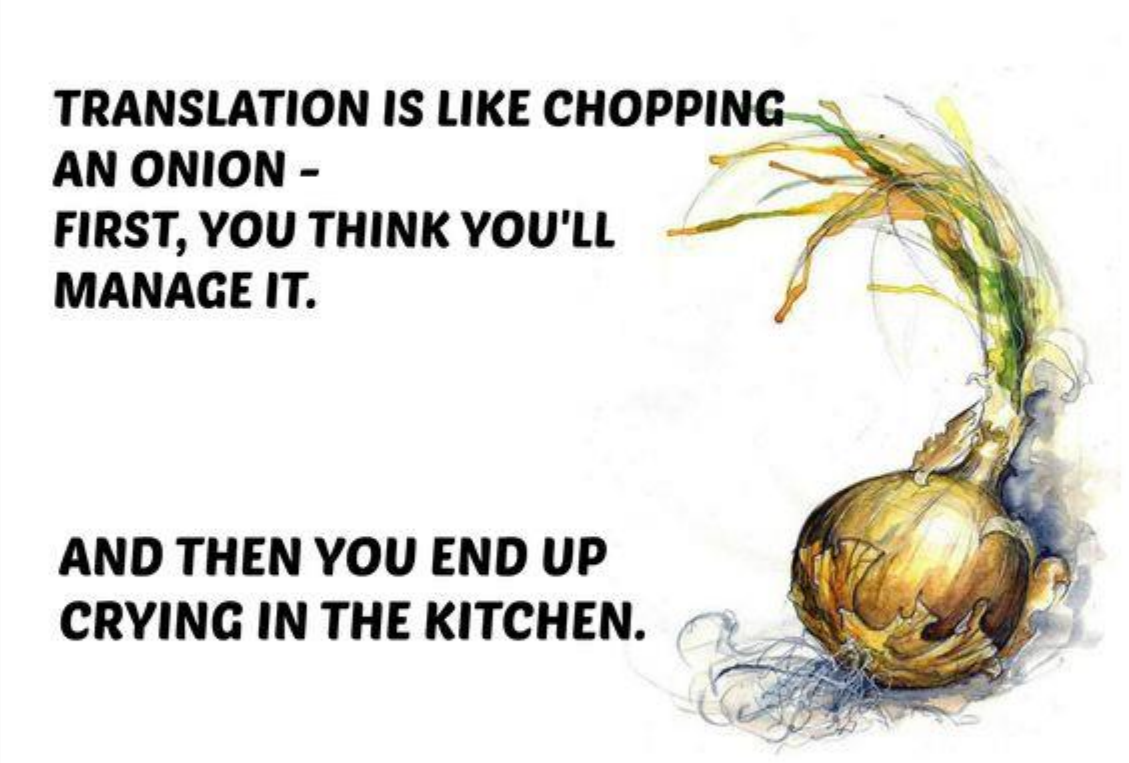


TRANSLATION IS LIKE CHOPPING AN ONION –
FIRST, YOU THINK YOU'LL MANAGE IT.

AND THEN YOU END UP CRYING IN THE KITCHEN.

Source: Vera Hembitskaya

CS 2731
Introduction to Natural Language Processing

Session 20: Machine translation part 1

Michael Miller Yoder

November 6, 2024

University of Pittsburgh | School of Computing and Information

Course logistics: homework

- Homework 4 is **due tomorrow, Thu Nov 7**
  - Part 1: Do part-of-speech tagging manually with the Viterbi algorithm
  - Part 2: Fine-tune BERT-based models for part-of-speech tagging in English and Norwegian
    - Copy and fill in a skeleton Colab notebook

Course logistics: project

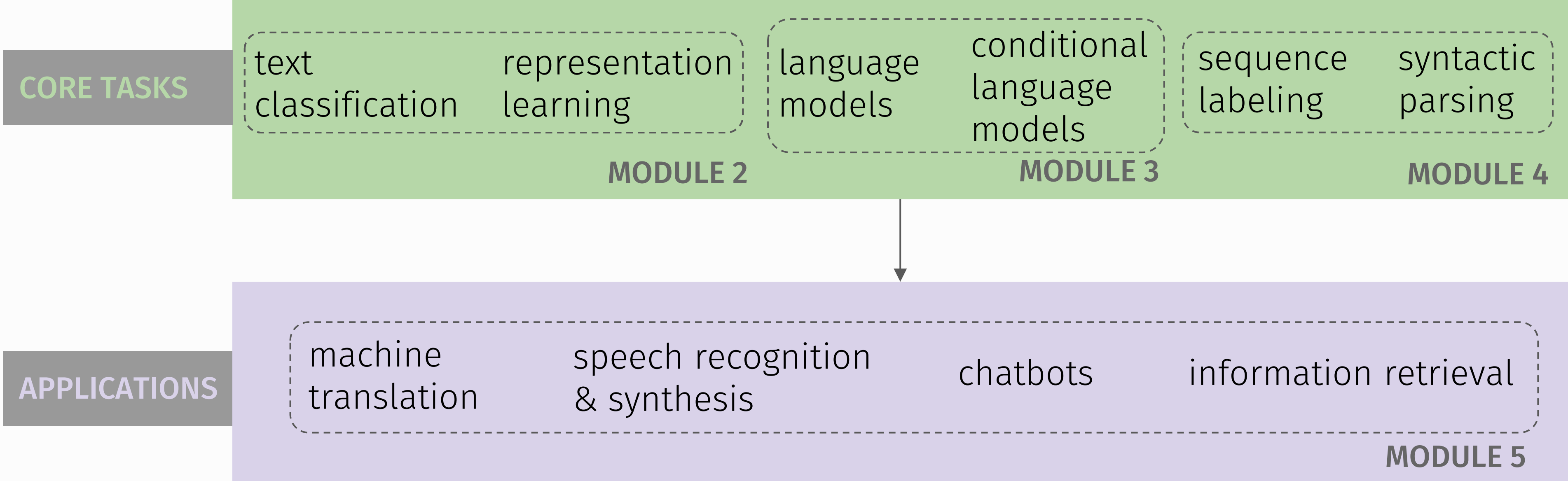
- Project progress report **due next Thu Nov 14**
  - Max 3 pages, ACL format
  - Try to get **something** functional (has input and output, even if the output is not great)
  - See project website for instructions
- Project peer review **due next Thu Nov 14**
  - Will be released today
  - Form where you will review your own and your teammates' contributions so far
  - Will not be used for grading, just for addressing any issues

