Topic: Introduction in Spoken Language

Part II

Part I
- Spoken Language Systems
- Spoken Language Structure
- Speech Perception
- Speech Production

Part II
- Phonetics and Phonology
- Syllables and Words
- Syntax and Semantics
- Transcription of Spoken Speech

Reference: Huang et al. Chapter 1 and 2; Waibel/Lee Chapter 9

April 25, 2013
Overview (I)

- Phonetics and Phonology
  - Phonetics vs. Phonology
  - Phone vs. Phoneme
    - Analogy to text coding
  - Phonetic Alphabets
    - IPA Scheme for Consonants
    - IPA Scheme for Vowels
- Vowels
  - The Vowel-Quadrangle
  - Different Shapes of the Vocal Tract
  - Formants (F1 and F2)
  - Diphthongs
- Consonants
  - Classification of Consonants
  - Vocal Tract Shapes of Plosives / Nasals / Fricatives & Spectrograms
  - English Consonants
Overview (II)

• Phonetics and Phonology
  – Sound Inventories among Languages
  – Sound Unit Sharing across Languages
  – Representations of Phonemes
    • Allophones
    • Coarticulation
    • Phones and Context
    • Prosody
    • Spontaneously spoken speech

• Syllables and Words
  – Syllables
  – Words
Overview (III)

• Syntax and Semantics
  – Syntax
  – Semantics
  – Language & Compactness

• Transcription of Spoken Speech
  – Transcription vs. Transliteration
Overview

- Phonetics and Phonology
  - Phonetics vs. Phonology
  - Phone vs. Phoneme
    - Analogy to text coding
  - Phonetic Alphabets
    - IPA Scheme for Consonants
    - IPA Scheme for Vowels
  - Vowels
    - The Vowel-Quadrangle
    - Different Shapes of the Vocal Tract
    - Formants (F1 and F2)
    - Diphthongs
  - Consonants
    - Classification of Consonants
    - Vocal Tract Shapes of Plosives / Nasals / Fricatives & Spectrograms
    - English Consonants
Phonetics vs. Phonology

- **Phonetics**: Study of the *production, classification* and *transcription* of speech sounds
  - Focus is on the unique (since dependent from speaker, ...) acoustic realization of speech sounds

- **Phonology**: Study of the *distribution* and *patterning* of speech sounds in a language, pronunciation
  - Focus is on finding gross characteristics of speech sounds that are adequate for description and classification of words (in a dictionary)
Phone vs. Phoneme

- No relation between object and its sign, no relation between pronunciation and meaning
- Phonetics: *Speech sounds do not have an inherent meaning*
- But: We need to have a characteristics of speech sounds to describe and classify words and their pronunciation
- 2 basic concepts:
  - **Phoneme**: a phoneme is the smallest speech unit which differentiates the meaning of a word pair (*minimal pair*)
    Examples:
    /bat/ vs. /pat/ $\rightarrow$ /b/ and /p/ are phonemes (linguistically distinct sounds)
    /l/ and /r/ are phonemes in English, but not in Japanese
  - **Phone**: a phoneme might have different *acoustic realizations* according to context, speaker, language, ...
- Writing convention
  - /phoneme/ vs. [phone]
Phone vs. Phoneme – Analogy to text coding

Analogy to the coding of text characters:

- The **grapheme** does *not* specify the size, shape, or orientation on the screen
- The **phoneme** does *not* specify the acoustic realization of a sound

<table>
<thead>
<tr>
<th>Form</th>
<th>Genuine Abstraction</th>
<th>Particular Realization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>U+0041 (unicode) = Grapheme</td>
<td>Α,Α,Α,Α,υ = Glyphe</td>
</tr>
<tr>
<td>Speech</td>
<td>/t/ = Phoneme</td>
<td>[t] = Phone</td>
</tr>
</tbody>
</table>

- The exact definition of a **grapheme** depends on the character code set
- The definition of a **phoneme** depends on the language
  (However even for the same language, linguists often disagree on the phoneme set)
Phonetic Alphabets

- **IPA**: International Phonetic Alphabet
  - designed by the International Phonetic Association
  - Inventory for all sounds of the world’s languages
  - Example: /ˈɔutəməˌtɪʃə/ /ˈspaːɾəˌkɛrkenəŋ/

