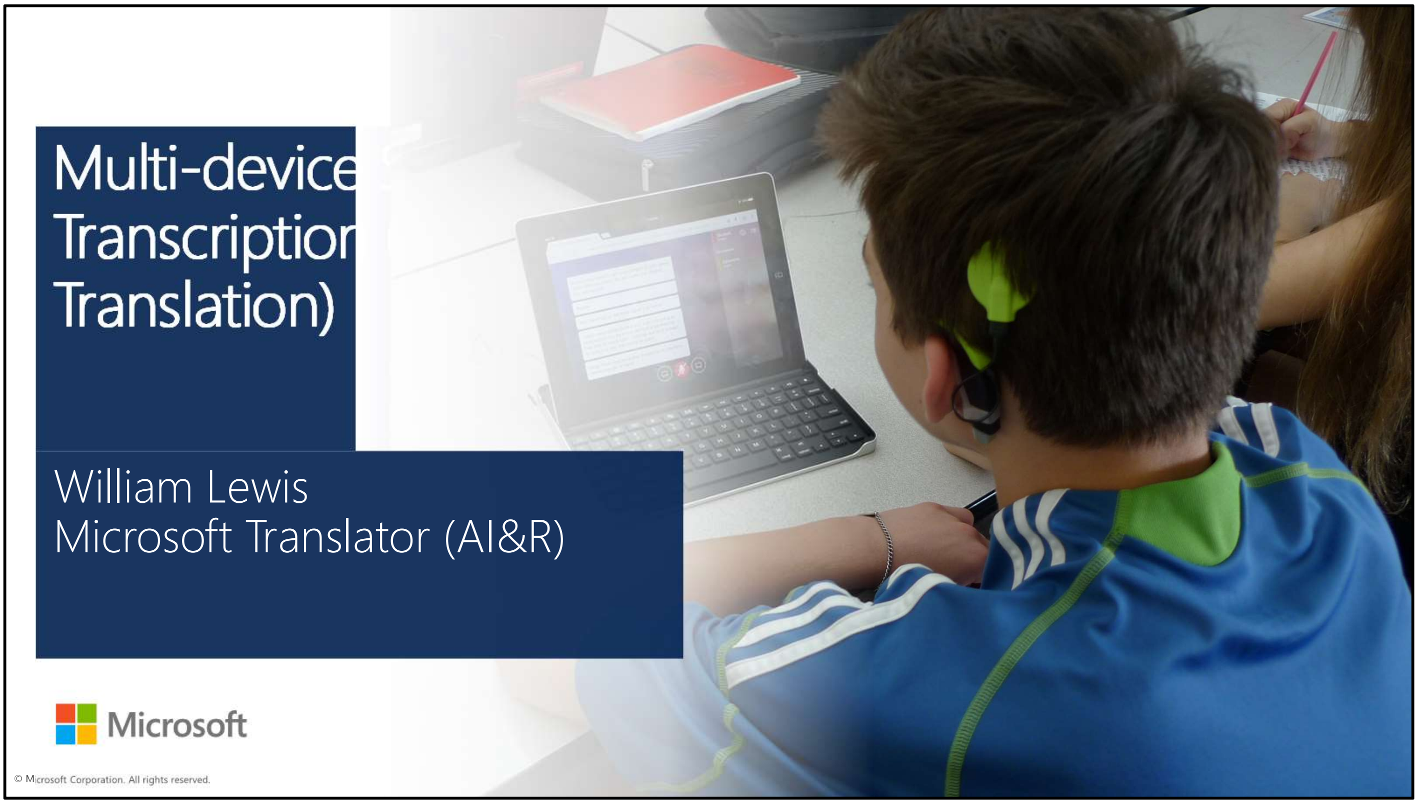


Multi-device Transcriptior Translation)

William Lewis
Microsoft Translator (AI&R)

Microsoft

Although our team works on Machine Translation, most specifically and recently, Speech Translation, quality speech transcription is a crucial byproduct. As I go through the talk, recognize that a lot of our work is focused on generating the best speech transcription we can, the net result being higher quality translation as a result.

To join the conversation, either type in the room code, or scan the QR code with your phone.

What you're seeing on the screen is transcription from our PowerPoint add-in, called the Presentation Translator, which was customized a few minutes before using our service (I'll go into that later). The transcriptions provided can be sent to any devices running our service by connecting to the specific QR code that is listed on the screen, or by typing the URL.

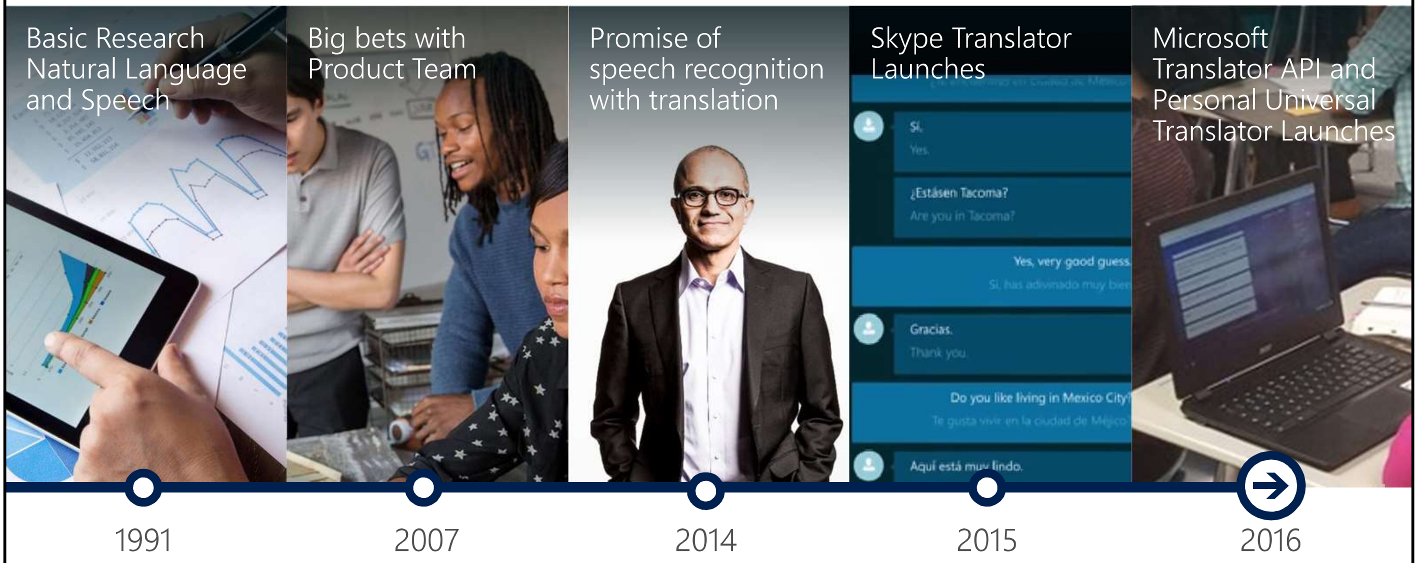
Some of the FCC folks who are here:

- Eliot Greenwald, Patrick Webre, Karen Strauss, Robert Aldrich, Michael Scott, and Sue Bahr.

The MS folks who are here:

- Will Lewis, Paula Boyd, Lissa Shook, and Bradley Baird

The History of Microsoft Translator



Basic Research Natural Language and Speech

1991

Big bets with Product Team

2007

Promise of speech recognition with translation

2014

Skype Translator Launches

Sí.
Yes.

¿Estásen Tacoma?
Are you in Tacoma?

Yes, very good guess.
Sí, has adivinado muy bien

Gracias.
Thank you.

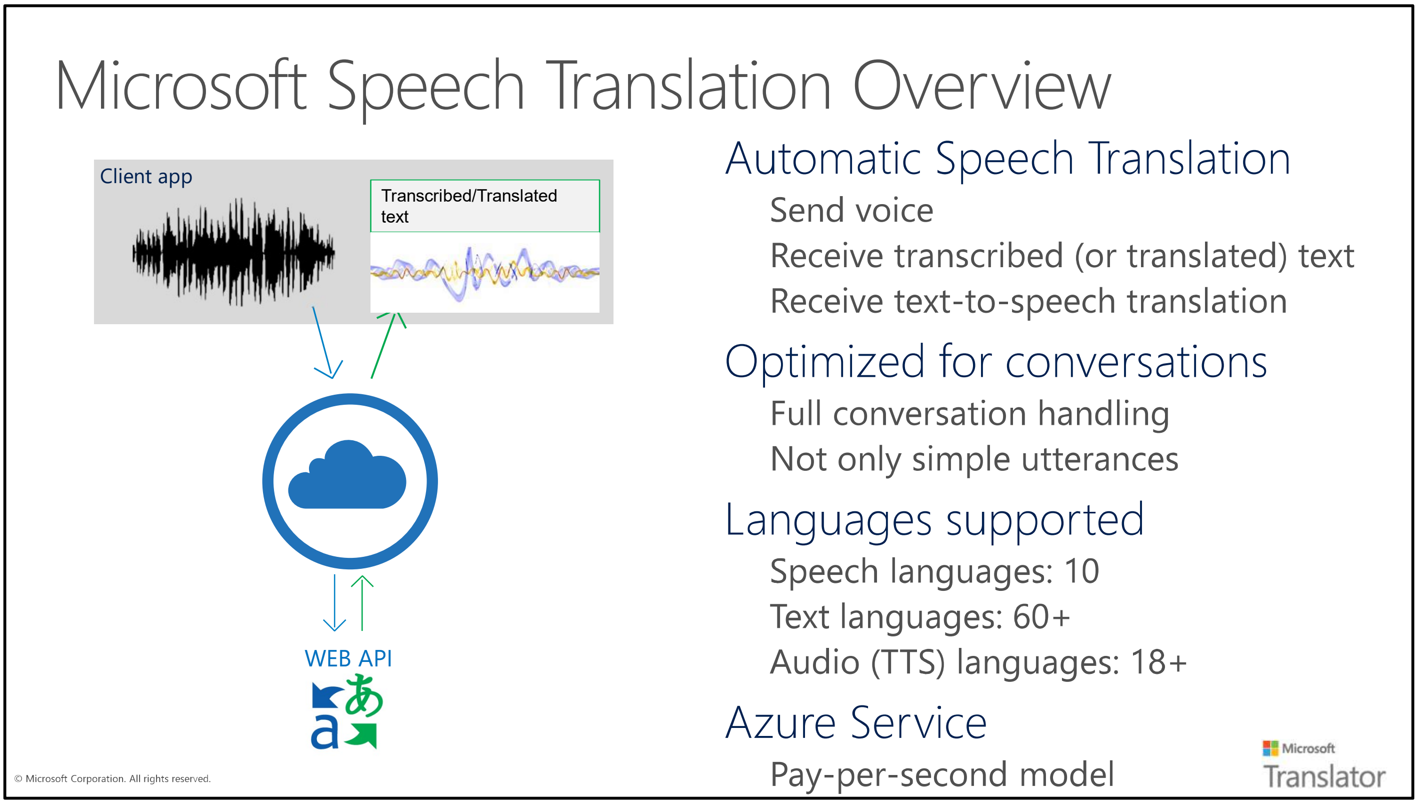
Do you like living in Mexico City?
Te gusta vivir en la ciudad de México

Aquí está muy lindo.

2015

Microsoft Translator API and Personal Universal Translator Launches

2016



Microsoft Speech Translation Overview

Automatic Speech Translation

Send voice

Receive transcribed (or translated) text

Receive text-to-speech translation

Optimized for conversations

Full conversation handling

Not only simple utterances

Languages supported

Speech languages: 10

Text languages: 60+

Audio (TTS) languages: 18+

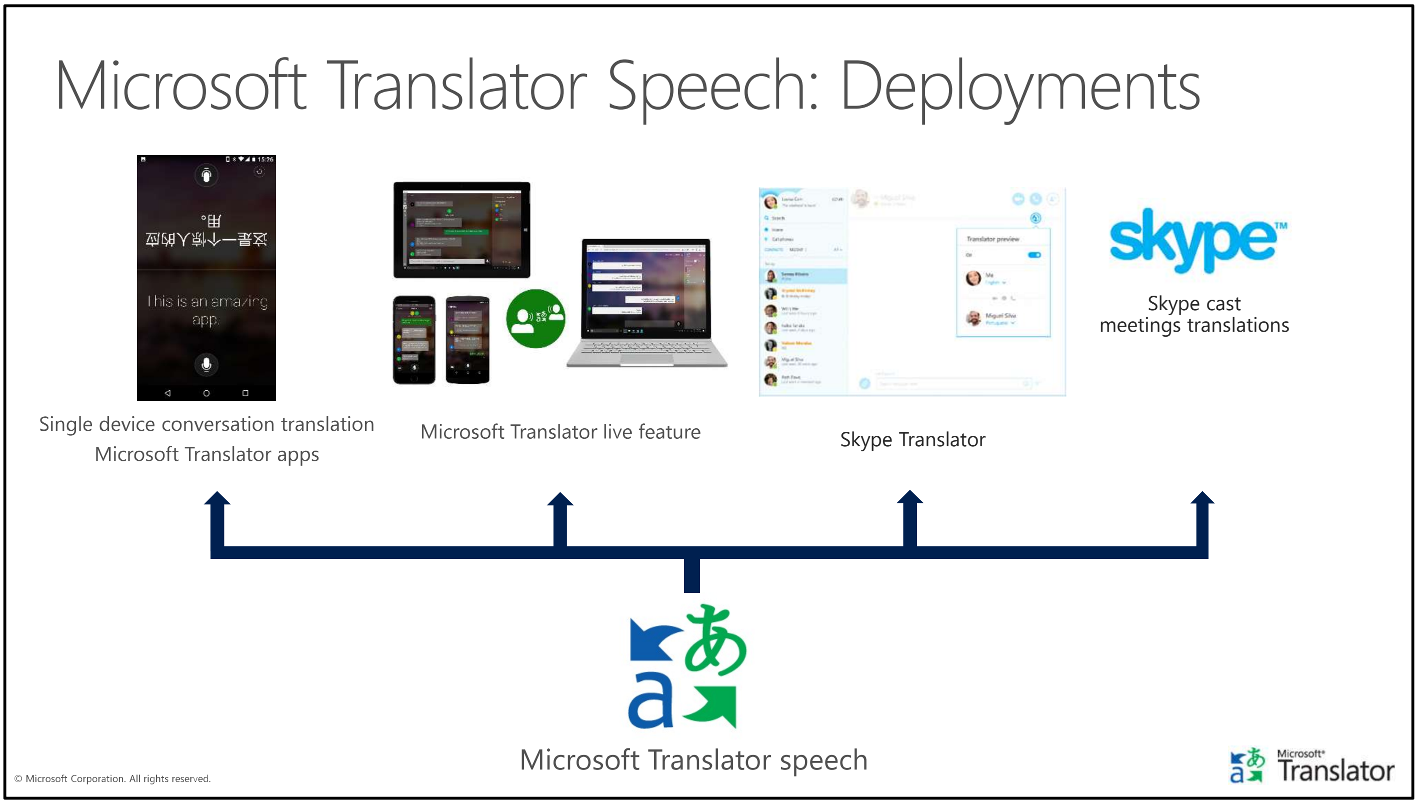
Azure Service

Pay-per-second model

Microsoft Translator

A subset of these 60+ languages, currently 18, also have “voices”, meaning you can listen to the audio in another language, as the speaker speaks in one of the 10 input languages. Handles conversations and spontaneous speech, including a variety of disfluencies, etc.