Muddiest point

What topic or concept was the least clear to you from last lecture?

- Syntax
- Dependency grammar
- Dependency treebanks
- Transition-based dependency parsing
- Evaluation of dependency parsing

Core tasks and applications of NLP



CORE TASKS

text classification

representation learning

MODULE 2

language models

conditional language models

MODULE 3

sequence labeling

syntactic parsing

MODULE 4

APPLICATIONS

machine translation

speech recognition & synthesis

chatbots

information retrieval

MODULE 5

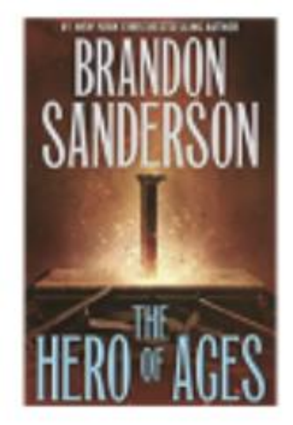
Overview: Machine translation part 1

- History of machine translation (MT)
- Translation in practice
- Exercise: translate some Tajik
- Why is translation difficult?
- Parallel corpora

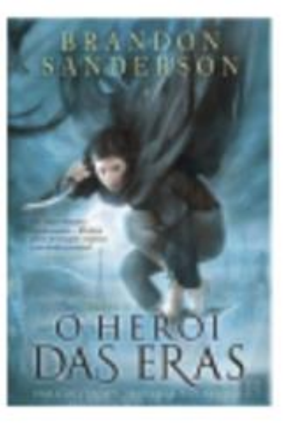
Translation

- Mapping a “text” in a source language to a target language

“I went to the store to buy eggs” → “Eu fui à loja comprar ovos”