- **Worldbet**:
  - 1:1 mapping of IPA symbol set into ASCII-7 symbol set to make it suitable for computer
  - James Hieronymous
  - Example: /ˈɔutəmə.tɪʃə/ /ˈspaːɾə.ˈkɛrkenəŋ/

- **Sampa**:
  - Also ASCII-7 symbol set but originally designed for German and later for other Indo-European languages
  - Recently extended to more languages (X-Sampa)
  - Example: /ˈɔutəmə.tɪʃə/ /ˈspaːɾə.ˈkɛrkenəŋ/
# IPA Scheme for Consonants

**The International Phonetic Alphabet (revised to 1993)**

<table>
<thead>
<tr>
<th>CONSONANTS (PULMONIC)</th>
<th>Bilabial</th>
<th>Labiodental</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Postalveolar</th>
<th>Retroflex</th>
<th>Palatal</th>
<th>Velar</th>
<th>Uvular</th>
<th>Pharyngeal</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plosive</td>
<td>p b</td>
<td></td>
<td></td>
<td>t d</td>
<td></td>
<td>t d</td>
<td>c f j</td>
<td>k g q</td>
<td>g</td>
<td></td>
<td>?</td>
</tr>
<tr>
<td>Nasal</td>
<td>m mj</td>
<td></td>
<td></td>
<td>n</td>
<td></td>
<td>n j</td>
<td>n n j</td>
<td>n</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trill</td>
<td>B</td>
<td></td>
<td></td>
<td>r</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>Tap or Flap</td>
<td>r</td>
<td></td>
<td></td>
<td>t</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fricative</td>
<td>φ β f v</td>
<td>θ ø s z</td>
<td></td>
<td>s z</td>
<td>c i x y</td>
<td>χ κ h f h f</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximant</td>
<td>u j l j w</td>
<td></td>
<td></td>
<td>l</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral approximant</td>
<td>l</td>
<td></td>
<td></td>
<td>l</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

---

**Location where sound is produced**

<table>
<thead>
<tr>
<th>Location where sound is produced</th>
<th>Influence of air flow</th>
</tr>
</thead>
</table>

*IPA Scheme for Consonants*
IPA Scheme for Vowels

- Horizontal position of the tongue
- Vertical position of the tongue
- Rounded shape of the lips
- Unrounded shape of the lips

Where symbols appear in pairs, the one to the right represents a rounded vowel.
Vowels

Vowels are characterized basically by three parameters, the tongue placement and shape and the lips shape.

1. Vertical Position of the Tongue:
The higher the tongue is placed (vocal tract is more closed) the higher a vowel will sound.
Example: [i:] in BEAT is higher (more closed) than [e] in BET.

2. Horizontal Position of the Tongue:
The more at the front the highest point of the tongue is, the "brighter" the vowel will sound.
Example: [i:] in BEAT is a front vowel and [u:] in BOOT is a back vowel.

3. Shape of the Lips:
Depending on the shape of the lips, we call a vowel rounded or unrounded.
Example: [æ] in BAT is unrounded and [o] in BOAT is rounded.
### The Vowel-Quadrangle

**Horizontal position of the tongue (Ort der Verengung / Behinderung des Luftzugs):**

- BEAT
- BAIT
- BIT
- BET
- BAT

**Vertical position of the tongue (Höhe der Zunge am Gaumen):**

- Close
- Close-mid
- Open-mid
- Open

**Unrounded shape of the lips:**

- a
- o

**Rounded shape of the lips:**

- å
- ø

**VOWELS:**

- Front: i, y, I, Y
- Central: u, u, ʉ, ɻ
- Back: w, u, æ, ɒ

Where symbols appear in pairs, the one to the right represents a rounded vowel.
Different Shapes of the Vocal Tract

- **BEAT**: front
- **BIT**: close
- **BAIT**: open
- **BET**: back

Horizontal position of the tongue
Vertical position of the tongue
Formants

- The *resonance frequencies* of an acoustic tube (*vocal tract*) are called **formants**
- Formants are the distinguishing *frequency components* of human speech
- The information that humans require to distinguish between vowels can be represented quantitatively by the frequency content of the vowel sounds
- The formant with the lowest frequency is called $F_1$, the 2nd $F_2$, and the 3rd $F_3$
- Most often the two first formants, $F_1$ and $F_2$, are enough to disambiguate the vowel.
- **Spectrograms** are used to visualize formants.

