# Human Communication 1
# Lecture 22

# Transcribing and Annotating Dialogue
# How and Why to create a Spoken Language Corpus in Computer Science

# Dialogue

Before we can start creating a language spoken corpus we need to know what dialogue is

One definition says it is:

A conversation between 2 or more people

# Dialogue = Conversation

A conversation between 2 or more people

What does this definition imply?

- There are communication messages exchanged between 2 or more people
- These messages are encoded in sound strings
- The sound is travelling through the space between the dialogue participants
- The participants perceive the sound, encode and understand it

# Dialogue = Conversation = Communication

A conversation between 2 or more people

- Is it only the sound that transforms a conversation into communication?

  No, there is more

- Silence during speech can communicate a message as well as sound
- When we can see our dialogue partner/s, we are able to pick up nonverbal communication clues

# Transcribing and Annotating

What do I need to transcribe speech sound?

- Focusing on speech: All speech sound you can hear need to be transcribed

What do I need to annotate?

- All metadata you want to add to your transcription needs to be annotate
  - In case of speech that can be no-human noise, other human sound than speech (coughing, laughing, filled pauses like errr, mhm, etc....)

# How to start your spoken language corpus

In computer science

- First you need to know what you want to achieve
- What does your system need to learn?
- For example you might want to develop a mobile translation system for the translation of spontaneous speech in face-to-face situations

# The idea

What comes next?

- What do you want your system to do?
  - In this case you want a system that recognises, translates and produces natural utterances

Next?

- Even if you would love to have a system that can recognise all world languages you want robust results quickly and you are also bound to a certain amount of project time/money

# Languages

- Therefore you need to restrict your ambitions to specific languages at first

  Which languages?
- This leads you to the group you think will mainly use your system:
- Businessman
  - Why? → Companies can afford to buy such a system, it might not be inexpensive for a broader costumer group in the beginning

# 3 languages

3 languages spoke in the most influential industrial states

- English/German/Japanese
- Discuss why not Chinese!!

# Creating a Scenario

You have your languages

Now you start creating a scenario

Why?

- You want a spoken language corpus that includes certain words and speech acts to create robust data

  → Meaning you want a variety of different speech acts and words but you also want a nice distribution that consists of the most used speech acts and words

# Scenario 1

- In our case a good scenario would be dialogue about arranging a meeting, since we are targeting business people
- How do you create a scenario that gives you mainly natural spoken language?
- You don't give your future participants a fully scripted dialogue, you just give them the scenario
- E.g.: You are on the phone with a business partner, you want to arrange a meeting on a certain day to a certain time in a certain city

# Scenario 2

- Day, time and city are given

  How to arrange the meeting and how to express this in the dialogue will be up to the participant

  Will this totally be up to the participant?

- Not always, since it is a dialogue one of the dialogue partners can be from your project and s/he can guide the dialogue in a certain direction

# Recording

- Assume you have everything: Idea, languages you want to record, scenario, participants
- Now you can start your recordings
- You will need a soundproof room
- A microphone and a computer (if you are recording for a multi modal corpus you also need camcorder/webcam)
- 2 or more participants, all speaking the **same** language

# Transcribing

- Your data is recorded and digitalised
- Meaning you are ready to transcribe!
  - We remember that we need to decide if we want a phonetic or orthographic transcript
- Let's go with the orthographic
- We can start transcribing
  - Problems:

Spontaneous speech is difficult to punctuate correctly. Transcribers will have to rely on their hearing and intuition as native speakers as well as on their grammatical knowledge in order to structure utterances and sentences in a logical way.