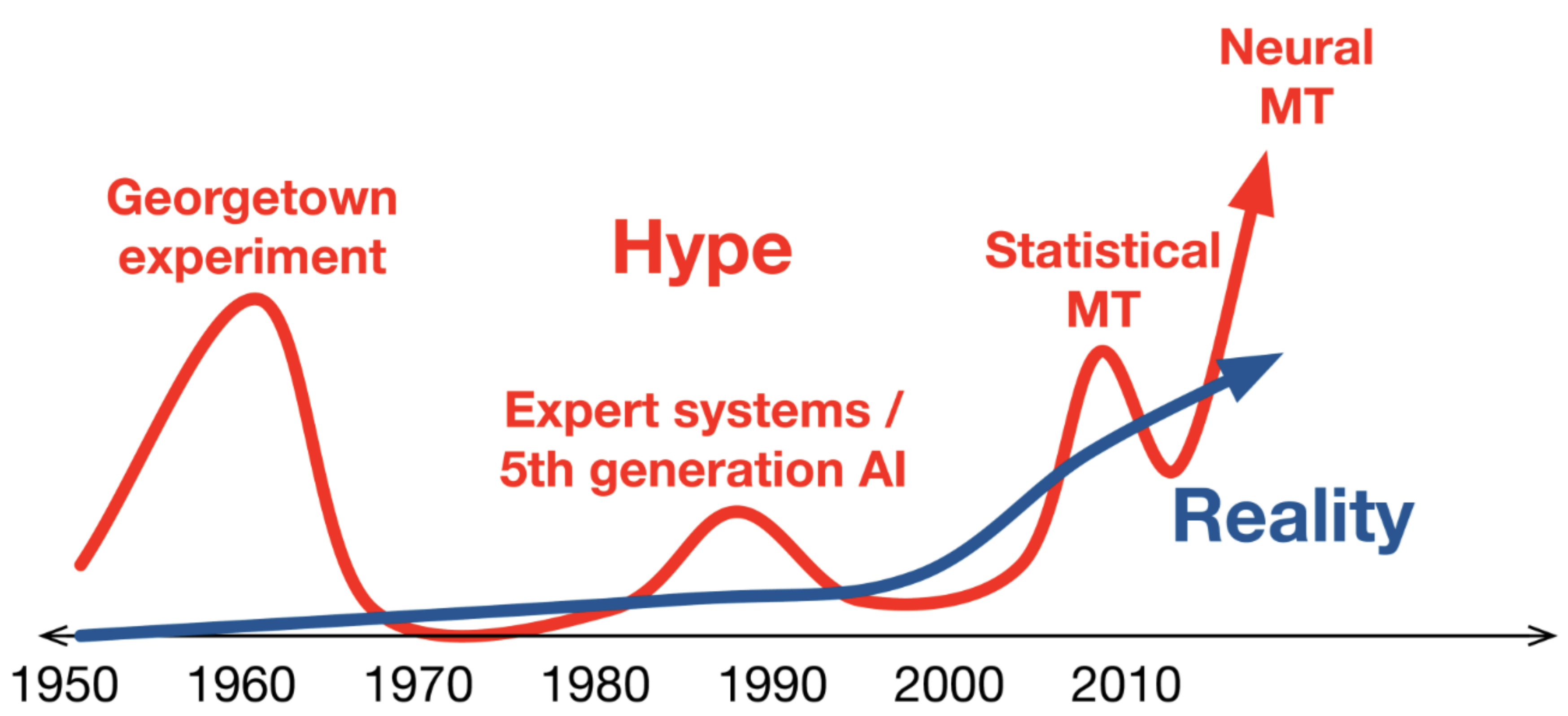
→



Slide adapted from Lori Levin

History of machine translation

MT history: hype vs reality



Slide adapted from Mohit Iyyer, Richard Socher

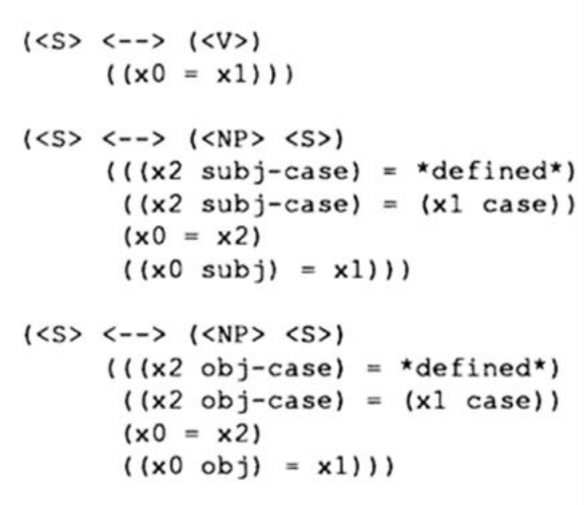
When did people start using computers to translate?

- Roughly 1950s
- Research stopped in the US for about 15-20 years after a 1967 report deemed it impossible
- Research resumed in the US in the early 1980s

Slide adapted from Lori Levin, Chris Manning

What did early MT systems look like?

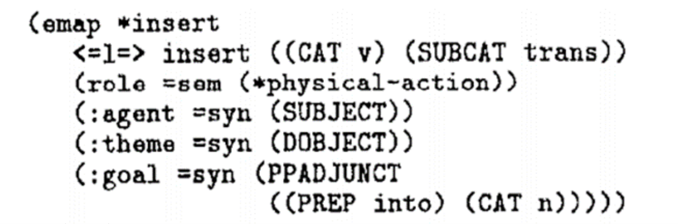
Human linguists wrote elaborate rules involving syntax, semantics, etc



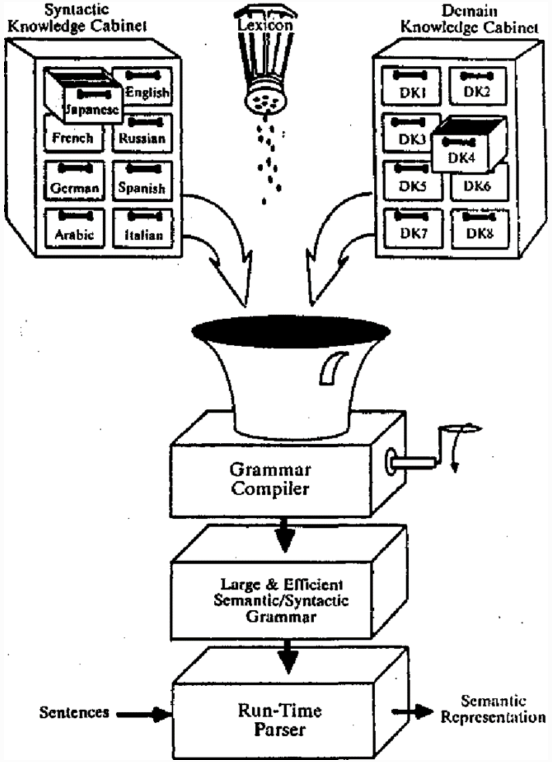
```
(<S> <--> (<V>)
     ((x0 = x1)))

(<S> <--> (<NP> <S>)
     (((x2 subj-case) = *defined*)
      ((x2 subj-case) = (x1 case))
      (x0 = x2)
      ((x0 subj) = x1)))

(<S> <--> (<NP> <S>)
     (((x2 obj-case) = *defined*)
      ((x2 obj-case) = (x1 case))
      (x0 = x2)
      ((x0 obj) = x1)))
```