Spectrogram of 3 vowels
y-axis: Frequency
Dark: High energy
$F_1$, $F_2$: Formants

- $F_1$: major resonance of the pharyngeal cavity
- $F_2$: major resonance of the oral cavity
Formants F1 and F2

This is a plot of measured formants for different vowels from different speakers:

The so called vowel-triangle expresses which vowels have which formants in average:

F1: major resonance of the pharyngeal cavity
F2: major resonance of the oral cavity
Formants F1 and F2

- The part from glottis to tongue (pharyngeal part) is longer than the forward part from the tongue (oral part), thus the resonance is lower.
- Tongue placement and oral tract shape determine F1 and F2.
- Rounding the lips has the effect of extending the forward part thus lowering F2.
- Example: beat
  - tongue is far forward
  - longer pharyngeal part
  - lower F1,
  - forward part extremely short
  - higher F2

F1: major resonance of the pharyngeal cavity
F2: major resonance of the oral cavity
Diphthongs

The characteristics F1 and F2 are sometimes called **formant targets**

- Vowels which have *one* specific target are called **monophthongs**

- Vowels which combine two distinct sets of targets are called **diphthongs**

Here the initial vowel target glides smoothly to the final configuration

- Some languages (like Mandarin) even have **triphthongs**

(Diphthong often as long as monophthong)
As opposed to vowel, consonants are characterized by constriction or obstruction in the pharyngeal and/or oral cavities.

- Consonants are classified by manner and place of articulation.
- **Manner of articulation** refers to the articulation mechanism.
- **Place** refers to the location of the major constriction.

Other characteristics:

- **Sonority:** continuous voicing - liquids (*rat, lean*), glides (*yes*).
  - Non-sonority requires (close to) complete obstruction.

- **Voicing:** even non-sonorant consonants may have some voicing. Before the obstruction occurs, in some consonants the vocal folds are vibrating thus making the sound voiced, e.g. *Z/S, ZH/SH, B/P, D/T, G/K, V/F*.

- **Aspiration:** consonants can be **aspirated** (e.g. T in *THOMAS*), they can be strong (fortes) or weak (lenes).
Classification of Consonants

Manner of Articulation:

• **Plosive:** Closure in oral cavity - completely blocked (B D G P T K)
  (Verschlusslaut)
• **Nasal:** Velum closes oral cavity – air goes through nose (M N NG)
• **Fricative:** turbulent airstream noise, constriction (F V S Z SH ZH)
  (Reibelaut)
• **Lateral:** (vowel-like) the air passes left and right of the tongue (L)
• **Retroflex:** (vowel-like) tip of tongue vibrates also curled back (vibrated R)
• **Glide:** vowel-like (Y, W)
• **Affricate:** combination of plosive + fricative (PF TS)

Place of Articulation:

- **Labial:** (Lippen) m, p, b, w
- **Dental:** (Zähne) th, dh
- **Alveolar:** (die Zahnleiste betreffend) t, d, n, s, z, r, l
- **Palatal:** (harter Gaumen) sh, zh, y
- **Velar:** (weicher Gaumen) k, g, ng
Another plosive is the so called **glottal stop** where the airflow is interrupted by closing the vocal cords in the glottis.
Plosives – Spectrograms

"bab", "dad", "gag"

The examples of this lecture are from
- http://home.cc.umanitoba.ca/~robh/howto.html
Vocal Tract Shapes of Nasals

- Lip Closure
- Lips-Teeth Closure
- Velar Closure

TOM  TON  TONGUE
Nasals can be identified by areas of little or no spectral energy (fully reflected frequencies result in wave cancelation).