Pricing here: https://azure.microsoft.com/en-us/pricing/details/cognitive-services/speech-api/, ~$1-2/hour



Microsoft Translator Speech: Deployments

Single device conversation translation
Microsoft Translator apps

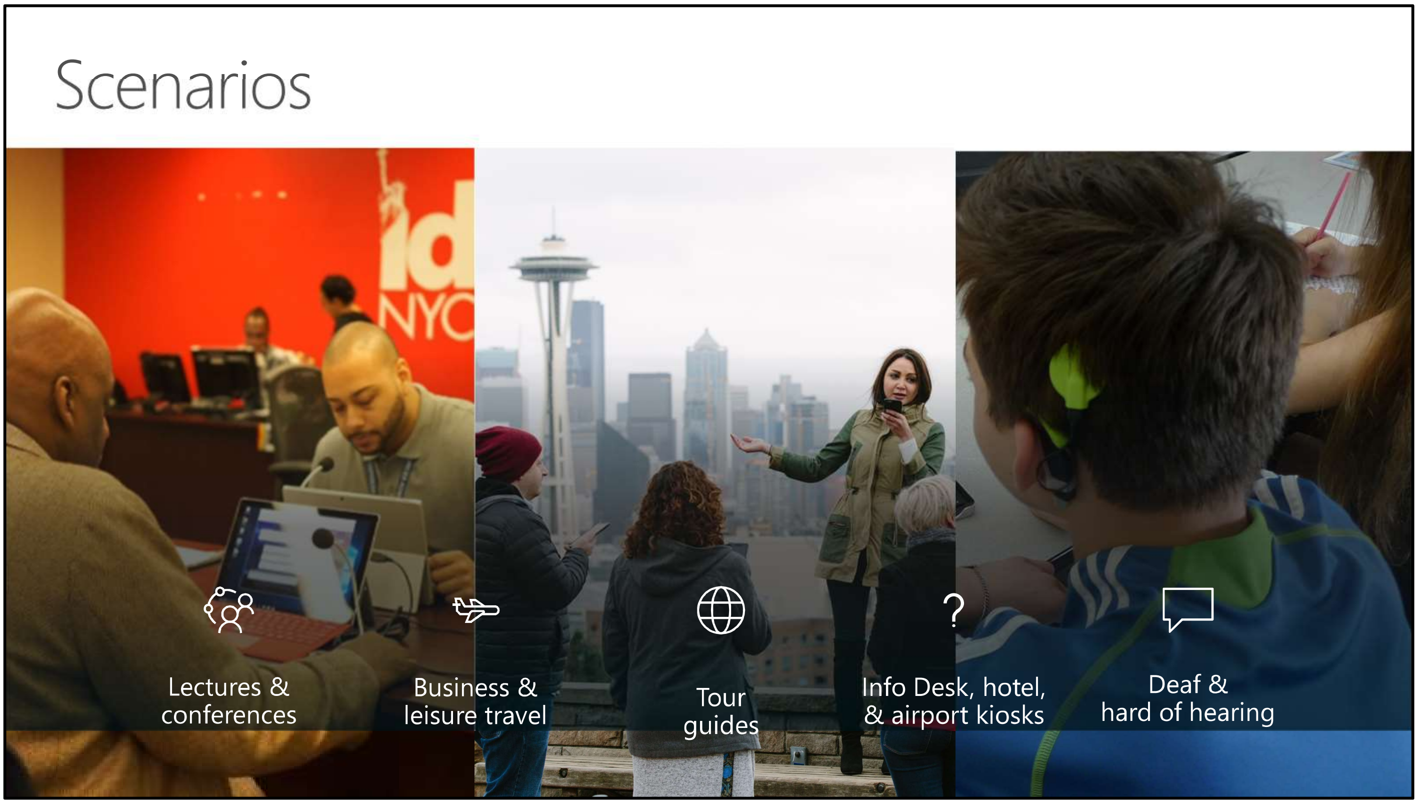
Microsoft Translator live feature

Skype Translator

Skype cast
meetings translations

Microsoft Translator speech

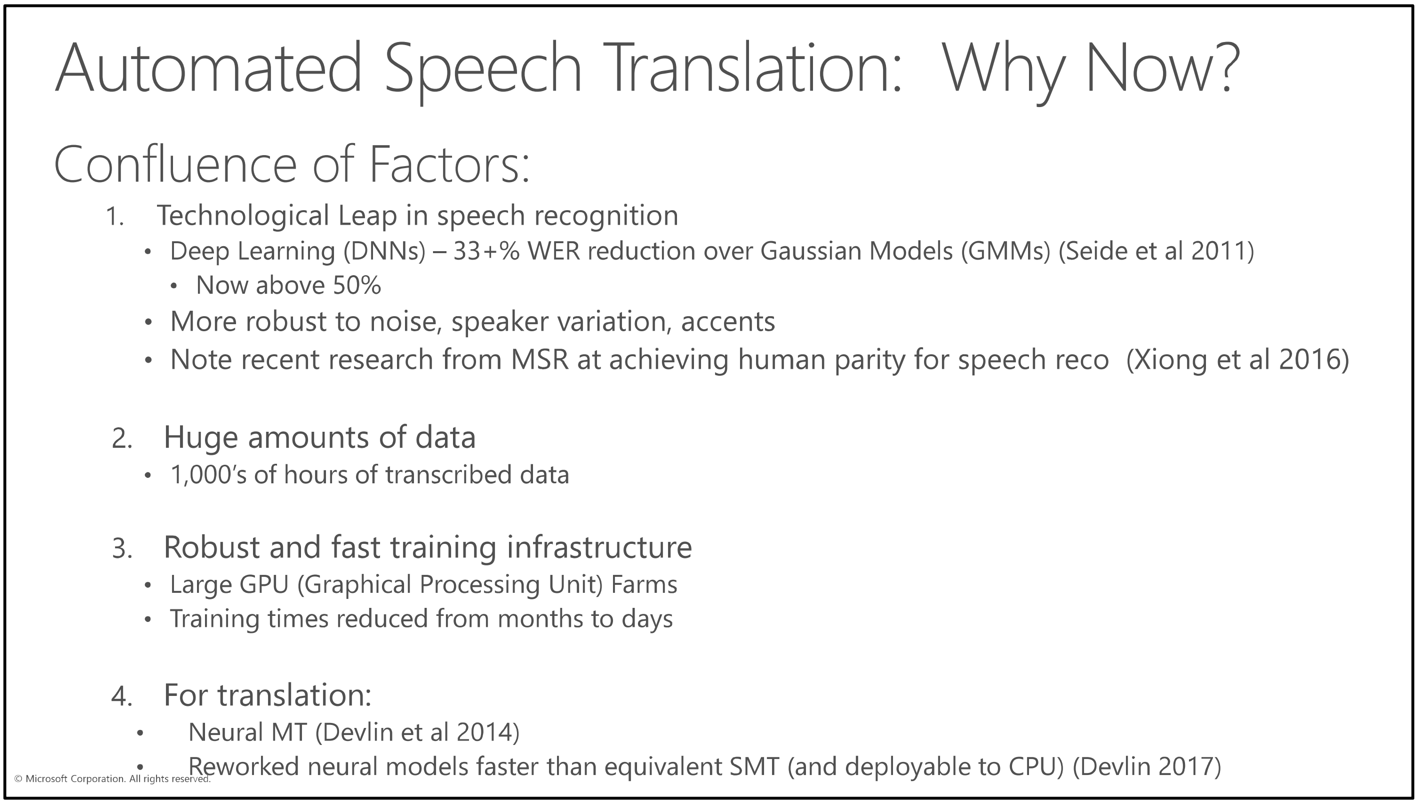
We develop both an developer infrastructure for first and third party developers and enterprises. More than just prototyping: filling perceived need through agile development cycles (witness Skype Translator and Microsoft Translator live).



Scenarios

- Lectures & conferences
- Business & leisure travel
- Tour guides
- Info Desk, hotel, & airport kiosks
- Deaf & hard of hearing

Conversational Speech Recognition



Automated Speech Translation: Why Now?

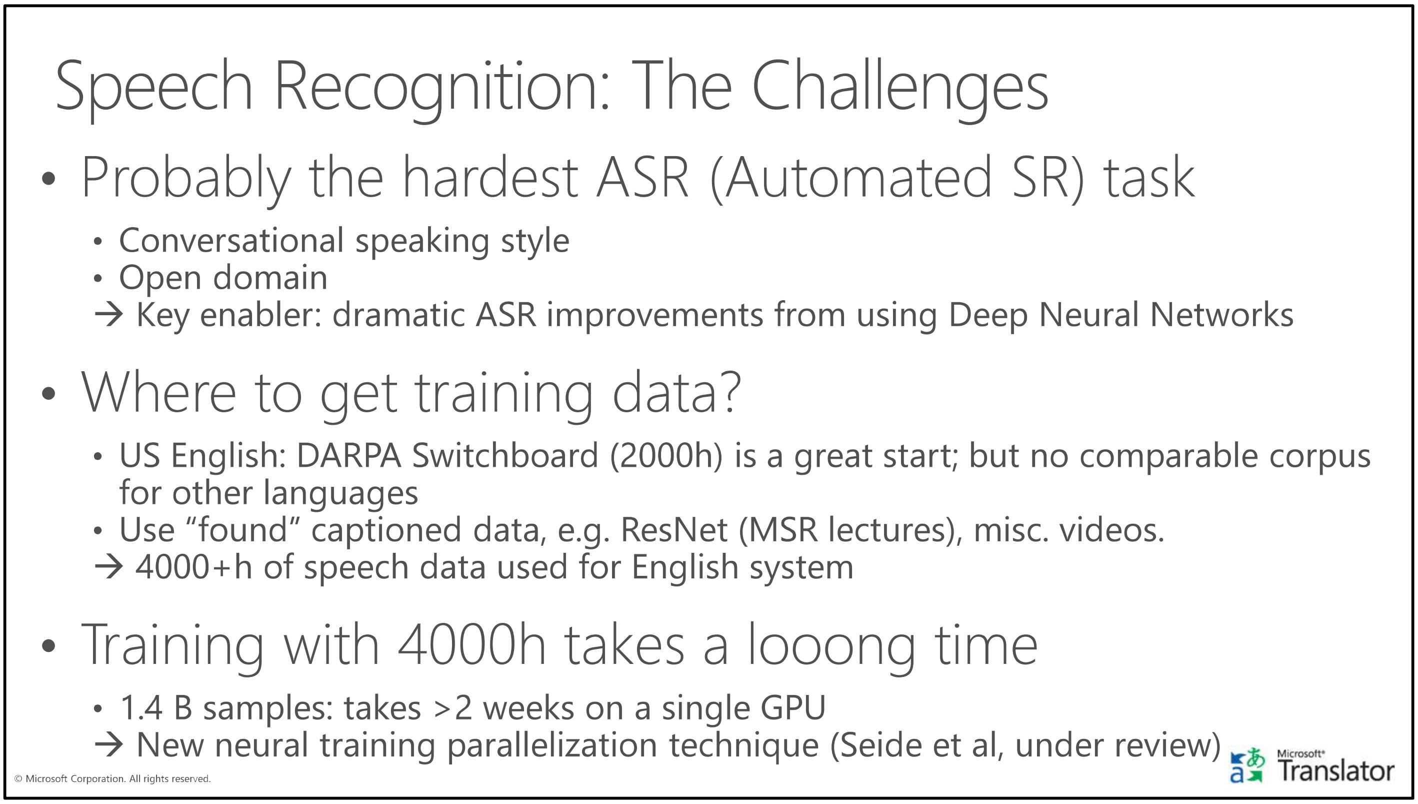
Confluence of Factors:

1. Technological Leap in speech recognition
   - Deep Learning (DNNs) – 33+% WER reduction over Gaussian Models (GMMs) (Seide et al 2011)
     - Now above 50%
   - More robust to noise, speaker variation, accents
   - Note recent research from MSR at achieving human parity for speech reco (Xiong et al 2016)

2. Huge amounts of data
   - 1,000’s of hours of transcribed data

3. Robust and fast training infrastructure
   - Large GPU (Graphical Processing Unit) Farms
   - Training times reduced from months to days

4. For translation:
   - Neural MT (Devlin et al 2014)
   - Reworked neural models faster than equivalent SMT (and deployable to CPU) (Devlin 2017)



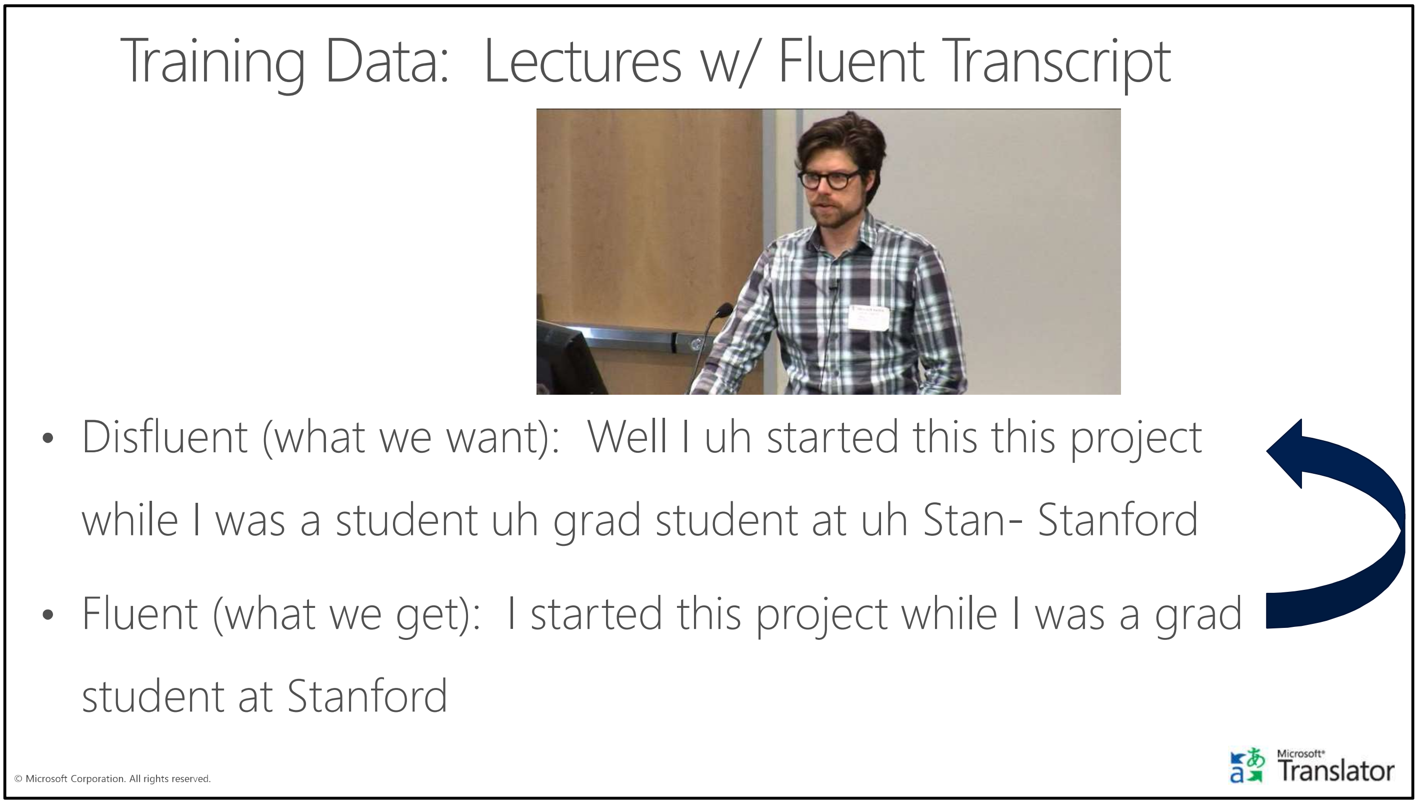
Speech Recognition: The Challenges

- Probably the hardest ASR (Automated SR) task
  - Conversational speaking style
  - Open domain

  → Key enabler: dramatic ASR improvements from using Deep Neural Networks
- Where to get training data?
  - US English: DARPA Switchboard (2000h) is a great start; but no comparable corpus for other languages
  - Use “found” captioned data, e.g. ResNet (MSR lectures), misc. videos.