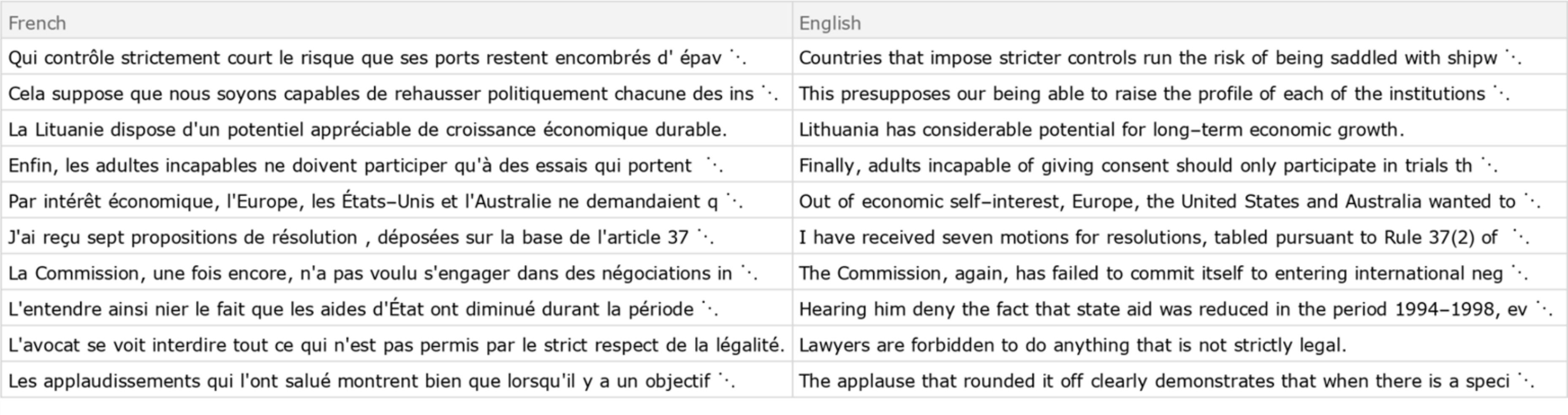
```
(emap *insert
   <=1=> insert ((CAT v) (SUBCAT trans))
   (role =sem (*physical-action))
   (:agent =syn (SUBJECT))
   (:theme =syn (DOBJECT))
   (:goal =syn (PPADJUNCT
                    ((PREP into) (CAT n)))))
```



Slide adapted from Lori Levin

Learning to translate from data

Since the late 1980s, Machine Translation researchers have been using parallel corpora to train Machine Translation systems.

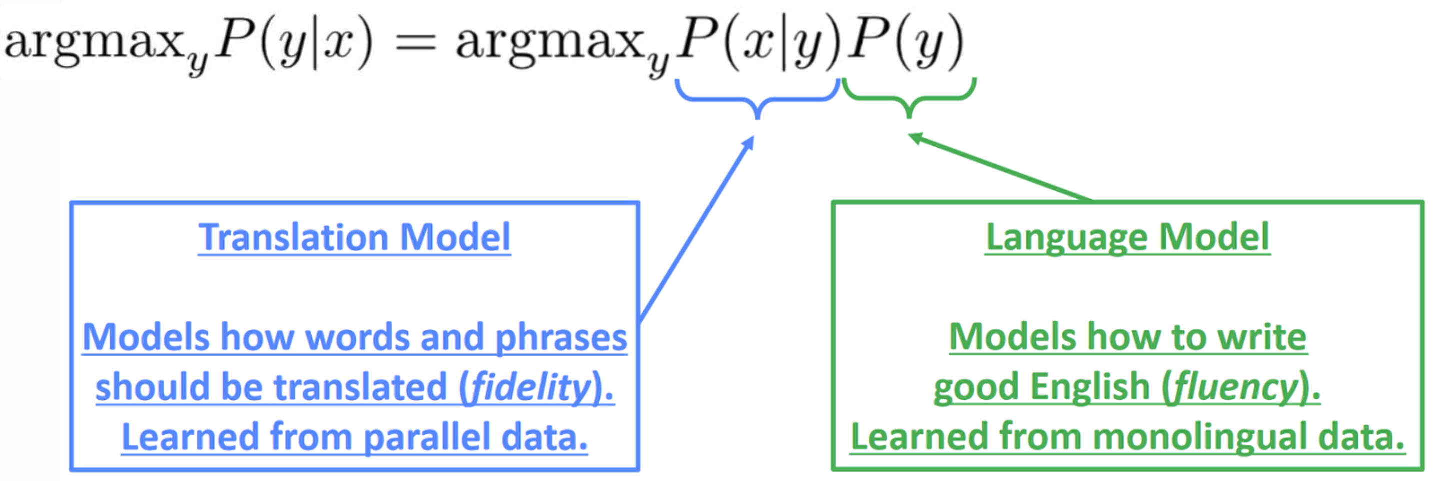


French	English
Qui contrôle strictement court le risque que ses ports restent encombrés d' épav ⋱.	Countries that impose stricter controls run the risk of being saddled with shipw ⋱.
Cela suppose que nous soyons capables de rehausser politiquement chacune des ins ⋱.	This presupposes our being able to raise the profile of each of the institutions ⋱.
La Lituanie dispose d'un potentiel appréciable de croissance économique durable.	Lithuania has considerable potential for long-term economic growth.
Enfin, les adultes incapables ne doivent participer qu'à des essais qui portent ⋱.	Finally, adults incapable of giving consent should only participate in trials th ⋱.
Par intérêt économique, l'Europe, les États-Unis et l'Australie ne demandaient q ⋱.	Out of economic self-interest, Europe, the United States and Australia wanted to ⋱.
J'ai reçu sept propositions de résolution , déposées sur la base de l'article 37 ⋱.	I have received seven motions for resolutions, tabled pursuant to Rule 37(2) of ⋱.
La Commission, une fois encore, n'a pas voulu s'engager dans des négociations in ⋱.	The Commission, again, has failed to commit itself to entering international neg ⋱.
L'entendre ainsi nier le fait que les aides d'État ont diminué durant la période ⋱.	Hearing him deny the fact that state aid was reduced in the period 1994-1998, ev ⋱.
L'avocat se voit interdire tout ce qui n'est pas permis par le strict respect de la légalité.	Lawyers are forbidden to do anything that is not strictly legal.
Les applaudissements qui l'ont salué montrent bien que lorsqu'il y a un objectif ⋱.	The applause that rounded it off clearly demonstrates that when there is a speci ⋱.