In "dinner", "dimmer", "dinger", the final nasals have identifiable formants that are less in amplitude than in the vowel, and the regions between them are blank.
Introduction in Spoken Language - 25

Vocal Tract Shapes of Fricatives

Lip-Teeth Friction

Tongue-Teeth
Tongue-Alveoli

Palatal Friction
Alveolar Friction

Palatal Friction

FAN, VAN
SUE, ZOO
VISION, VICTIOUS
YOU

Additionally there is a glottal fricative \( /\text{h}/ \) as in HOUSE. Other languages often also have aspirated velar and palatal fricatives.
Fricatives – Spectrograms
English does not make full use of all possible mechanisms.

Other languages require even more mechanisms:

- Chinese: tonal language (Mandarin 4 tones + neutral)
- Japanese: vowel length is distinctive
- Spanish: trilled vs. implosive r
Speech Signal and Spectrogram

- Speech signal and spectrogram of the word “phonetician“
- Containing examples for vocals, nasals, plosives, and fricatives
Overview

• Phonetics and Phonology
  – Sound Inventories among Languages
  – Sound Unit Sharing across Languages
  – Representations of Phonemes
    • Allophones
    • Coarticulation
    • Phones and Context
    • Prosody
    • Spontaneously spoken speech

• Syllables and Words
  – Syllables
  – Words
Sound Inventories among Languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Phone set</th>
<th>Dict</th>
<th>Read Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>∑ C V</td>
<td>C V</td>
<td>C V</td>
</tr>
<tr>
<td>Chinese</td>
<td>48 48.9 51.1</td>
<td>55.9 44.1</td>
<td>55.0 45.0</td>
</tr>
<tr>
<td>Croatian</td>
<td>30 83.3 16.6</td>
<td>54.8 45.2</td>
<td>52.5 47.5</td>
</tr>
<tr>
<td>German</td>
<td>43 51.2 48.8</td>
<td>61.0 39.0</td>
<td>60.5 39.5</td>
</tr>
<tr>
<td>Japanese</td>
<td>31 67.7 32.3</td>
<td>48.2 51.8</td>
<td>51.4 48.6</td>
</tr>
<tr>
<td>Korean</td>
<td>41 56.1 43.9</td>
<td>54.9 45.1</td>
<td>54.6 45.4</td>
</tr>
<tr>
<td>Portuguese</td>
<td>46 45.6 54.3</td>
<td>47.7 52.3</td>
<td>50.1 49.9</td>
</tr>
<tr>
<td>Russian</td>
<td>47 78.7 21.3</td>
<td>56.1 43.9</td>
<td>55.9 44.1</td>
</tr>
<tr>
<td>Spanish</td>
<td>40 60.0 40.0</td>
<td>53.9 46.1</td>
<td>54.0 46.0</td>
</tr>
<tr>
<td>Turkish</td>
<td>29 72.4 27.6</td>
<td>53.5 46.5</td>
<td>53.2 46.8</td>
</tr>
</tbody>
</table>

Consonant (C) to Vowel (V) ratio and phone-based error rates for nine languages

- Phone error rate correlates with number of phonemes used to model a language

Source: Multilingual Speech Processing, Schultz & Kirchhoff (ed.), Chapter 4, p.86
Sound Unit Sharing across Languages (1)

1) Build **universal sound inventory** based on IPA: 485 sounds are reduced to 162 IPA-sound classes

2) Each sound class is represented by one “phoneme” which is trained through **data sharing** across languages
   - $m, n, s, l$ occur in all languages
   - $p, b, t, d, k, g, f$ and $i, u, e, a, o$ occur in almost all languages
   - no sharing of triphthongs and palatal consonants
**Sound Unit Sharing across Languages (2)**