  → 4000+h of speech data used for English system
- Training with 4000h takes a looong time
  - 1.4 B samples: takes >2 weeks on a single GPU

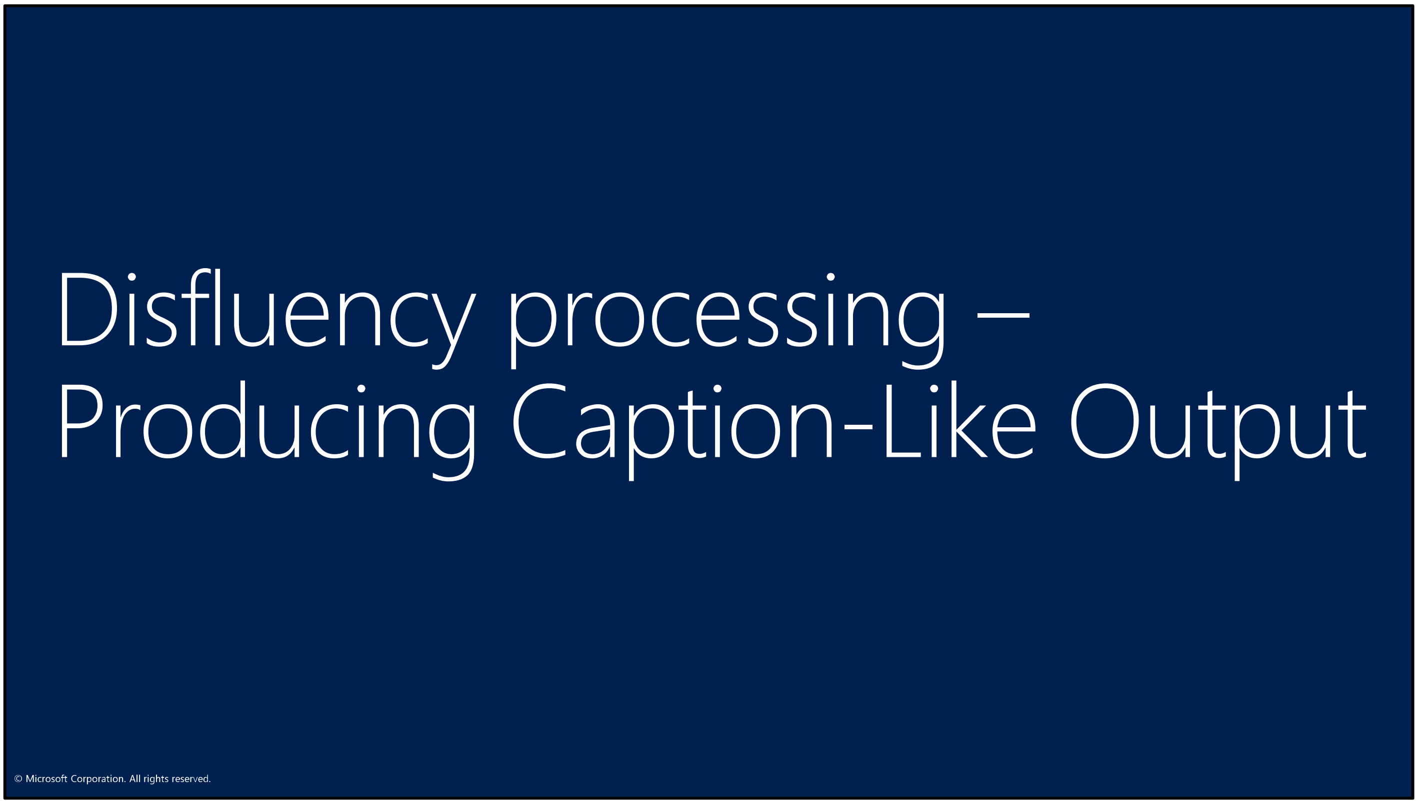
  → New neural training parallelization technique (Seide et al, under review)



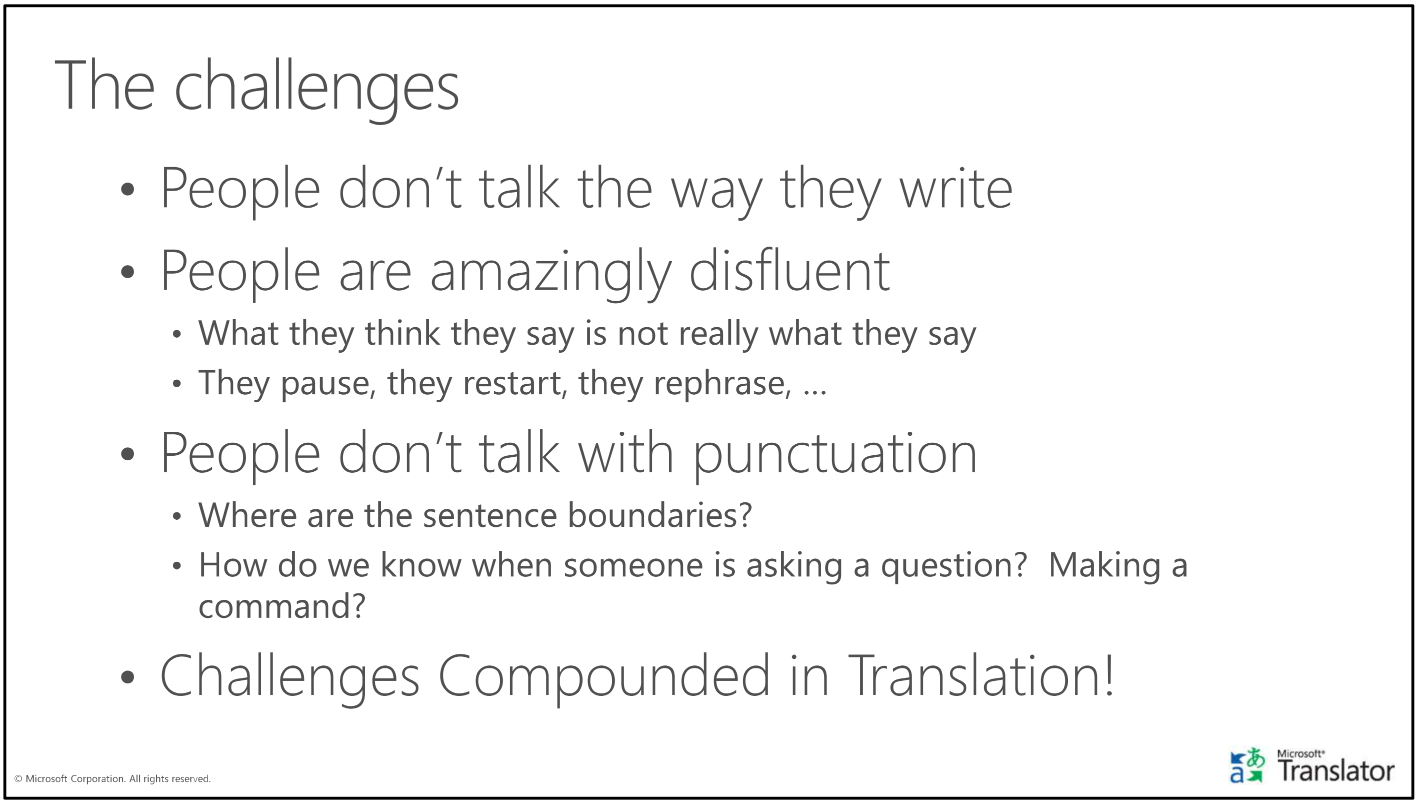
Training Data: Lectures w/ Fluent Transcript

- Disfluent (what we want): Well I uh started this this project while I was a student uh grad student at uh Stan- Stanford
- Fluent (what we get): I started this project while I was a grad student at Stanford

We have a huge amount of lecture data at MSR, what we call ResNet. Tens of thousands of hours of transcribed speech content, across many different speakers, voices and accents. The problem is that the transcriptions aren't right for training speech recognition.

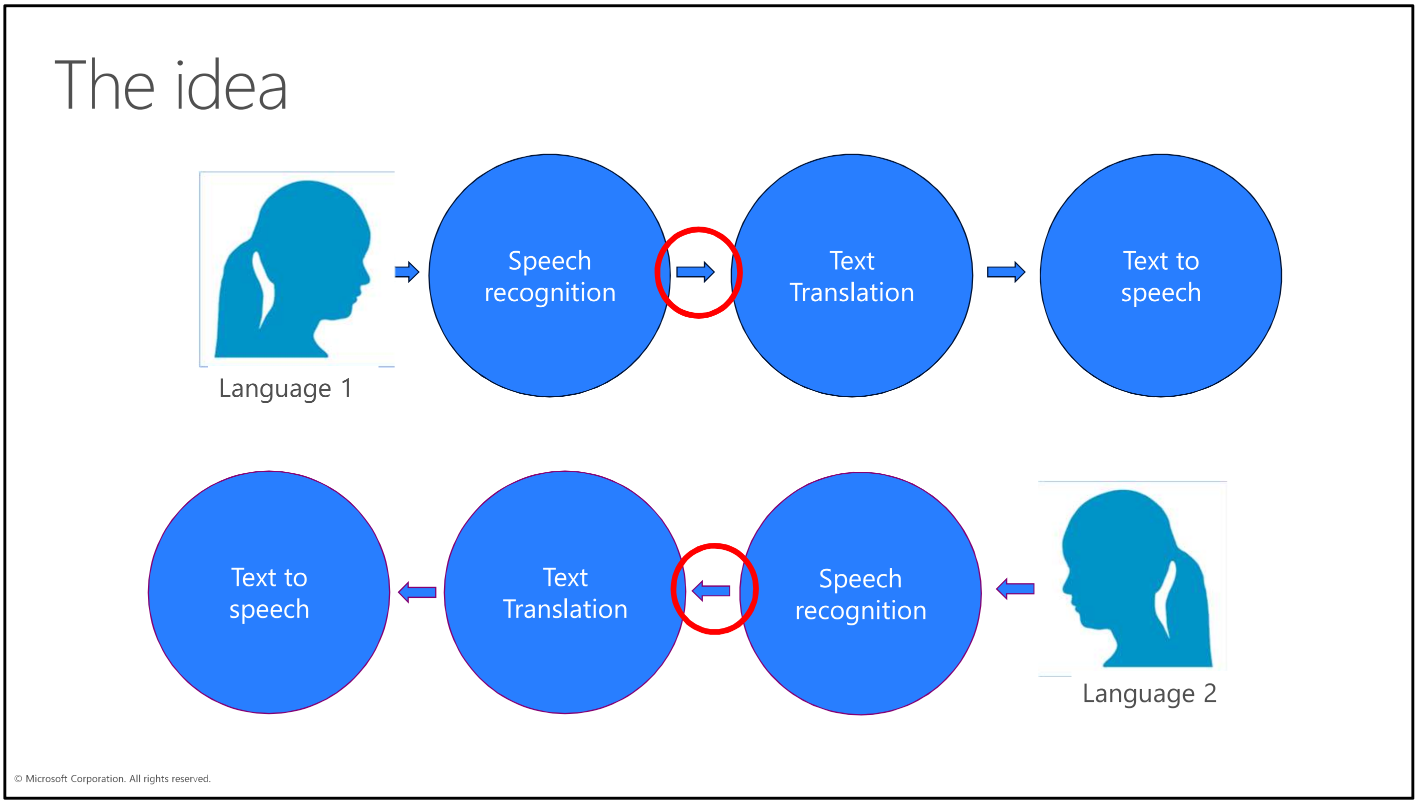


Disfluency processing – Producing Caption-Like Output



The challenges

- People don’t talk the way they write
- People are amazingly disfluent
  - What they think they say is not really what they say
  - They pause, they restart, they rephrase, …
- People don’t talk with punctuation
  - Where are the sentence boundaries?
  - How do we know when someone is asking a question? Making a command?
- Challenges Compounded in Translation!