Figure credit: Wolfram Data Repository
Slide adapted from Lori Levin

Statistical machine translation (1990s-2010s)

- Core idea: Learn a probabilistic model from data
- For French -> English, we want to find best English sentence *y*, given French sentence *x*
- Use Bayes' Rule to break this down into two components to be learned separately:



$$\text{argmax}_y P(y|x) = \text{argmax}_y \underbrace{P(x|y)}\,\underbrace{P(y)}$$

Translation Model

Models how words and phrases should be translated (*fidelity*). Learned from parallel data.

Language Model

Models how to write good English (*fluency*). Learned from monolingual data.

Slide adapted from Chris Manning

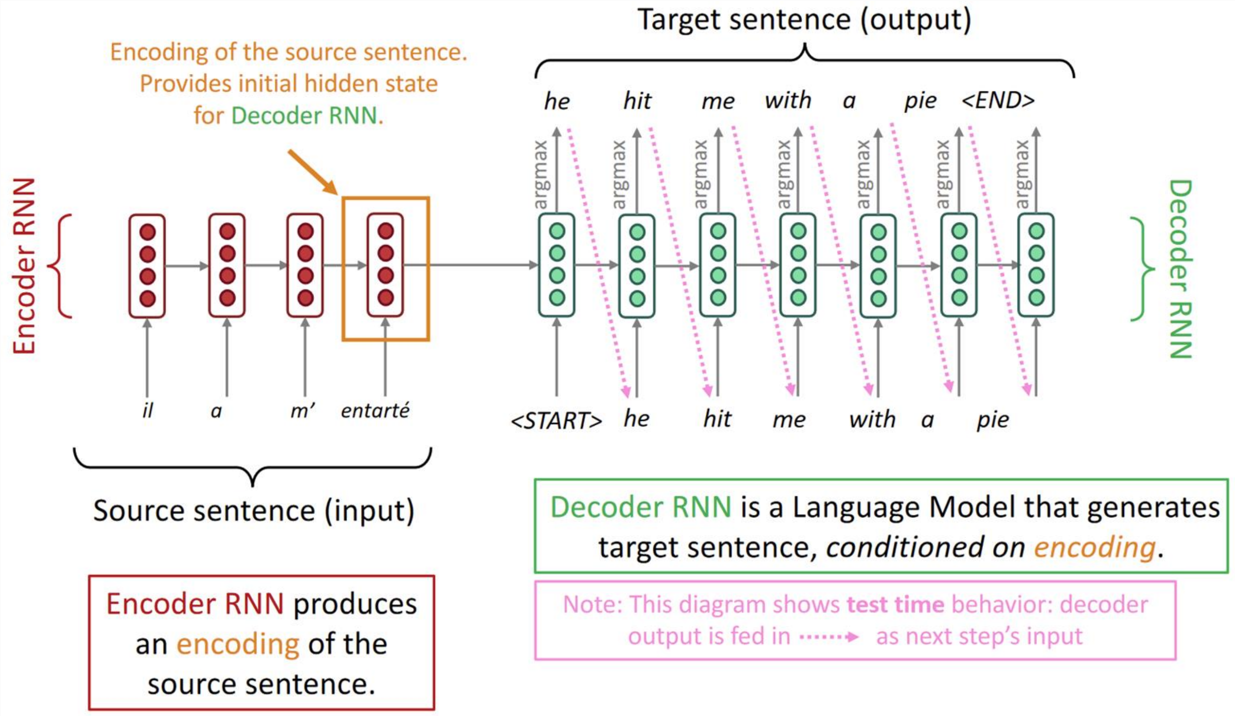
Statistical machine translation (1990s-2010s)

- The best SMT systems were extremely complex
    - Hundreds of important details
- Systems had many separately-designed subcomponents
    - Lots of feature engineering
    - Need to design features to capture particular language phenomena
- Required compiling and maintaining extra resources, like tables of equivalent phrases
    - Lots of human effort to maintain
- Repeated effort for each language pair

Slide adapted from Chris Manning

Neural machine translation (2010s on)

- Single end-to-end neural network
- Encoder-decoder (sequence-to-sequence, seq2seq) framework