<table>
<thead>
<tr>
<th>Shared by</th>
<th>#</th>
<th>Modeled Phonemes (IPA symbols)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>4</td>
<td>Consonants: m,n,s,l</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Vowels: i,u,e</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>η,ν,ζ,j</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>r,h,tf</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>η,ν,ζ,j</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>i,y,ε,ω</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>η,ν,ζ,j</td>
</tr>
<tr>
<td>All</td>
<td>83</td>
<td>Polyphonemes shared across ≥ 2 languages</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Consonants: p,b,t,d,k,g,f</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vowels: a,o</td>
</tr>
<tr>
<td>CH</td>
<td>15</td>
<td>Consonants: tʃ,ɹ,ʃ,ç,ç,ç,h</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vowels: i,ί,ι,ι,υ,υ,υ</td>
</tr>
<tr>
<td>EN</td>
<td>5</td>
<td>ι,ι,ι,ι,ι,ι,ι,ι,ι,ι,ι,ι</td>
</tr>
<tr>
<td>FR</td>
<td>5</td>
<td>ι,ι,ι,ι,ι,ι,ι,ι,ι,ι,ι,ι</td>
</tr>
<tr>
<td>GE</td>
<td>3</td>
<td>ι,ι,ι,ι,ι,ι,ι,ι,ι,ι,ι,ι</td>
</tr>
<tr>
<td>JA</td>
<td>2</td>
<td>ι,ι,ι,ι,ι,ι,ι,ι,ι,ι,ι,ι</td>
</tr>
<tr>
<td>KO</td>
<td>14</td>
<td>Consonants: p,ɬ,ɬ,ɬ,ɬ,ɬ,ɬ,h</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vowels: i,ᵀ,ᵀ,ᵀ,ᵀ,ᵀ,ᵀ</td>
</tr>
<tr>
<td>KR</td>
<td>1</td>
<td>η</td>
</tr>
<tr>
<td>PO</td>
<td>8</td>
<td>ι,ι,ι,ι,ι,ι,ι,ι,ι,ι,ι,ι</td>
</tr>
<tr>
<td>RU</td>
<td>15</td>
<td>Consonants: p,ɬ,ɬ,ɬ,ɬ,ɬ,ɬ,h</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vowels: j,_jet,jet,jet,jet</td>
</tr>
<tr>
<td>SP</td>
<td>2</td>
<td>η,ν,ζ,j</td>
</tr>
<tr>
<td>SW</td>
<td>9</td>
<td>η,ν,ζ,j</td>
</tr>
<tr>
<td>TU</td>
<td>0</td>
<td>Silence and noises shared across languages</td>
</tr>
</tbody>
</table>

**Share factor:**
Average number of languages sharing the units of a global unit set
Representations of Phonemes

- Why do phonemes differ in their actual representation (pronunciation)?
- How?

1)  Coarticulation - Context
2)  Coarticulation with Varies Speaking Rate
Coarticulation - Allophones

- Phonemes are often modified in a systematic way by its phonetic neighborhood
- This process is called **coarticulation**
- When the variation resulting from the coarticulatory process can be perceived, the modified phonemes are called **allophones**
- Allophonic differences are categorical, i.e. they can be understood and denoted by a small number of symbols
- Example:
  /l/ in “clear”, “light”, “like”:
  Front part of tongue clearly touches the alveolar ridge
  **But**: dark /l/ in “kill”
  Tongue is often not touching any longer but stiffened in the mouth
  → Both are allophones of /l/ conditioned by the position in the syllable (initial vs. final position)
Coarticulation

- In continuously spoken speech (with varying speaking rate)
  - Formant targets are less likely to be reached
  - Stress patterns might be deleted
  - Modification of sounds occur (assimilation)
  - Sounds are completely deleted (elision)

- **Principle of efficiency**
  - *Minimize the articulatory effort*
    (but keep the information at its maximum)
  - Increase speaking rate (speaker dependent)
  - Reduce articulatory effort

- Sounds *within* syllables influence one another‘s realization more than *across* syllable boundaries
Phones and Context
Prosody

• Besides regular intonation of each sound, a phrase can have its own melody.

• The prosody carries information about:
  – **Intention of the utterance** (question, command, statement)
  – **Relevance** (putting focus of attention to a specific part of it)
  – **Resolving syntactic / semantic ambiguities**
  – Describing the current **mood / emotions** of the speaker

• Enriching an utterance with prosodic information:
  – **Intonation (pitch)**: produces a "melody"
  – **Pauses**: are used as markers for focus of attention or disambiguation
  – **Stress**: increase of loudness and pitch
  – **Rhythm**: the alternation of high power (sonorant) and low power sounds
Spontaneously spoken speech

- Used in everyday life to communicate with other humans
- ... in opposition to planned speech
  - speaking while thinking and vice versa
  
  - Example:
    "I think we uhm we could meet maybe at hmm ah yeah maybe September 19th [pause] or uhm [laughter] in fact I have another mee- ah another meeting at this day so ah how about uhm how about September 21st."