The idea

Language 1

Speech recognition

Text Translation

Text to speech

Text to speech

Text Translation

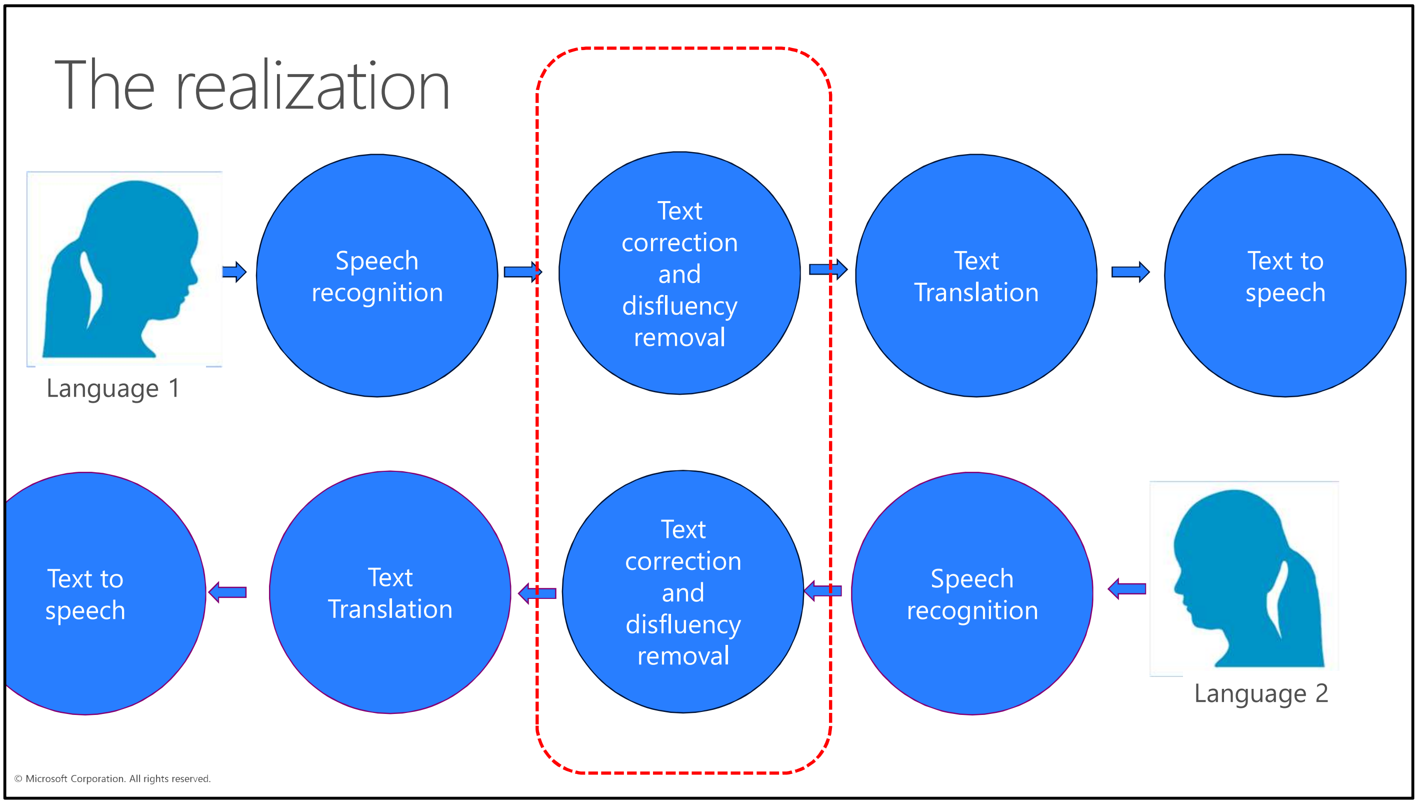
Speech recognition

Language 2

You can see the different components needed for S2S: SR > MT > (TTS)

Animation: **The Mismatch Problem:** The first serious problem we had to deal with was getting the speech recognition to talk with the MT, to generate transcripts that could be translated correctly. But at the same time, if we process the data well, we will generate more fluent, "caption-like" transcripts, which is desirable in speech transcription settings, e.g., live captioning.

The realization



Language 1

Speech recognition

Text correction and disfluency removal

Text Translation

Text to speech

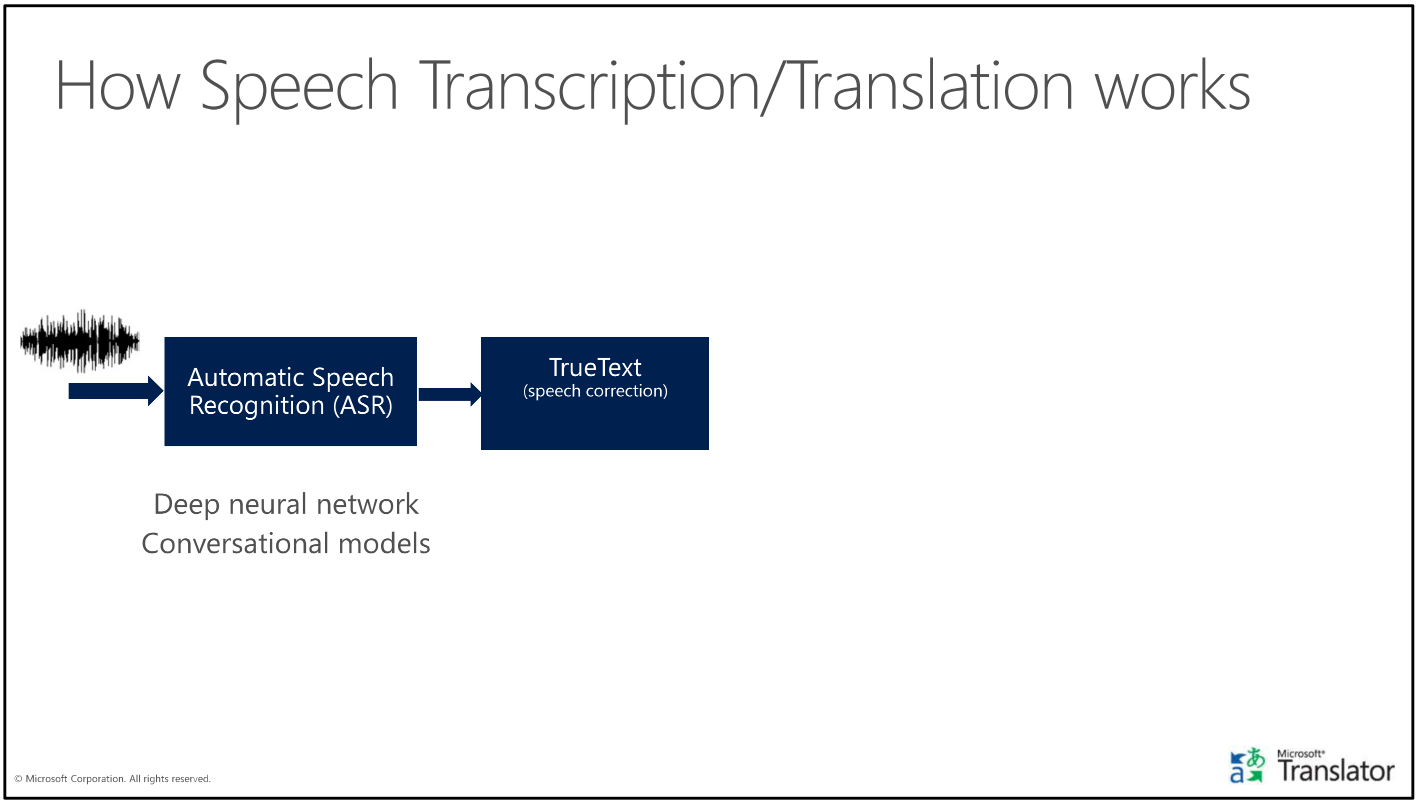
Text to speech

Text Translation

Text correction and disfluency removal

Speech recognition

Language 2



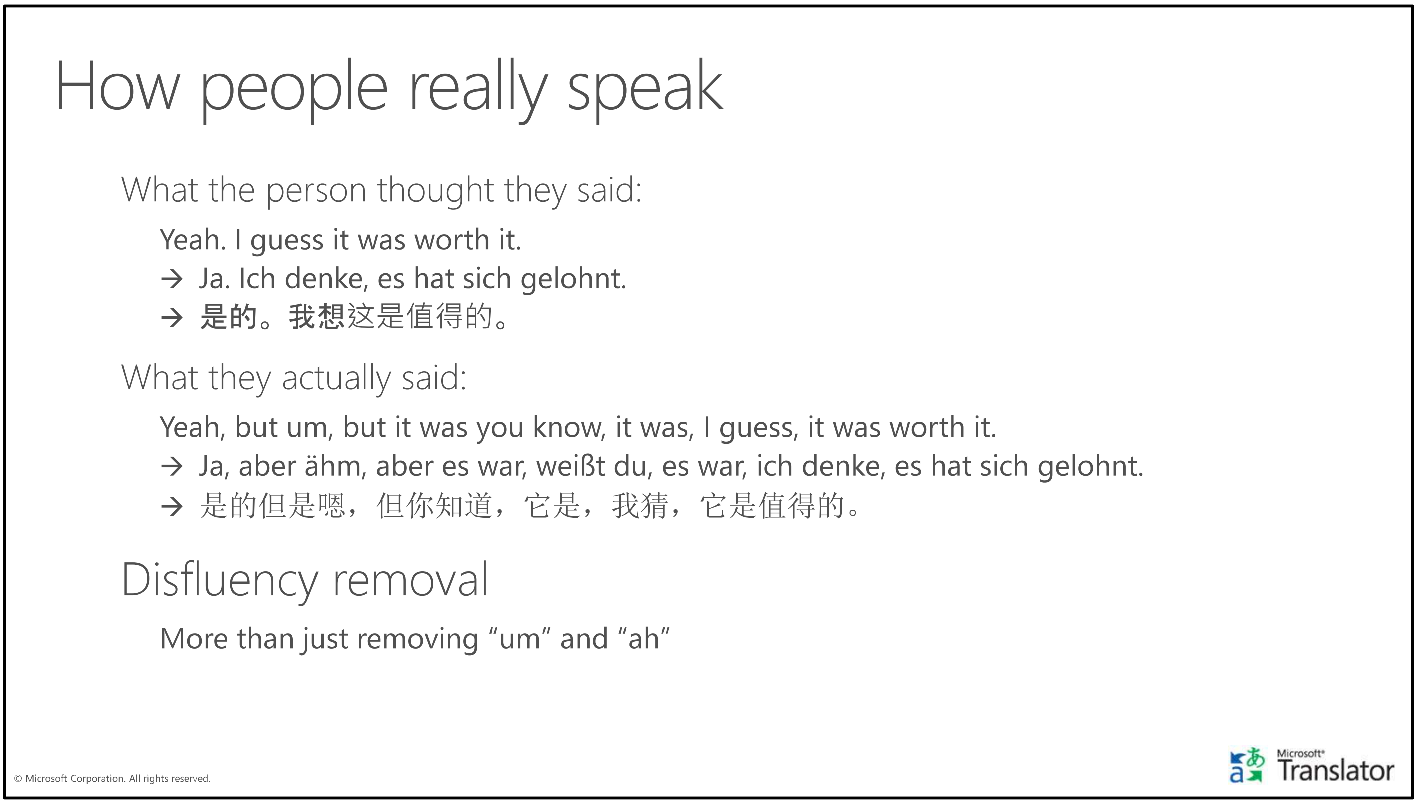
How Speech Transcription/Translation works

Automatic Speech Recognition (ASR)

TrueText
(speech correction)

Deep neural network
Conversational models

TrueText is really crucial to processing the speech signal into something that’s readable and translateable.



How people really speak

What the person thought they said:

Yeah. I guess it was worth it.
→ Ja. Ich denke, es hat sich gelohnt.
→ 是的。我想这是值得的。

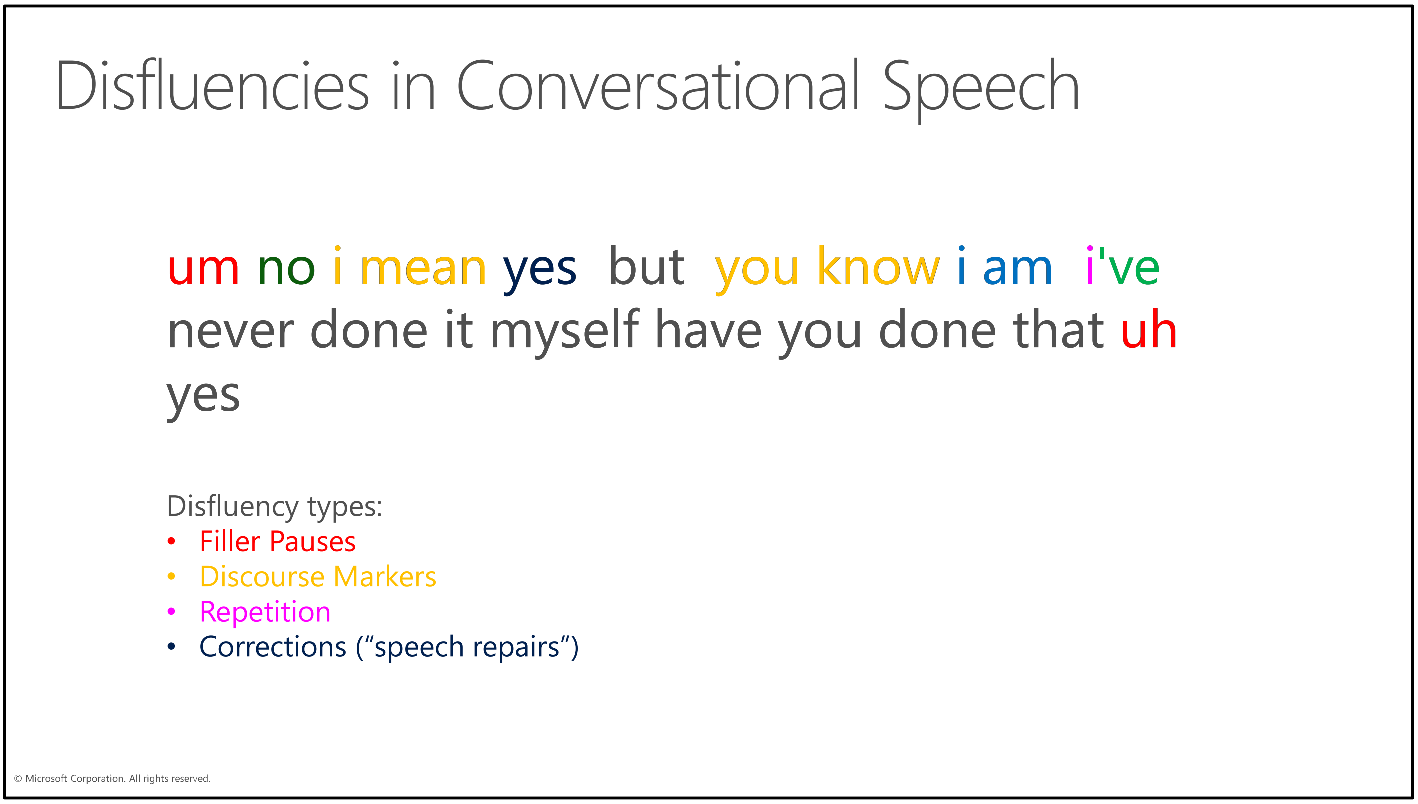
What they actually said:

Yeah, but um, but it was you know, it was, I guess, it was worth it.
→ Ja, aber ähm, aber es war, weißt du, es war, ich denke, es hat sich gelohnt.
→ 是的但是嗯，但你知道，它是，我猜，它是值得的。

Disfluency removal

More than just removing “um” and “ah”

[READ SLIDE] So, if we take the raw ASR output and just throw it at MT, it doesn’t work so well. We need components to process the ASR, remove disfluencies, etc. and make it more palatable to MT. Likewise we need to adapt MT to handle this kind of input. Crucially, we make the transcripts more fluent, more readable, and more usable, both for MT, but also for reading by humans (e.g., in DHH scenarios).