# Annotating

- By annotation our speech data we are adding metadata to our corpus
- In the case of our system we will need to come up with a system that contains all labels of your metadata, you can call it transliteration conventions
- In your conventions you are collecting all labels you find important for your annotation and you also define every label thoroughly

# Annotation Labels

- What labels do I need?
- You can look at categories that might be interesting for your purpose e.g.:

To label phenomenons of spontaneous speech such as false starts, corrections, repetitions, reductions and filled pauses. They also indicate technical artefacts like recording interruptions or microphone noises, as well as speaker interference

# Possible Categories of Annotation Labels

lexical units

syntactic-semantic formation

nonverbal articulatory production

noises

pauses

acoustic interference

comments

special comments

prosody

# Examples for labels

$ → spelling  label added in front of letter: $S $M $I $T $H

& → acronym  label added in front of word: &UN , &$B $B $C

~ → proper noun  label added in front of word:  ~Cameo

# → number  label added in front of word:  #one

* → neologism  label added in front of word: *spanglish

!KEY → command word

label added in front of word:  !KEYcommand

<Z> → lengthening  label added where needed:  a<Z>pple, lemon<Z>

# Basic requirements

a) automatic processing:

consistent file structure, informative and consistent turn names, consistent transliterations, explicit symbolisation, ASCII symbols, parsable transliteration conventions

b) textual requirements

written record of all audible elements of the dialogue, syntactic-semantic labels, labelling of speaker and noise interference, labelling of certain word categories (proper names, numbers) and failures, maintenance of readability

c) the process of transliteration (= transcribing + annotation)

straightforward usability of all conventions, simple and easy-to-understand conventions, even for non-experts

# Limits

- You can set limits by deciding if you want a broad or a narrow transcription

An annotation on the level of a broad transcription does not provide:

- a phonological transliteration
- a phonetic transcription
- a time correlation of the speech signal

For labelling nonverbal productions transcribers can only choose from categories. A pronunciation or event that is particularly noticeable can be referred to by a so-called pronunciation or local comment. Pronunciation comments are an attempt to mirror the phonetic characteristics of pronunciation variation and disfluencies by means of orthography.

# More limits

Speech acts:

- We are able to feed computers with speech acts like greetings, etc.
- Usually in computer science dialogue acts are used:
- These are generic speech acts consisting of e.g. :

A meta question: "What can I say?"
A statement
A request
They are also domain specific → hence scenarios

# Why do we need spoken language corpora

- A lot of research and training is based on the data that is collected in a spoken language corpora
- Speech synthesis
- Speech recognition
  - Interaction between acoustics and linguistics
  - Semantical constructions
  - Lexicon and morphology
  - Pragmatic communicational information

# Challenges

Challenges in spoken language corpora are many. One basic challenge is in design methodology, how to design compact corpora that can be used in a variety of applications; how to design comparable corpora in a variety of languages; how to select (or sample) speakers so as to have a representative population with regard to many factors including accent, dialect, and speaking style; how to create generic dialogue corpora so as to minimize the need for task or application specific data; how to select statistically representative test data for system evaluation. Another major challenge centers on developing standards for transcribing speech data at different levels and across languages: establishing symbol sets, alignment conventions, defining levels of transcription (acoustic, phonetic, phonemic, word and other levels), conventions for prosody and tone, conventions for quality control (such as having independent labelers transcribe the same speech data for reliability statistics).

Lori Lamel & Ronald Cole 1995

# Multimodal corpora 1

Advantages:

- All possible communication channels (hand gesture, facial expression, body movement, emotions) can be better picked up when using video that allows us to create a multimodal corpus
- It enables us to analyse human communication, either with another human or a machine, better
- ...

# Multimodal corpora 2

Challenges

- All that apply to a spoken language corpus
- More time consuming since we have to create annotation labels for each channel you would like to annotate
- Since not everybody is using the same labels and definitions it might be hard to decode and not everybody might agree with labels and their definitions

# Historical Fun Fact

Annotating microrythms in a dialogue

- William Condon spent a year and a half to decode a 4 ½ second segment of film, in which a woman says to a man and a child, over dinner: *"You all should come around every night .We never had a dinner time like this in months."*
- He broke the film in individual frames each about 1/45 of a second
- He found out that the wife is turning her head exactly as the husbands hand comes up : He proofed that in dialogue we are synchronising our body movement with each other