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: Stemming, to return pages with related forms.
- Language modelling: Can make predictions about unseen words

zhenshina devochke dala knigu
woman+NOM girl+DAT gave book+ACC
'the woman gave the girl a book'

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

Stem vs. Root vs. Lemma

- Stem
  - the part of the word that is common to all its inflected variants
  - stem of produce and production is produc
- Root
  - primary lexical unit of a word, cannot be reduced into smaller constituents
  - the root of interrupt is rupt (no meaning by itself; but related to rupture)
- Lemma
  - the canonical form, dictionary form, or citation form of a set of words
  - fly, flies, flew and flying are forms of same lexeme, with fly as lemma

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: meaning changing particles
- 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
- Some consonants of the lemma may be changed or removed
  - walk+ed
  - frame+d
  - emit+ted
  - carr(-y)+ied
- Can be just spelling, or for phonetic reasons (dogs vs. foxes)
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)

- Probably a cognitive explanation: regular morphology allows us to express new meanings without memorizing new words. (But then: why irregulars?)

Inflectional Morphology

- In English
  - Nouns are inflected for count (plural: +s) and for possessive case (+’s)
  - Verbs are inflected for tense (+ed, +ing) and a special 3rd person singular present form (+s)
  - Adjective are inflected in comparative (+er) and superlative (+est)
  - Determiners are not inflected

- In German
  - Nouns are inflected for count and case
  - Verbs are inflected for tense, person, and count
  - Adjectives are inflected for count, gender, and definiteness
  - Determiners are inflected 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>dem</td>
<td>die</td>
</tr>
</tbody>
</table>

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

Morphology across languages

- Alternatives
  - Some languages have no verb tenses
    → use explicit time references (yesterday)
  - Case inflection determines roles of noun phrase
    → use fixed word order instead
  - Cased noun phrases often play the same role as prepositional phrases

- Trade-off between morphology and syntax, but always some redundancy
**Inflectional vs. Derivational Morphology**

- **Derivational morphology**
  - change part of speech or meaning of a word
  - not driven by syntactic relations outside the word

- **Inflectional morphology**
  - does not change basic meaning or part of speech
  - expresses grammatic features or indicates relations between different words
  - applies to all words of the same part of speech

- Inflectional morphology occurs outside derivational morphology

---

**Derivational Morphology**

- Changing the verb back to a noun
  
  wordify → wordification (2,350 hits on Google)

- A person who engages in wordification
  
  wordification → wordificator (8 hits on Google)

- A person who wordifies
  
  wordify → wordifier (2,820 hits on Google)

- What is the difference between a wordifier and a wordificator?

---

**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 266k hits. (Why?)

---

**Derivational Morphology**

- Turning wordification into a ideology:
  
  wordification → wordificationism

- 1 hit on Google

  I think you’re confusing the term ”Democracy” with ”Capitalism”; I think you mean ”Has Capitalism failed”? No. It hasn’t.
  
  I agree, Hambone; I’m just trying to correct the wordificationism. Where in the world did you get the word ”wordificationism”? Not in the Merriam-Webster dictionary, not in the Thesaurus...
**Derivational Morphology**

- A adherent of wordificationism

  \[\text{wordificationism} \rightarrow \text{wordificationist}\]

- 0 hits on Google (except now these slides!)

- We created a new word!

**Compounds**

- Creating new words by merging multiple words

- Rare in English

  \[\text{home work} \rightarrow \text{homework}\]

  \[\text{web site} \rightarrow \text{website}\]

- More common in other languages

**Acronyms**

- Guardian:

  David Cameron plans to save millions by cutting quangos

- What is a quango?

- An Australian animal?

- No:

  quasi non-governmental organization

- Another example: 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
Different Languages

- Languages differ a lot in morphology

- Examples from The World Atlas of Language Structures Online (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: surface form
  - wanted: yes/no decision if it is in the language

- Generation
  - given: lemma and morphological properties
  - wanted: surface form

- Analysis
  - given: surface form
  - wanted: lemma and morphological properties

Word Lists

- Simple Solution
  - create a list of all surface forms 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

Generated language: \{acac, acbc, aacc, aabb, bacc, babb\}
Finite State Machines: Loops

START
a
b

b

a

b

START END

a

b

a

b

b

c

a

b

b

END

loop → infinite language: \{ babb, bbabb, bbbabb, bbbbbabb, ... \}

Regular Languages

• Finite state machines produce regular languages
• Easy to deal with:
  determining if a word can be generated by the language straight-forward
• Not all languages can be generated by a finite state machine
  example: \(a^nb^n = \{ ab, aabb, aaabbb, aaaaabbb, ... \}\)
  (would require an infinite number of states)

Regular Expressions

• Common feature of programming languages

• Examples
  – ls *.jpg
  – if \($word = /[^\[A-Z].*/\) \{ $name = 1; \}"
  – if \($name = /[^\[WB\]ill/\) \{ print "Will or Bill" \}

• Regular expressions have same power as finite state machines
  → define regular languages
  → can be encoded as finite state machines

Chomsky Hierarchy

• Four major types of formal languages

3. regular (generated by finite state machines)
2. context-free (will be covered in later lectures on syntax)
1. context-sensitive
0. recursive enumerable (anything a computer program can produce)

• Languages higher up in the Chomsky hierarchy increasingly complex
  – can describe more languages
  – harder to compute
    for instance: for type-0 it is not generally possible to determine if a specified
    word can be generated by the language
Reminder

Lab on Thursday: make sure you get and test your DICE login before then and go through the DICE introduction.