Disfluencies in Conversational Speech

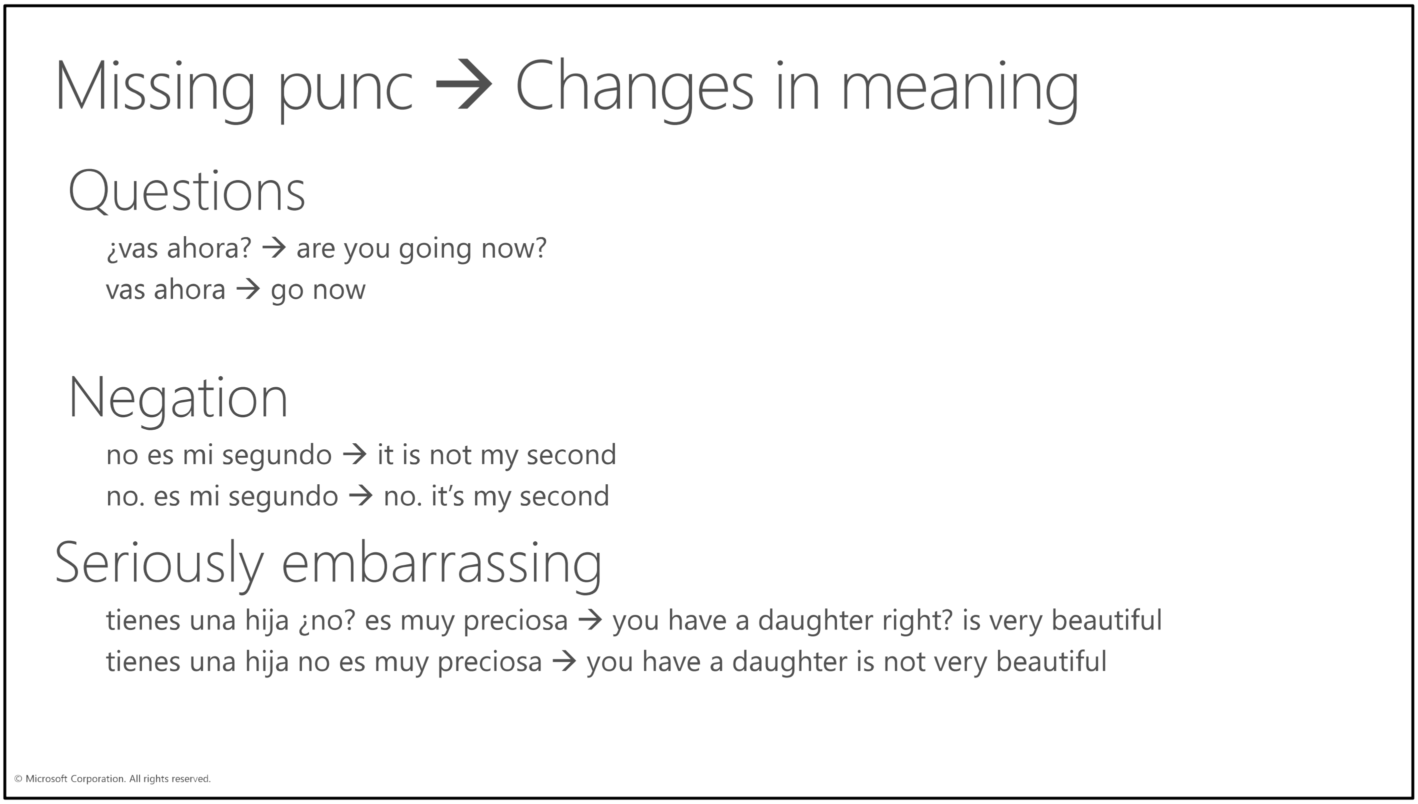
um no i mean yes but you know i am i've never done it myself have you done that uh yes

Disfluency types:

- Filler Pauses
- Discourse Markers
- Repetition
- Corrections (“speech repairs”)

So let’s take a closer look at the different types of disfluencies – first you have uh, your, um, fillers, then, you know, I mean, your discourse markers and and and repetition, and finally correct--, I mean, speech repairs, where people go back and repeat, I mean, change what they said.

In this example, here the speaker changed no to yes, and “I am” to “I have”.



Missing punc → Changes in meaning

Questions

¿vas ahora? → are you going now?

vas ahora → go now

Negation

no es mi segundo → it is not my second

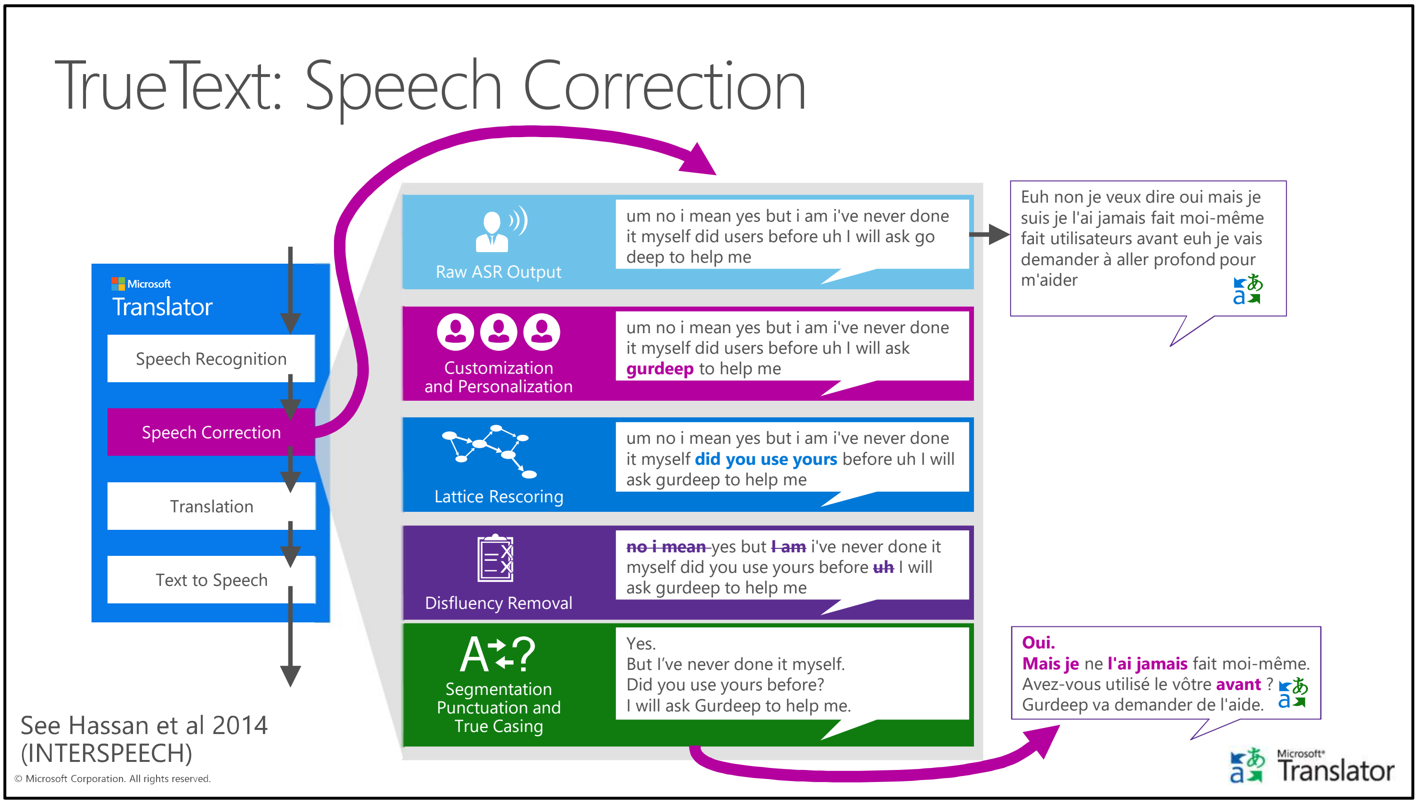
no. es mi segundo → no. it’s my second

Seriously embarrassing

tienes una hija ¿no? es muy preciosa → you have a daughter right? is very beautiful

tienes una hija no es muy preciosa → you have a daughter is not very beautiful

Another thing that is missing is punctuation -- If we don’t get punctuation right, we are risking a lot more than just word salad. You may know the example “Let’s eat grandma”, where a missing comma could lead to a tragic outcome. When translating, the problem gets worse.



TrueText: Speech Correction

Microsoft Translator

Speech Recognition

Speech Correction

Translation

Text to Speech

Raw ASR Output

um no i mean yes but i am i've never done it myself did users before uh I will ask go deep to help me

Euh non je veux dire oui mais je suis je l'ai jamais fait moi-même fait utilisateurs avant euh je vais demander à aller profond pour m'aider

Customization and Personalization

um no i mean yes but i am i've never done it myself did users before uh I will ask **gurdeep** to help me

Lattice Rescoring

um no i mean yes but i am i've never done it myself **did you use yours** before uh I will ask gurdeep to help me

Disfluency Removal

~~**no i mean**~~ yes but ~~**I am**~~ i've never done it myself did you use yours before ~~**uh**~~ I will ask gurdeep to help me

Segmentation Punctuation and True Casing

Yes.
But I've never done it myself.
Did you use yours before?
I will ask Gurdeep to help me.

Oui.
Mais je ne **l'ai jamais** fait moi-même.
Avez-vous utilisé le vôtre **avant** ?
Gurdeep va demander de l'aide.

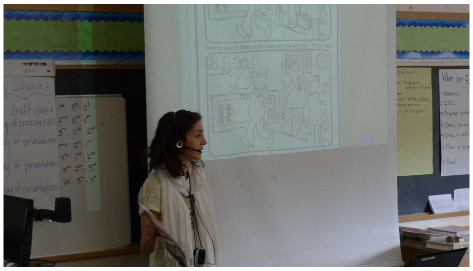
See Hassan et al 2014 (INTERSPEECH)

Microsoft Translator

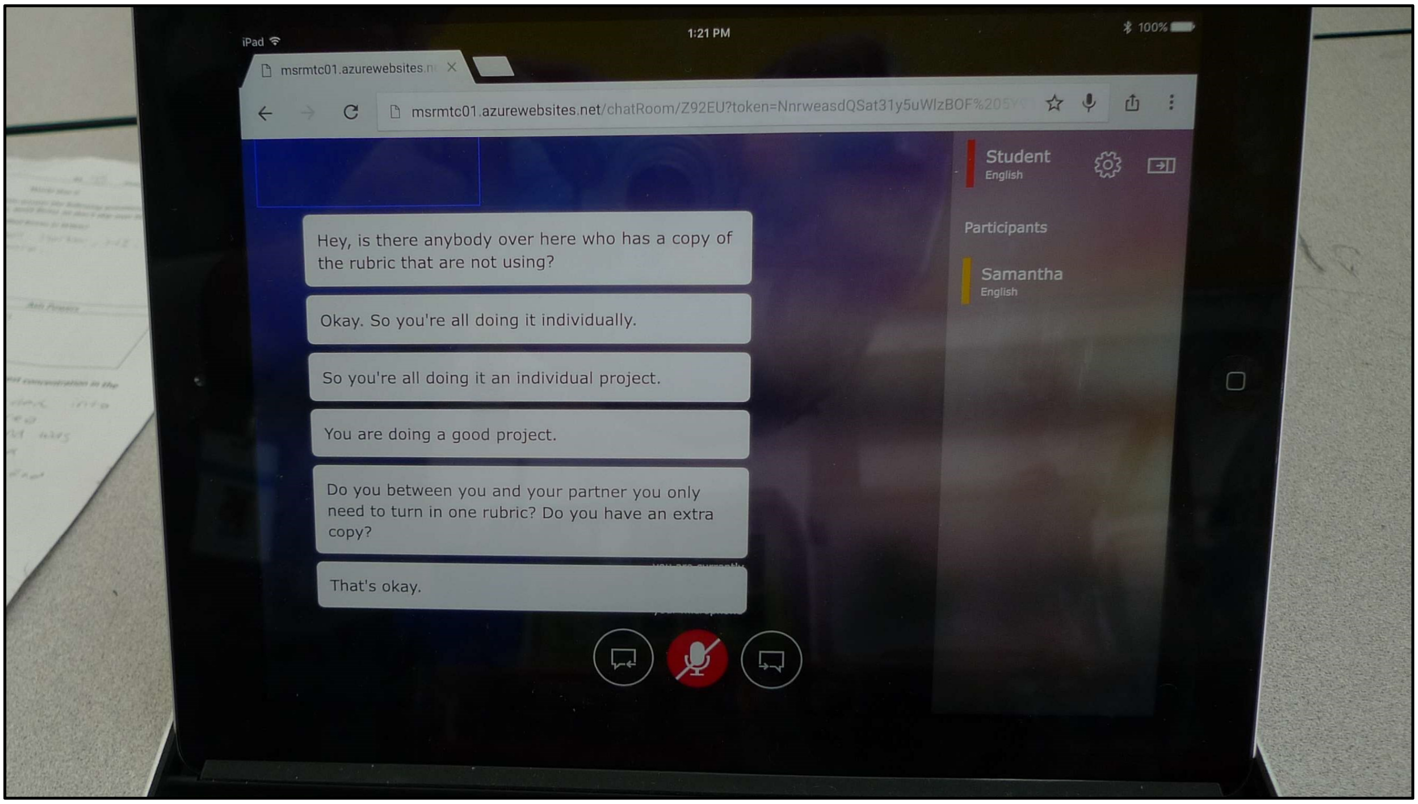
On your left, you see the pipeline for Speech Translation: Speech Recognition, Speech Correction (what we call TrueText), Machine Translation, and Text to Speech (if desire). Here we explore in-depth the means by which we process speech transcripts before handing them off to translation.

Transcription and Translation in the Classroom

Classroom setup



1. Teacher starts a “translation room” using her PC / laptop
2. Teacher “mics up” with a wireless headset
3. Teacher projects or shares conversation code with class
4. Students join room using the shared room code on their smart phones or school-provided computers
5. Each student can choose language of choice (for English-language-learners)
6. Student can save transcript at the end of the conversation



Hey, is there anybody over here who has a copy of the rubric that are not using?

Okay. So you're all doing it individually.