- false starts, repetitions, hesitations, filled and non-filled pauses, non-verbal noises
- ill-formed grammar, sentences
- lots of coarticulation effects \( \rightarrow \) speech disfluencies
Overview

• Phonetics and Phonology
  – Sound Inventories among Languages
  – Sound Unit Sharing across Languages
  – Representations of Phonemes
    • Allophones
    • Coarticulation
    • Phones and Context
    • Prosody
    • Spontaneously spoken speech

• Syllables and Words
  – Syllables
  – Words
Syllables and Words

- **Phonemes** are small blocks. They are easy to discriminate but they do not have a meaning by themselves.

- In order to contribute to language meaning, they must be organized into longer cohesive spans.

- The longer units must be combined in characteristic patterns to be meaningful.

- The patterns might be different in structure and length depending on the language.

- Usually these patterns are
  - Syllables
  - Words
Syllables (1)

- Concepts of syllables is thought to interpose between phone and word level
- Linguists define syllables as a unit with internal structure. It consists of:
  - Onset (initial consonant before the vowel peak – if any)
  - Rime containing:
    - Nucleus (vowel peak)
    - Coda (trailing consonant/s)
- In English syllables are centered around vowels (tom-cat); To split a word into syllables we have to judge about consonant affiliations
- Done either by articulatory or perceptual criteria (!unsolved)
Syllables (2)

- **Syllable centers** are thought of as *peaks in sonority* (high-amplitude, periodic sections of the speech waveform).
- The affiliation of consonants to syllables can be determined by comparing the sonority of each phoneme.
- Ranking of increasing sonority:
  
  stops < plosives < affricates < fricatives < nasals < approximants < vowels

- Example *verbal*:
  
  possible are *verb-al* or *ver-bal* but not *ve-rbal* since: $r > b < a > l$

- As long as sonority conditions are met, affiliation of consonant is ambiguous.

- There are also other criteria such as higher order considerations of the word structure.

  Example:

  *beekeeper* $\rightarrow$ *bee-keeper*: the structure blocks the affiliation.

  But *beaker* could be either *bea-ker* or *beak-er*. 
Words

• In Indo-European languages the concept of *word* is intuitively obvious – in the written form words are separated from each other by whitespaces.

• Loosely a *word* is defined as lexical item with a meaning (in a given community) that has the freedom of syntactic combination by its type (noun, verb, …).

• In languages like Japanese, Chinese or Thai no segmentation is given and the concept of a *word* above is no longer unambiguous.

• In spoken languages *words* are not marked by boundaries.

• However, some phrases include pauses 
  Example: „Never give all the heart, for love“
  = nevergivealltheheart // forlove“
  – these units are *intonation phrases*.
Lexical Part-of-Speech (POS)

- Assigning of a **word-type category** to each word form in order to summarize syntactical or pragmatic facts

- Typical set of POS categories:
  - Noun (refer to persons, places, things)
  - Verb (indicate relation between entities)
  - Adjective (specify noun references)
  - Adverb (specify verbal relations)
  - Interjection (express reaction)
  - Conjunction (join phrases)
  - Determiner (narrow noun references)
  - Preposition (denote spatial and temporal relations)
  - Pronoun (substitute for introduced noun phrases)

  - **Word Classes**: (Data-driven) process of grouping words together according to similarity of usage (semantic meaning) for Language Modeling

- **Content words**
- **Function words**
Morphology

- **Morphology:**
  patterns of word formation (inflection, derivation, compounds)

- **English morphology is relatively simple**
  Inflection: person and number agreement, tense marking
  Derivation: productive pre- and suffixes re-, pre-, -ism, -ish, -ity ...
  Compounds: usually max two roots are compounded

- **German (Compounds)**
  Donau-dampf-schiffahrts-gesellschafts-kapitän
  The captain of the company that operates the steamboats on the Donau River

- **Turkish (Inflection and Derivation)**
  behaving as if you were of those whom we might consider not converting into Ottoman
# Morphology & OOV