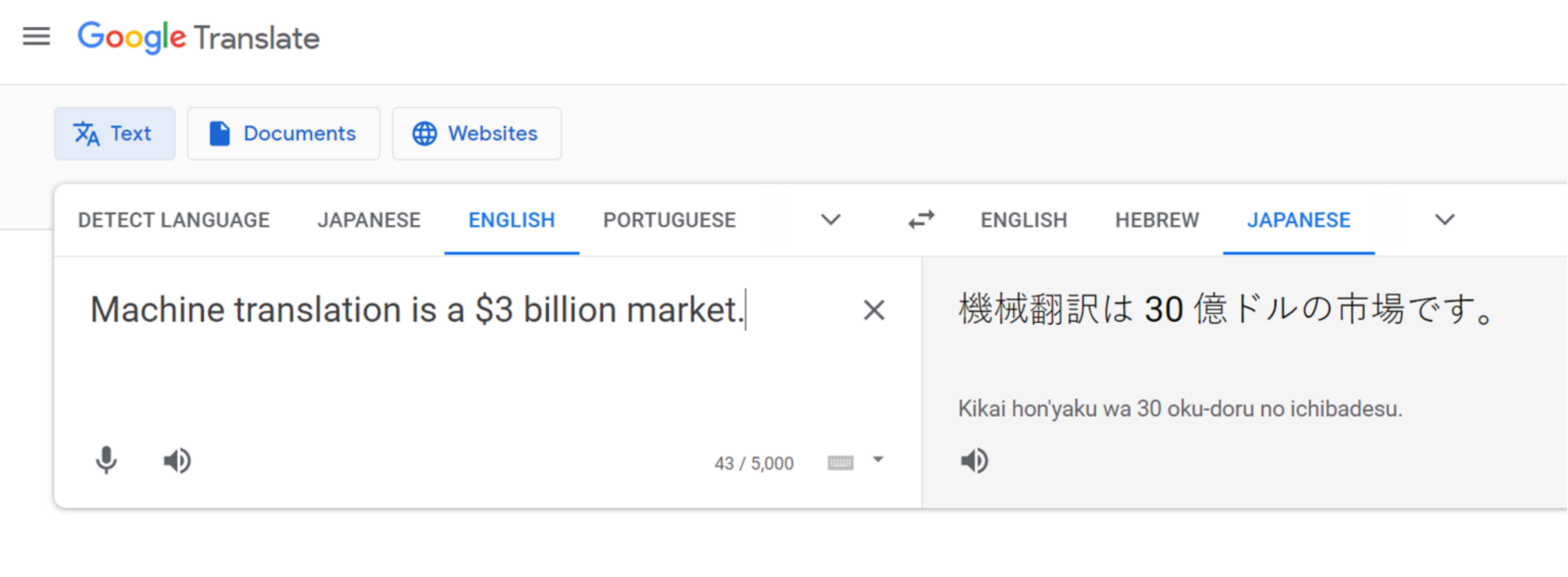

Target sentence (output)
Encoding of the source sentence. Provides initial hidden state for Decoder RNN.
he
hit
me
with
a
pie
<END>
argmax
Encoder RNN
Decoder RNN
il
a
m'
entarté
<START>
he
hit
me
with
a
pie
Source sentence (input)
Decoder RNN is a Language Model that generates target sentence, *conditioned on encoding*.
Encoder RNN produces an encoding of the source sentence.
Note: This diagram shows **test time** behavior: decoder output is fed in ·······▸ as next step's input

Slide adapted from Chris Manning

Translation in practice

Machine translation is a $3 billion market

Translation of text



Machine translation is a $3 billion market.

機械翻訳は 30 億ドルの市場です。

Kikai hon'yaku wa 30 oku-doru no ichibadesu.

Slide adapted from Lori Levin

Machine translation is a $3 billion market

Translation of speech

Person: Alexa, how do you say, “I hate this movie” in Japanese.

Alexa: “I hate this movie” in Japanese is “Kono eiga wa kirai da.”

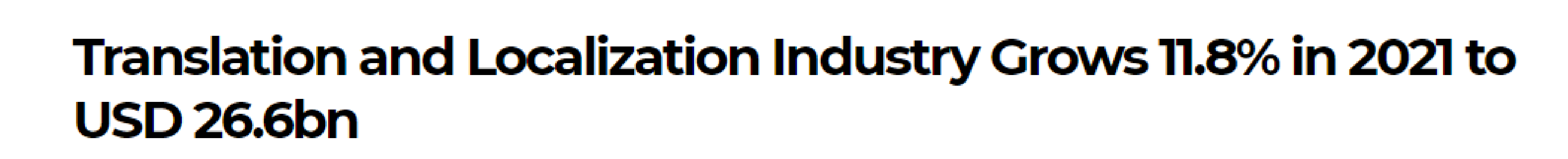
Person : Alexa, how do you say, “I hate this movie in Japanese” in Japanese.

Alexa: “I hate this movie in Japanese” in Japanese is “Kono eiga wa nihongo de wa kirai da.”

Real time translation of meetings is also now viable.

Slide adapted from Lori Levin

Most translation is still done by human translators

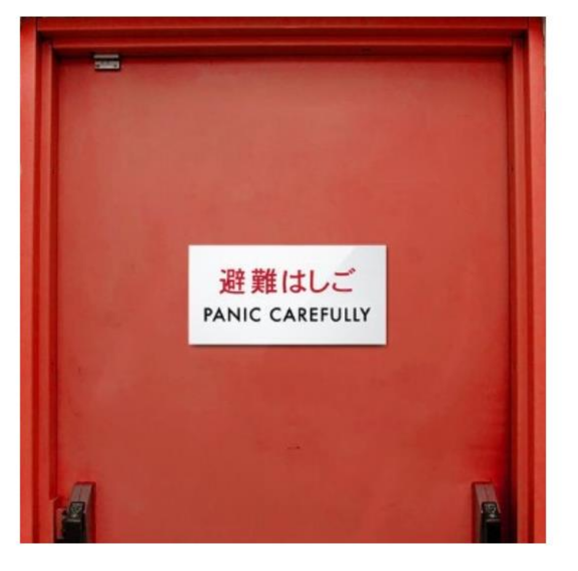


Translation and Localization Industry Grows 11.8% in 2021 to USD 26.6bn

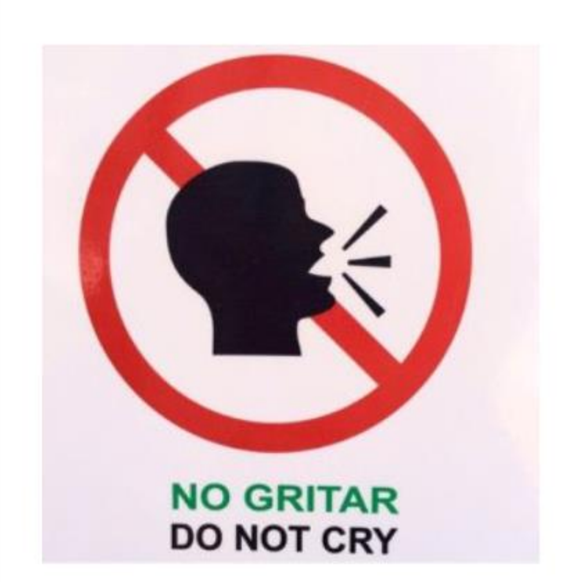
https://slator.com/translation-localization-industry-grows-in-2021-to-usd-26bn/
Image credit: AMN Healthcare
Slide adapted from Lori Levin