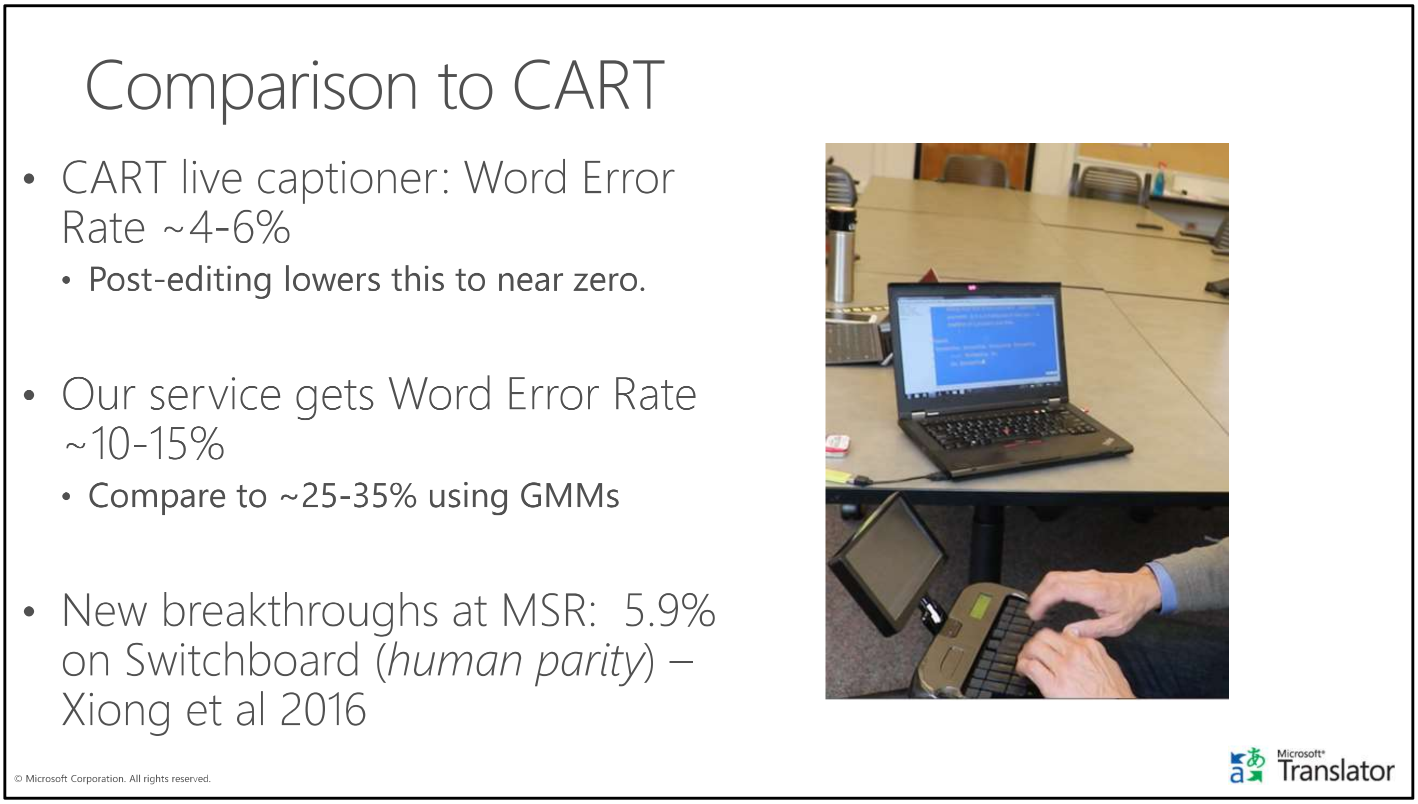
So you're all doing it an individual project.

You are doing a good project.

Do you between you and your partner you only need to turn in one rubric? Do you have an extra copy?

That's okay.

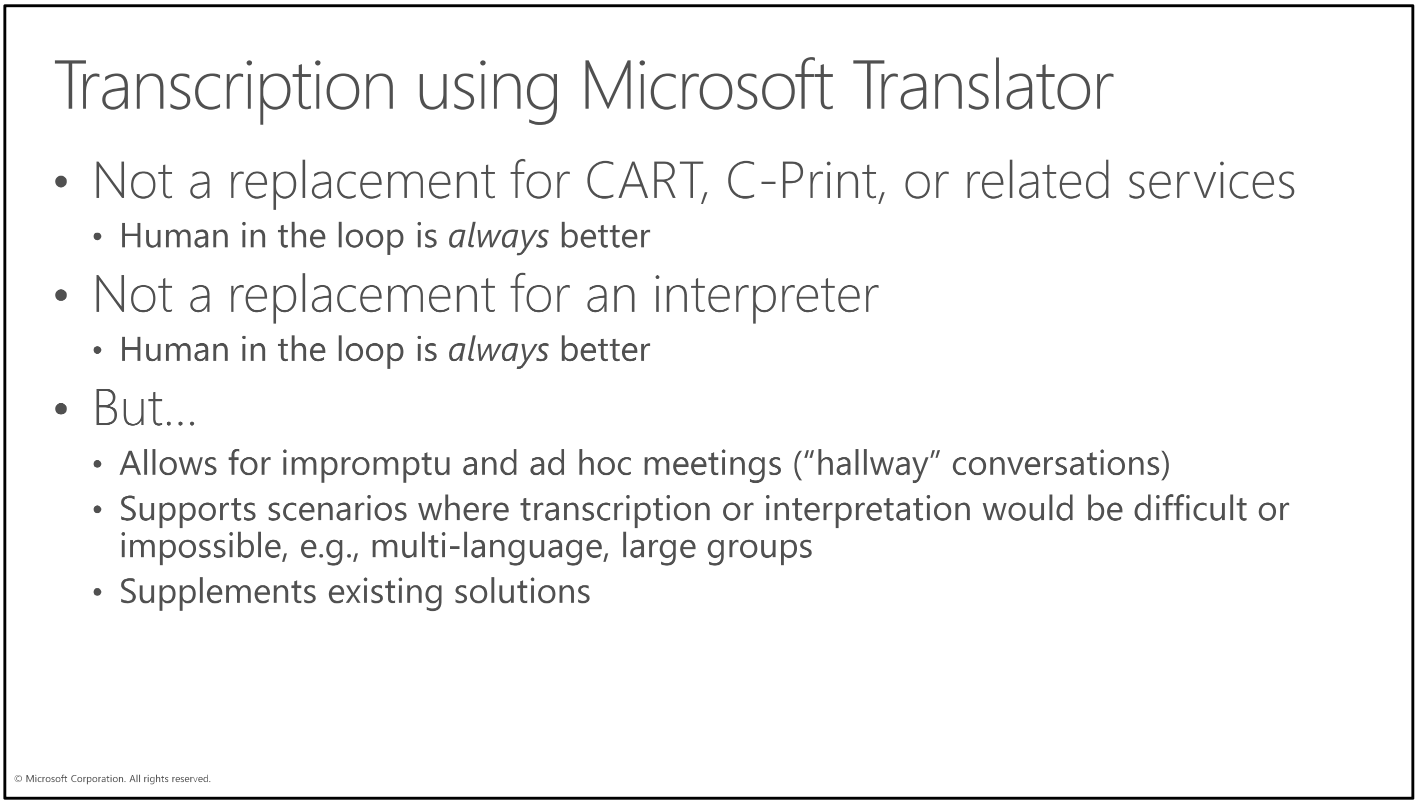
Multilingual class



Comparison to CART

- CART live captioner: Word Error Rate ~4-6%
  - Post-editing lowers this to near zero.
- Our service gets Word Error Rate ~10-15%
  - Compare to ~25-35% using GMMs
- New breakthroughs at MSR: 5.9% on Switchboard (*human parity*) – Xiong et al 2016

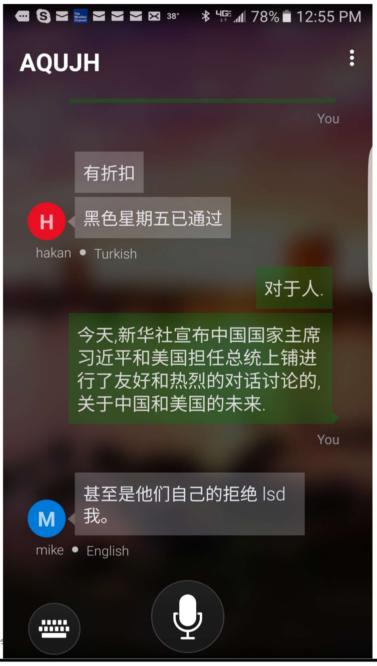
Microsoft Translator



Transcription using Microsoft Translator

- Not a replacement for CART, C-Print, or related services
  - Human in the loop is *always* better
- Not a replacement for an interpreter
  - Human in the loop is *always* better
- But…
  - Allows for impromptu and ad hoc meetings ("hallway" conversations)
  - Supports scenarios where transcription or interpretation would be difficult or impossible, e.g., multi-language, large groups
  - Supplements existing solutions

Examples



You

有折扣

黑色星期五已通过

hakan • Turkish

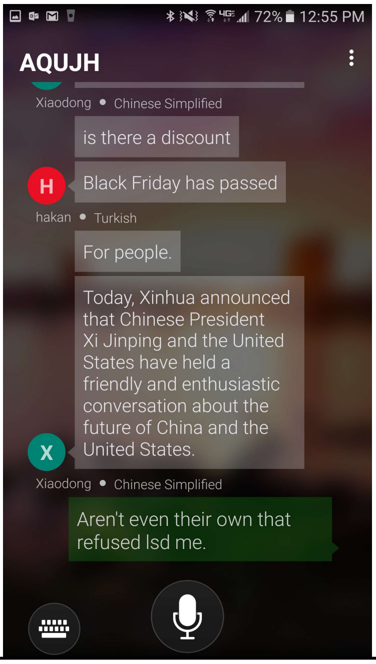
对于人.

今天,新华社宣布中国国家主席习近平和美国担任总统上铺进行了友好和热烈的对话讨论的,关于中国和美国的未来.

You

甚至是他们自己的拒绝 lsd 我。

mike • English



Xiaodong • Chinese Simplified

is there a discount

Black Friday has passed

hakan • Turkish

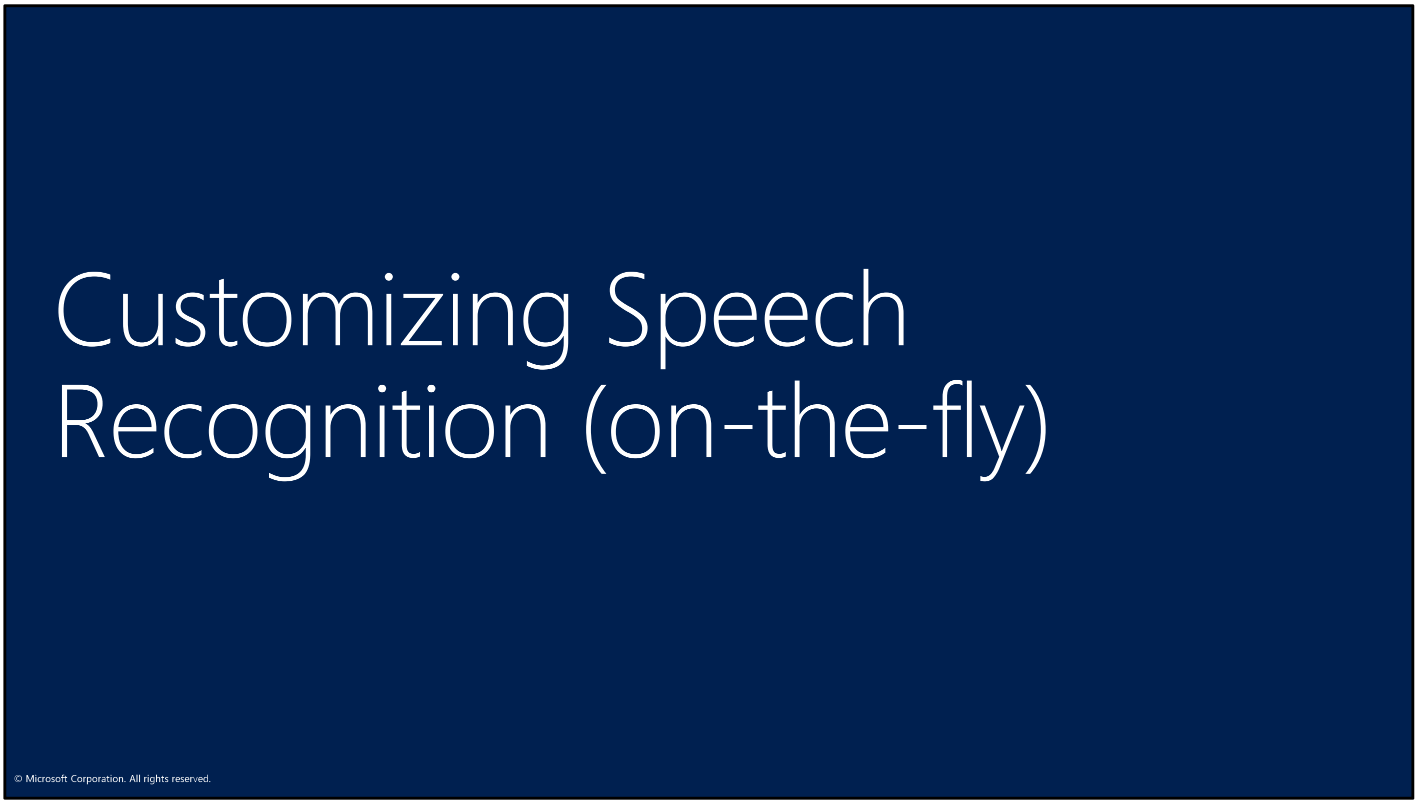
For people.

Today, Xinhua announced that Chinese President Xi Jinping and the United States have held a friendly and enthusiastic conversation about the future of China and the United States.

Xiaodong • Chinese Simplified

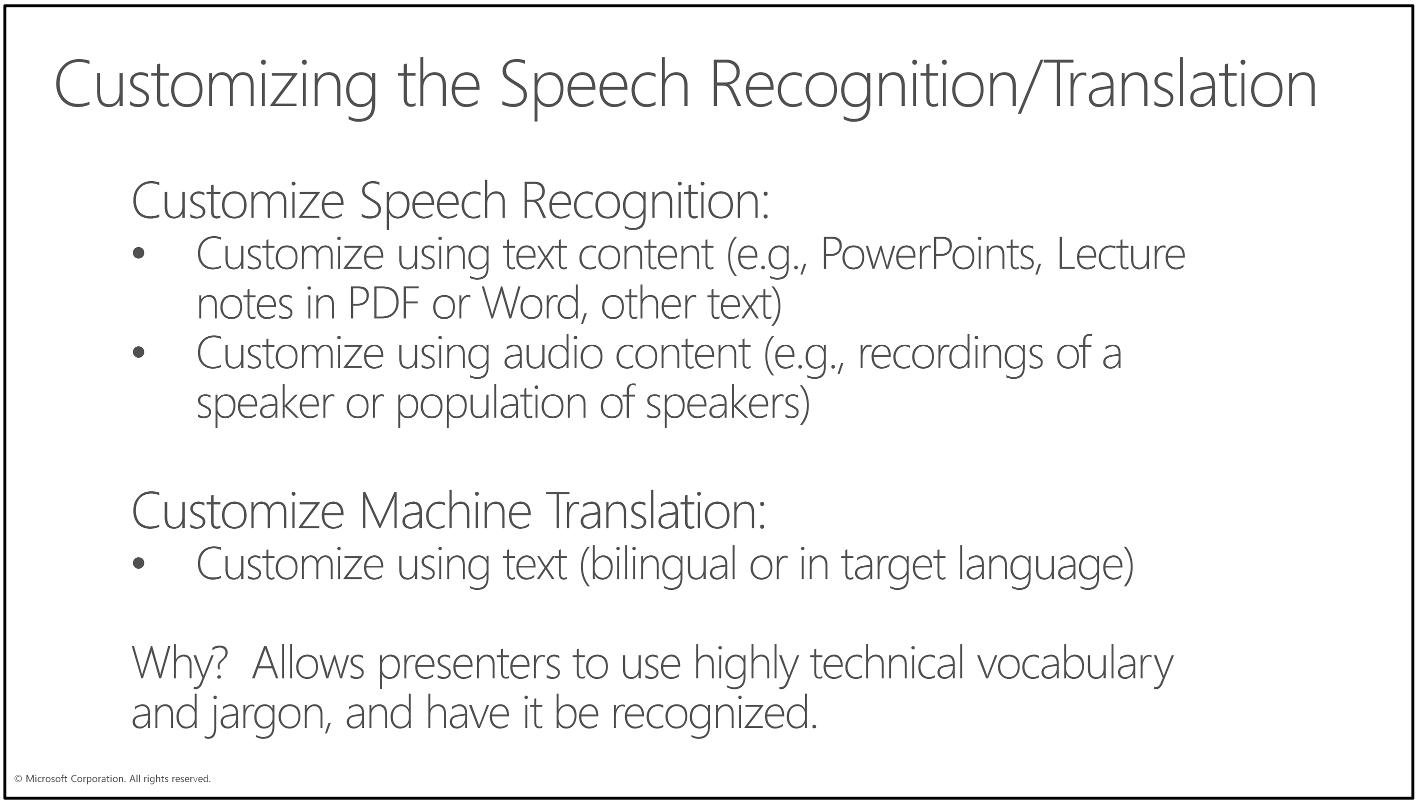
Aren't even their own that refused lsd me.