<table>
<thead>
<tr>
<th>Language</th>
<th>Vocabulary</th>
<th>OOV-Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Korean</td>
<td>64K</td>
<td>34.0%</td>
</tr>
<tr>
<td>Turkish</td>
<td>64K</td>
<td>13.5%</td>
</tr>
<tr>
<td>German</td>
<td>61K</td>
<td>4.4%</td>
</tr>
<tr>
<td>Portuguese</td>
<td>60K</td>
<td>4.3%</td>
</tr>
<tr>
<td>English</td>
<td>64K</td>
<td>0.3%</td>
</tr>
<tr>
<td>Korean (segmented)</td>
<td>64K</td>
<td>0.2%</td>
</tr>
<tr>
<td>Chinese (segmented)</td>
<td>60K</td>
<td>0%</td>
</tr>
<tr>
<td>Croatian</td>
<td>31K</td>
<td>13.6%</td>
</tr>
<tr>
<td>Spanish</td>
<td>30K</td>
<td>5.2%</td>
</tr>
<tr>
<td>French</td>
<td>30K</td>
<td>4.7%</td>
</tr>
<tr>
<td>Japanese (segmented)</td>
<td>22K</td>
<td>3.0%</td>
</tr>
</tbody>
</table>

Source: Multilingual Speech Processing, Schultz&Kirchhoff (ed.), Chapter 4, p.88
Overview

• Syntax and Semantics
  – Syntax
  – Semantics
  – Language & Compactness

• Transcription of Spoken Speech
  – Transcription vs. Transliteration
Syntax

- **Syntax** is the study of the
  - patterns of formation of sentences and phrases from words
  - rules for the formation of grammatical sentences
- Phrase schemata: create simple uniform template input from POS
- Parse Tree Representation
Semantics

- **Semantic** deals with the study of meaning (structure of meaning in language and changes in meaning over time)
- **Semantic Roles**: try to make sense of the participants in an event and provide vocabulary to ask w-questions
- **Lexical Semantics** is the level of meaning before words are composed into phrases and sentences
  - Practical problem: **Polysemy**: context-dependent resolution of word sense;
    Example: *bank* (*river bank, money in the bank*);
    (POS, mutual information, frequency analysis, a-priori)
- **Logical From**: To solve lexical, syntactic, and semantic ambiguities, we need external context
  - meta languages like the predicate logic are used to represent the logical form of a language
Figure 1.7: Compactness for nine EU-Languages: Vocabulary vs Word Tokens (left) and vs Graphemes (right)

Multilingual Speech Processing, Schultz&Kirchhoff (ed.), Chapter 4, p.90
Overview

• Syntax and Semantics
  – Syntax
  – Semantics
  – Language & Compactness

• Transcription of Spoken Speech
  – Transcription vs. Transliteration
Transcription of Spoken Speech

• To train and evaluate a speech recognizer, we need the **acoustic representation** of speech PLUS the corresponding **textual representation** of what was said.

• Textual representation is done by native experts.

• Representation can be given at three different levels:
  – **orthographic representation** of the spoken text (word level) often called „surface form“
  – **phonetic-canonical** representation of written words
  – **phonologic representation** of actual pronunciation

• Nowadays representations are only provided at the **word-level** (for efficiency purposes, i.e. time and costs):
  – Planned speech: ~ 6 x Real Time
  – Conversational speech: ~20 x Real Time
  – Careful transcriptions, xtalk, disfluencies: 40 x Real Time, multiple passes

• Since 2005, so-called **Quick Transcriptions (QT)**: Automatically produced by recognizer, then manually cross-checked by human expert.
Transcription vs. Transliteration

• To represent proper names there are two concepts:
  
  – **Transliteration**: the original form is transformed into the writing system of the target language “Gorbacev”
  
  – **Transcription**: quasi phonemic representation in the target language:
    
    • “Gorbachev” in English
    • “Gorbatschow” in German

• Advantages and Disadvantages:
  
  – **Transliteration** is somewhat arbitrary since sometimes the target language lacks the appropriate symbols, this results in the introduction of symbols and diacritics
  
  – **Transcription** is inconsistent since it depends on the actual target language
Thanks for your interest!