Post-editing and computer-assisted translation

- Checking and correcting of machine translation by humans is called **post-editing**



Evacuation Ladder



Do not yell

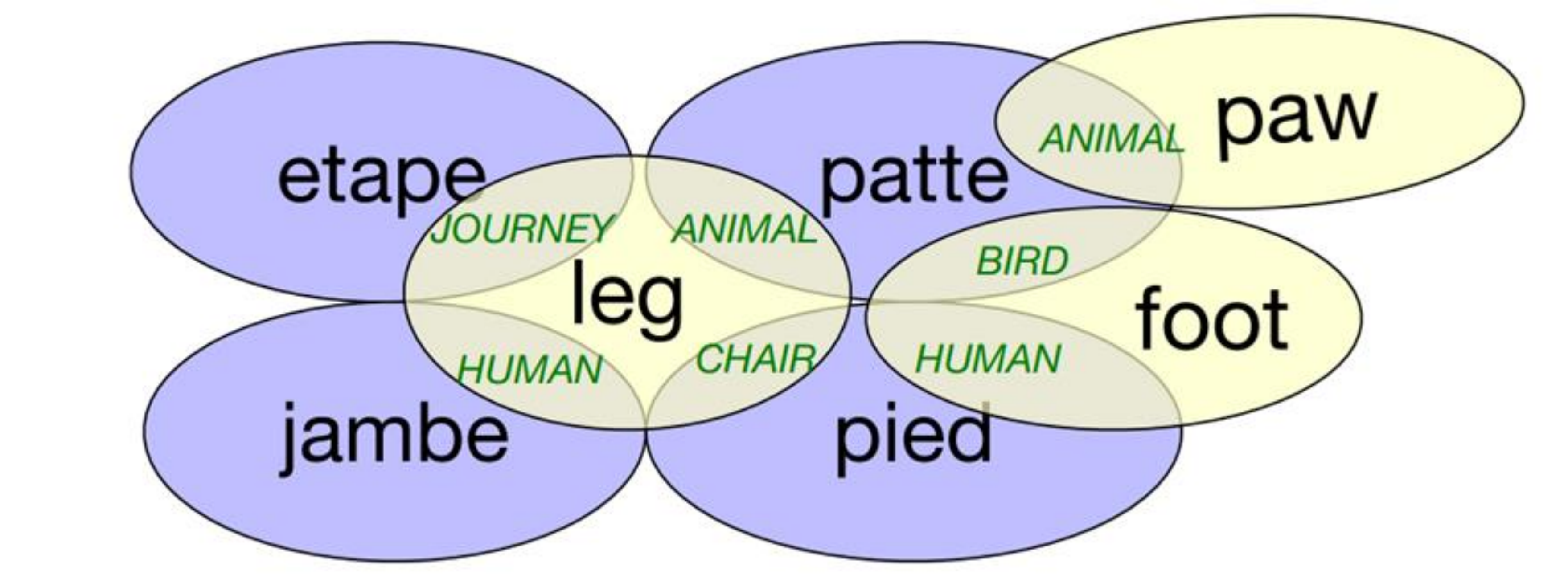
Slide adapted from Lori Levin

Images credit: https://www.languageconnections.com/blog/7-hilarious-machine-translation-mistakes/

Why is translation difficult?

Why not just look up each word in a dictionary and translate word-for-word?