Release of Microsoft Translator Live, from a hallway conversation between four people last month. Note that each individual who is talking is labeled (so it's clear who said what).



Customizing Speech Recognition (on-the-fly)

Users of our service can customize the speech recognition to their needs, for instance, customizing to specific enterprise or discipline jargon, to a particular usage or dialect, even to a specific speaker’s voice or a population of voices. The text customization has been integrated into the PowerPoint add-in.



Customizing the Speech Recognition/Translation

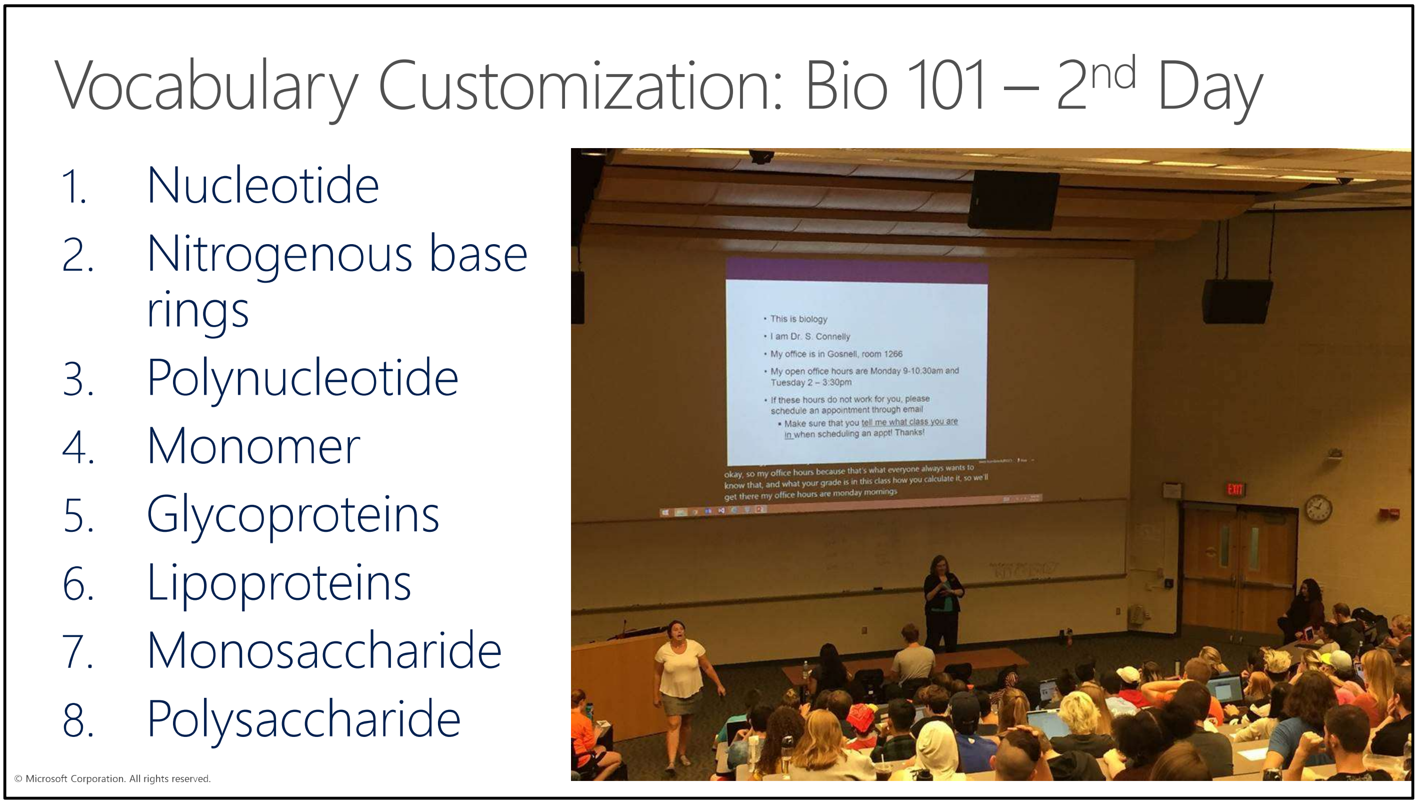
Customize Speech Recognition:

- Customize using text content (e.g., PowerPoints, Lecture notes in PDF or Word, other text)
- Customize using audio content (e.g., recordings of a speaker or population of speakers)

Customize Machine Translation:

- Customize using text (bilingual or in target language)

Why? Allows presenters to use highly technical vocabulary and jargon, and have it be recognized.



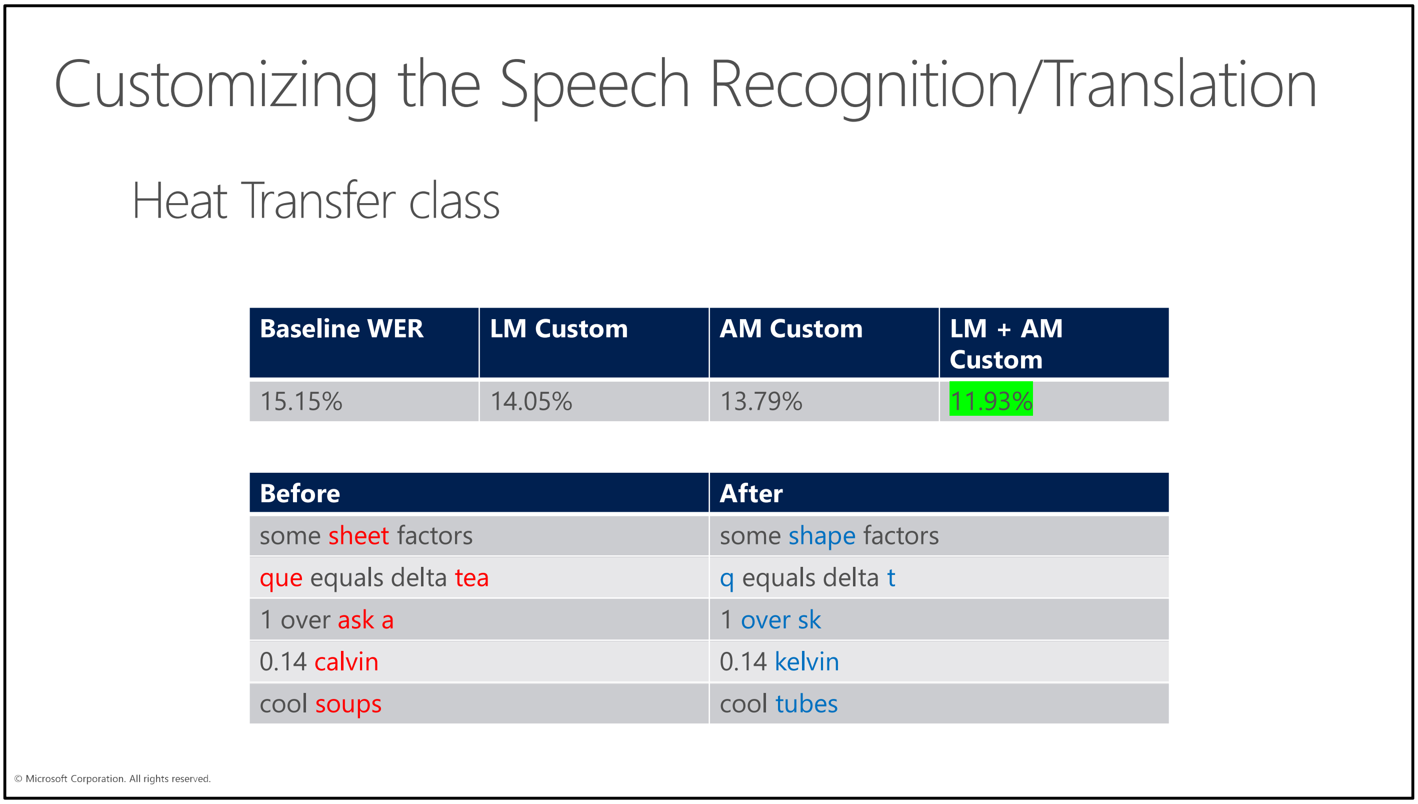
Vocabulary Customization: Bio 101 – 2nd Day

1. Nucleotide
2. Nitrogenous base rings
3. Polynucleotide
4. Monomer
5. Glycoproteins
6. Lipoproteins
7. Monosaccharide
8. Polysaccharide

- This is biology
- I am Dr. S. Connelly
- My office is in Gosnell, room 1266
- My open office hours are Monday 9-10:30am and Tuesday 2 – 3:30pm
- If these hours do not work for you, please schedule an appointment through email
  - Make sure that you tell me what class you are in when scheduling an appt! Thanks!

okay, so my office hours because that's what everyone always wants to know that, and what your grade is in this class how you calculate it, so we'll get there my office hours are monday mornings

1. When we come around to that the basic unit fur nucleic acids is a nucleotide.
2. Nitrogenous base rings actually allow them to form together nicely and make a nice straight chain.
3. And in these structures the basic monomer. It doesn't seem all that exciting. It's a fairly basic.
4. So the polymer is nucleic acids the monomer is a nucleotide.
5. Protein to fat glycoproteins are another one. So carbohydrate in a protein.
6. Okay, so we've already mentioned lipoproteins.
7. Sugars, you've probably heard the terms monosaccharide and polysaccharide before right.



Customizing the Speech Recognition/Translation

Heat Transfer class

Baseline WER	LM Custom	AM Custom	LM + AM Custom
15.15%	14.05%	13.79%	11.93%

Before	After
some sheet factors	some shape factors
que equals delta tea	q equals delta t
1 over ask a	1 over sk
0.14 calvin	0.14 kelvin
cool soups	cool tubes

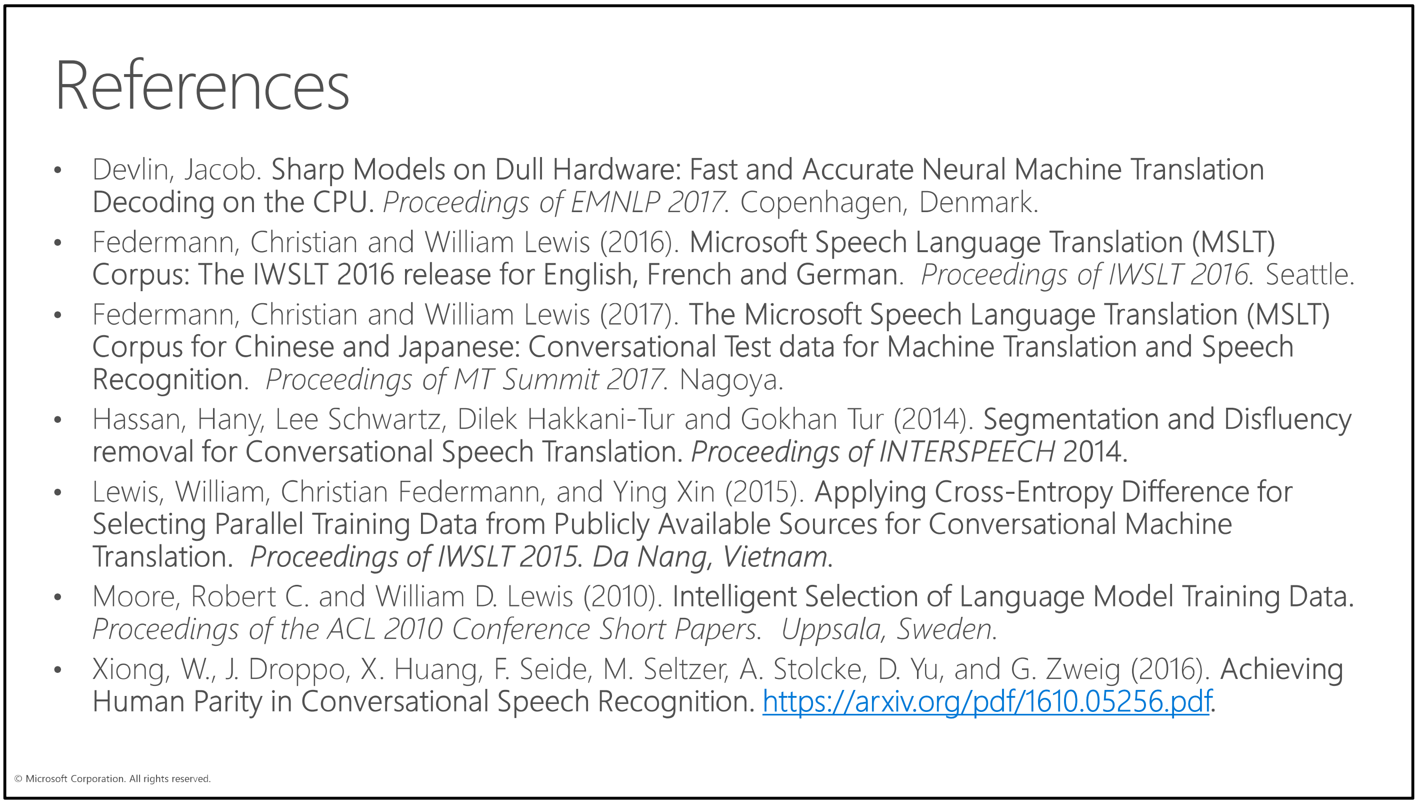
Microsoft Translator Strengths (speech)

- Integrated speech translation solution
  - End to end API: Audio in => Text and audio out
- Optimized for real-time real-life conversations translations
  - ASR expanded to support real life human to human conversations scope
    - Not simple human to machine command-type ASR
  - TrueText technology to handle the way people talk in real life
  - Expanded speech translation models
    - Handles spoken text translations as well as written text ones
- Deployed at scale on millions of users since Dec 2014
  - Skype Translator, Translator apps, 3rd party apps/solutions, etc.

