website}
  \]

- (Somewhat) rare in English

- More common in other languages (like German)


\[
\text{I think you're confusing the term “Democracy” with “Capitalism”; I think you mean “Has Capitalism failed”?}
\]

\[
\text{No. It hasn’t.}
\]

\[
\text{I agree, Hambone; I’m just trying to correct the wordificationism.}
\]

\[
\text{Where in the world did you get the word “wordificationism”? Not in the Merriam-Webster dictionary, not in the Thesaurus...}
\]
Acronyms/Initialisms

• Wikileaks / Guardian, document 2007-081-100110-0444:
  OGA operating in TF Catamount sector moved into Malekshay for operation. LN Shum Khan ran at the sight of the approaching CFA’s. CF utilized the escalation of force doctrine and shouted to stop, fired warning shots and then fired to wound. The LN was hit in the ankle and treated by Element medics on scene. It was determined through discussions with local Elders that the man was a deaf mute that was nervous of the CF operation. Solatia was made in the form of supplies and the Element mission progressed.

Morphology differs across languages

• Usually a trade-off between morphology and syntax
  – Some languages have no verb tenses
    → use explicit time references (yesterday)
  – Case inflection determines roles of noun phrase
    → use fixed word order instead
    → use prepositional phrases instead of cased noun phrases
  – Examples from the World Atlas of Language Structures (wals.info)
    – prefixes vs. suffixes
    – cases (zero to more than ten)
    – past tense remoteness distinctions
So...

How to deal with all this computationally?
What do we even want to be able to do?

Tasks

- Recognition
  - given: wordform (string of characters)
  - wanted: yes/no decision if it is in the language

- Generation
  - given: lemma and morphological properties
  - wanted: correctly inflected wordform

- Analysis
  - given: wordform
  - wanted: lemma and morphological properties

Word Lists

- Simple Solution
  - create a list of all wordforms and their morphological properties
  - solve tasks by checking against list

- But...
  - list can become very long
  - list fails to generalize for productive morphology

- Instead: use finite state machines
  (also called finite state automatons)
Finite State Machines: States

places we may find ourselves in

Finite State Machines: Transitions

moving between the states

Finite State Machines: Emissions

emissions: letters produced at each transition

Finite State Machines: Start and End

begin at start state, finish at end state
The language of an FSM

Every FSM defines a formal language:

- The set of strings that can be generated by moving from start to end states, emitting symbols on each transition.
- Equivalently, the set of strings that can be recognized by matching input characters to emission symbols.

The language of an FSM may be finite or infinite.

Regular Languages

- Languages produced by FSMs are called regular languages
- Many convenient properties (e.g., straightforward to determine if a word is in the language)
- Not all languages are regular
  example: \( a^n b^n = \{ ab, aabb, aaabbb, aaaaabbb, \ldots \} \)
  (would require an infinite number of states)
Regular Expressions

- Reg. languages can also be described with regular expressions.

- Every RegEx is equivalent to some FSM (and vice versa).
  Example: \( ac(ac|bc) \mid aa(cc|bb) \mid bb^*a(bb|cc) \)
  where ‘|’ means “or” and ‘\( x^* \)’ means “zero or more \( x \)’s”.

- RegExs are common in programming to describe sets of strings.
  - \( \text{ls} \ast \).jpg
  - if ($\text{word} =~ ^/[A-Z].*/$) { $\text{name} = 1; }
  - if ($\text{name} =~ /[WB]ill$/) { print "Will or Bill"; }

Chomsky Hierarchy

- Chomsky discussed four major classes of formal languages
  3. regular (generated by finite state machines, usually assumed sufficient to describe phonology and morphology)
  2. context-free (will be covered in later lectures on syntax)
  1. context-sensitive (possibly needed for some natural language phenomena)
  0. recursively enumerable (anything a computer program can produce)

- (There are also many classes of “sub-regular” languages.)

Summary

- Morphology: word formation processes (e.g., derivation/inflection).
- Varies a lot between languages.
- We want to be able to analyze substructure of words.
- To do so, we will use FSMs (more examples next class).
- FSMs: a way to describe regular languages.
- Reg. Langs also described by regular expressions.
Reminders

1. Labs on Wed/Thu:
   - Get and test your DICE login before then, and go through the DICE introduction.
   - Check the course web page to see which lab to go to.

2. Also, start going through probability tutorial (see Readings on web page).

3. Join the course discussion forum (Piazza)! Link is now on the course web page.