Appendices

References

Making MT more resilient to noisy transcribed input.
Part of the equation is training on data that is “in-domain”, rather, “in-style”.
We’ve also found Neural MT to be a better choice in the S2S workflow.

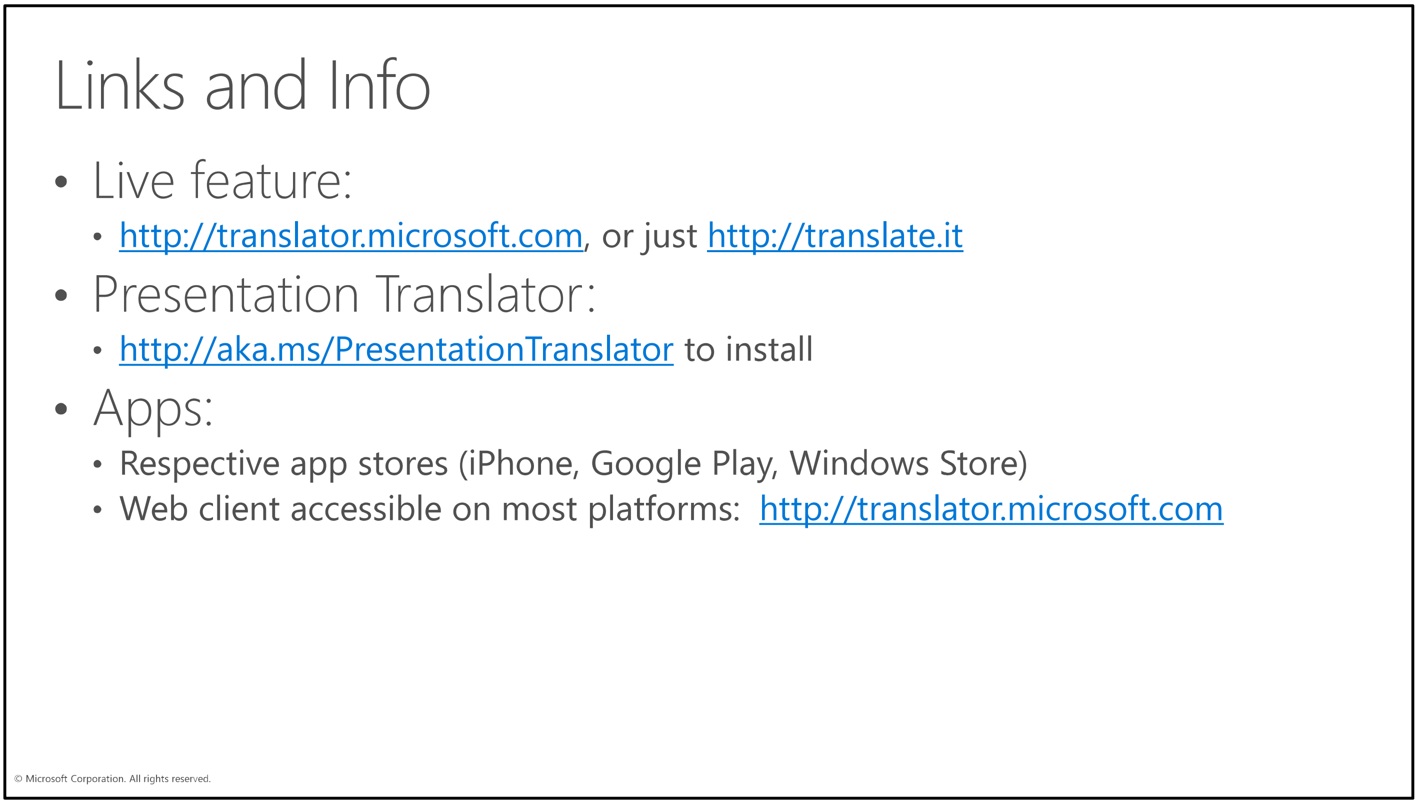


References

- Devlin, Jacob. Sharp Models on Dull Hardware: Fast and Accurate Neural Machine Translation Decoding on the CPU. *Proceedings of EMNLP 2017*. Copenhagen, Denmark.
- Federmann, Christian and William Lewis (2016). Microsoft Speech Language Translation (MSLT) Corpus: The IWSLT 2016 release for English, French and German. *Proceedings of IWSLT 2016*. Seattle.
- Federmann, Christian and William Lewis (2017). The Microsoft Speech Language Translation (MSLT) Corpus for Chinese and Japanese: Conversational Test data for Machine Translation and Speech Recognition. *Proceedings of MT Summit 2017*. Nagoya.
- Hassan, Hany, Lee Schwartz, Dilek Hakkani-Tur and Gokhan Tur (2014). Segmentation and Disfluency removal for Conversational Speech Translation. *Proceedings of INTERSPEECH* 2014.
- Lewis, William, Christian Federmann, and Ying Xin (2015). Applying Cross-Entropy Difference for Selecting Parallel Training Data from Publicly Available Sources for Conversational Machine Translation. *Proceedings of IWSLT 2015. Da Nang, Vietnam.*
- Moore, Robert C. and William D. Lewis (2010). Intelligent Selection of Language Model Training Data. *Proceedings of the ACL 2010 Conference Short Papers. Uppsala, Sweden.*
- Xiong, W., J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig (2016). Achieving Human Parity in Conversational Speech Recognition. https://arxiv.org/pdf/1610.05256.pdf.

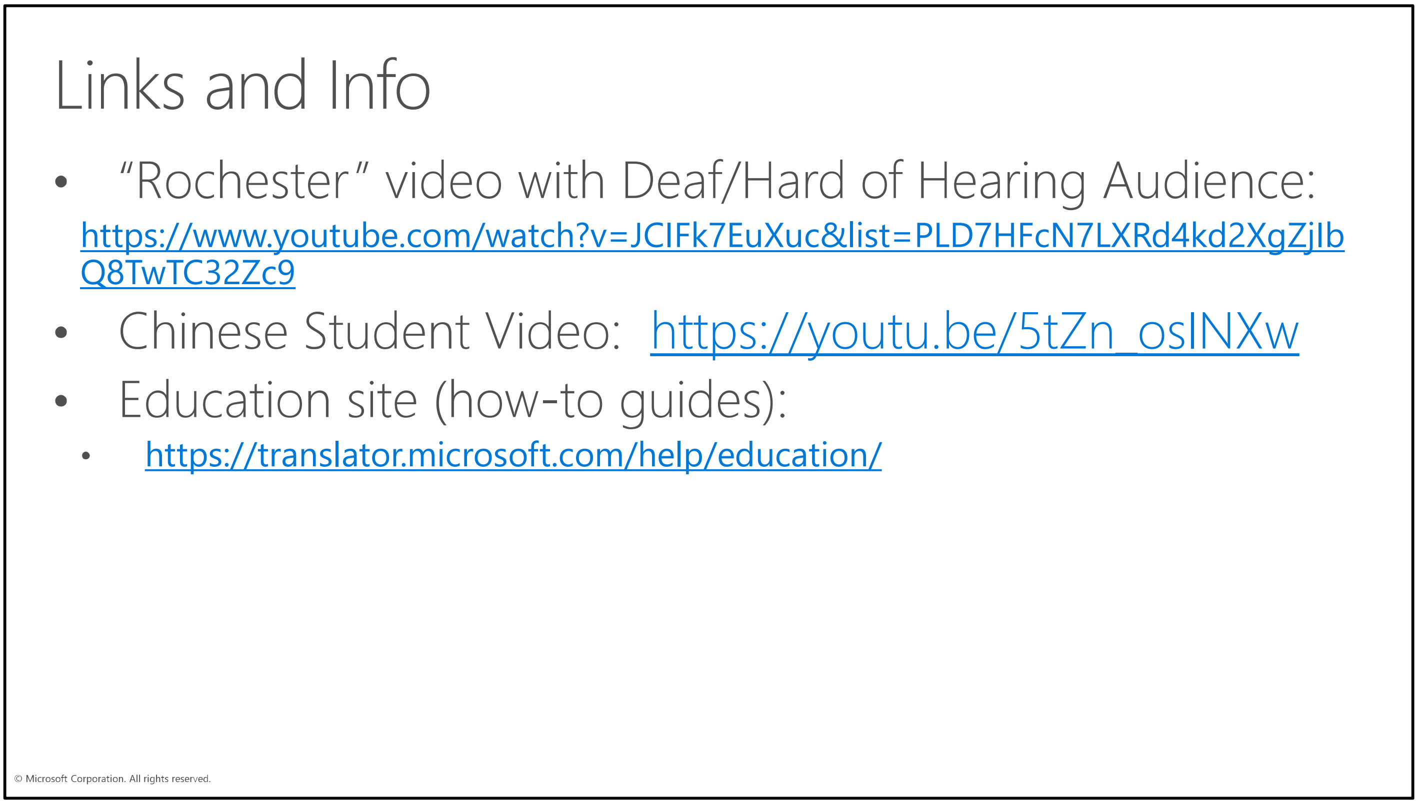
Links and Info

Making MT more resilient to noisy transcribed input.
Part of the equation is training on data that is “in-domain”, rather, “in-style”.
We’ve also found Neural MT to be a better choice in the S2S workflow.



Links and Info

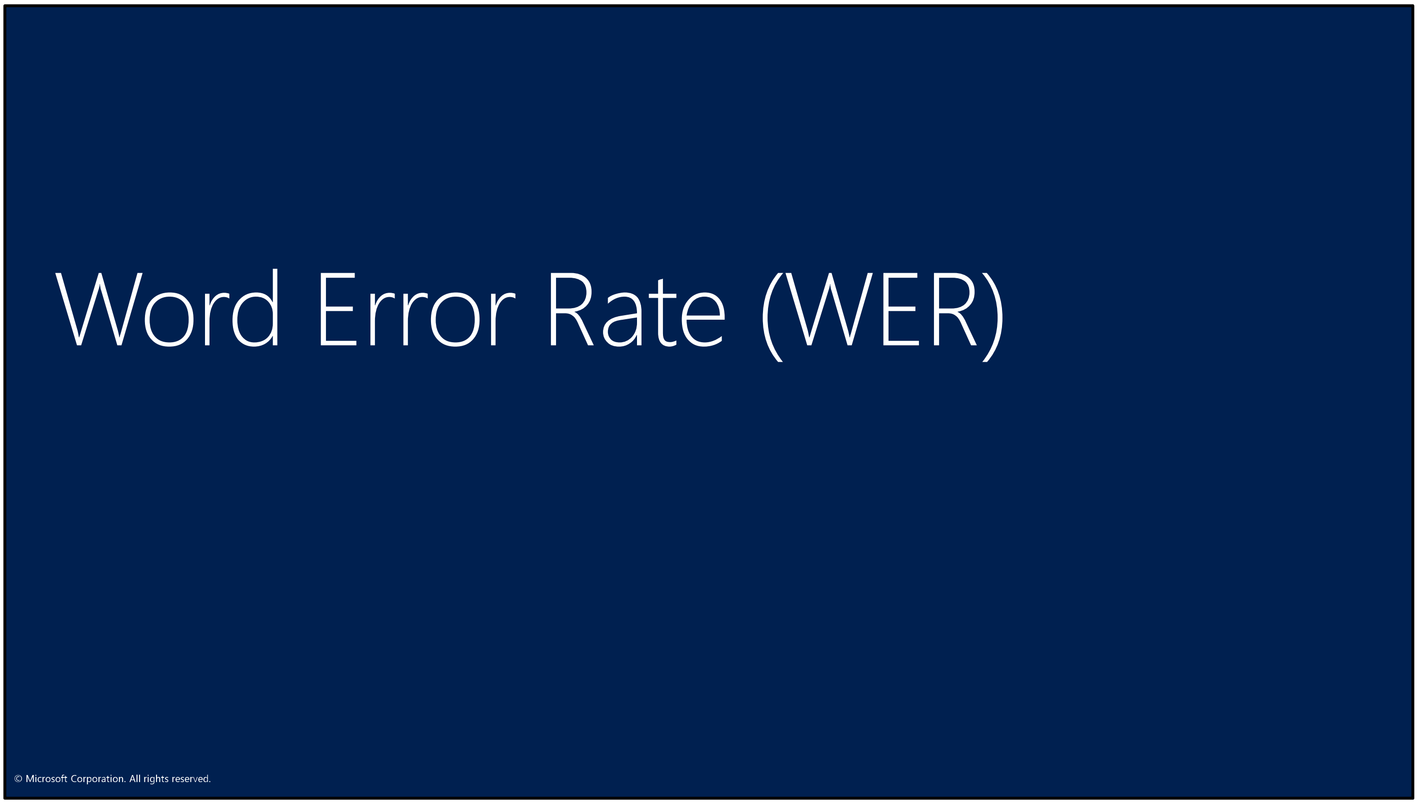
- Live feature:
  - http://translator.microsoft.com, or just http://translate.it
- Presentation Translator:
  - http://aka.ms/PresentationTranslator to install
- Apps:
  - Respective app stores (iPhone, Google Play, Windows Store)
  - Web client accessible on most platforms: http://translator.microsoft.com



Links and Info

- “Rochester” video with Deaf/Hard of Hearing Audience:

  https://www.youtube.com/watch?v=JCIFk7EuXuc&list=PLD7HFcN7LXRd4kd2XgZjIbQ8TwTC32Zc9
- Chinese Student Video: https://youtu.be/5tZn_osINXw
- Education site (how-to guides):
  - https://translator.microsoft.com/help/education/

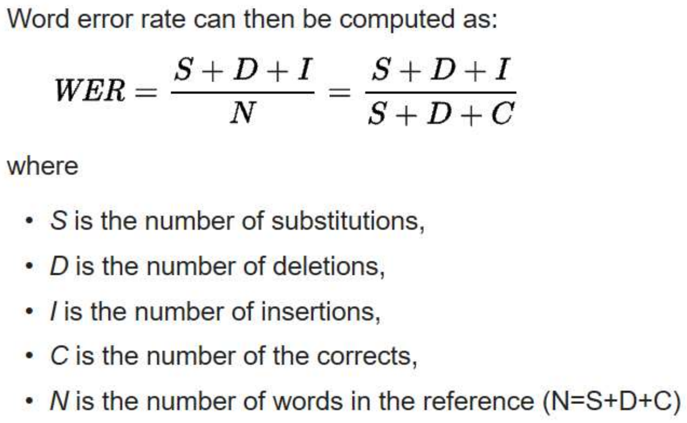


Word Error Rate (WER)

Making MT more resilient to noisy transcribed input.
Part of the equation is training on data that is "in-domain", rather, "in-style".
We've also found Neural MT to be a better choice in the S2S workflow.

Word Error Rate (WER)

Standard metric for measuring performance of speech recognition:



Word error rate can then be computed as:

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

where

- *S* is the number of substitutions,
- *D* is the number of deletions,
- *I* is the number of insertions,
- *C* is the number of the corrects,
- *N* is the number of words in the reference (N=S+D+C)

From Wikipedia: https://en.wikipedia.org/wiki/Word_error_rate

Multi-modality is particularly relevant to speech translation.