Many-to-many mappings of words



Slide adapted from Lori Levin, Jurafsky & Martin, Hutchins & Somers

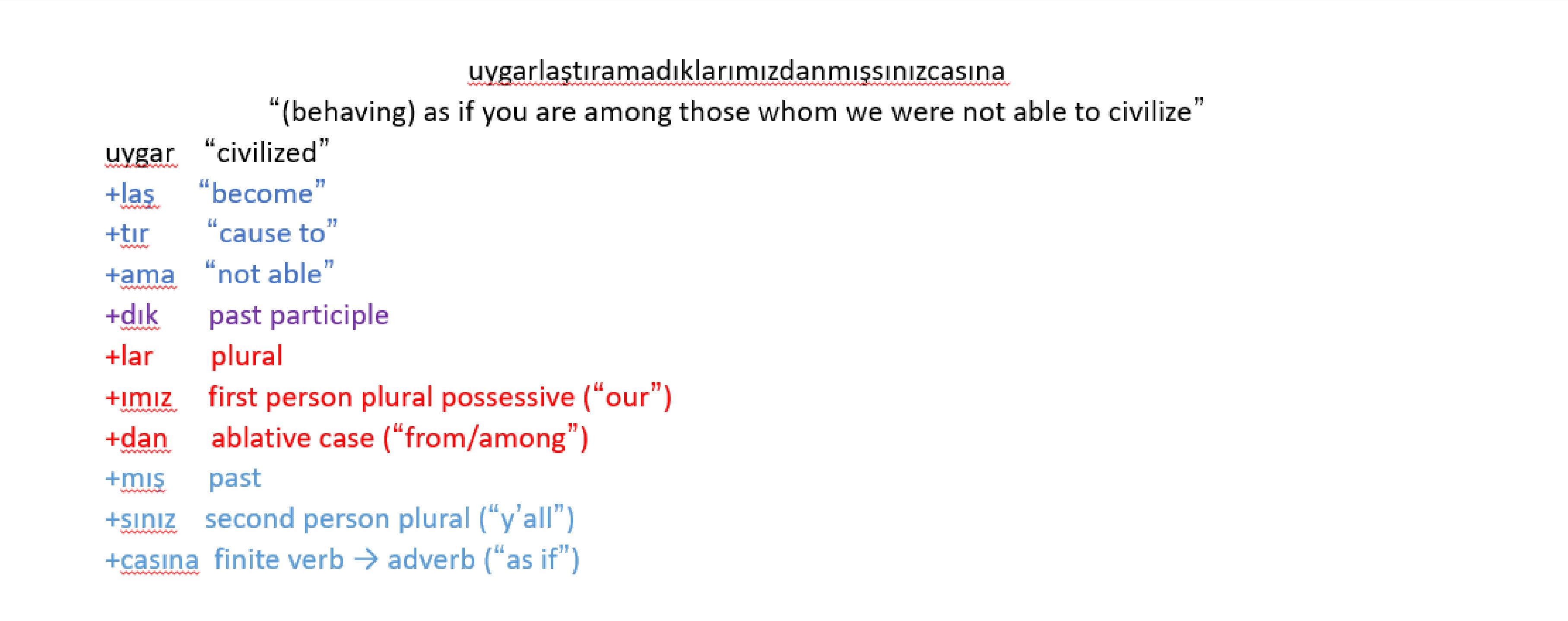
Why not translate word-for-word: grammar distinctions

The grammars of some languages make distinctions that other languages don't make:

- Russian *kniga* translates to English as *the book* or *a book*.
    - English grammar makes a distinction in definiteness
    - Russian grammar does not.
- English *it* translates to French *il*/*le* (masculine) or *elle*/*la* (feminine).
- English *a* translates to French as *un* (masculine) or *une* (feminine).
    - *Une chaise* (a chair) vs *un livre* (a book)
    - French grammar makes a distinction in gender
    - English grammar does not.

Slide adapted from Lori Levin

Why not translate word-for-word:
Different numbers of words to say the same thing



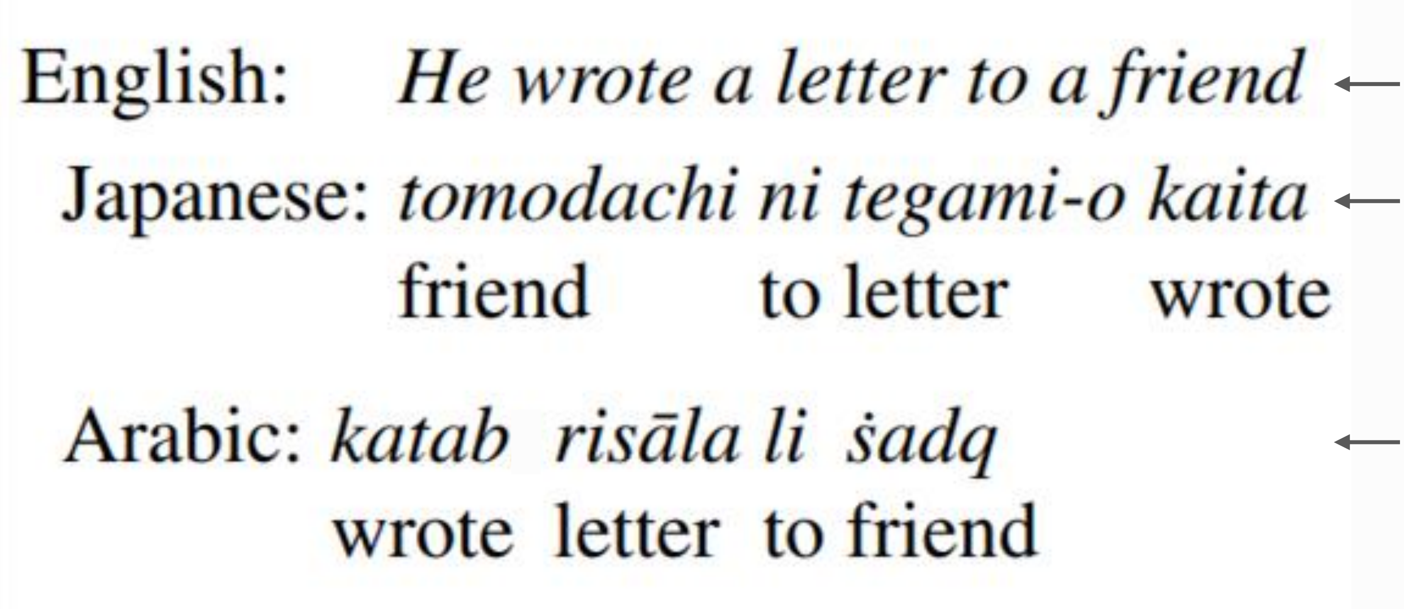
uygarlaştıramadıklarımızdanmışsınızcasına

"(behaving) as if you are among those whom we were not able to civilize"

uygar	"civilized"
+laş	"become"
+tır	"cause to"
+ama	"not able"
+dık	past participle
+lar	plural
+ımız	first person plural possessive ("our")
+dan	ablative case ("from/among")
+mış	past
+sınız	second person plural ("y'all")
+casına	finite verb → adverb ("as if")

Slide adapted from Lori Levin

Why not translate word-by-word: word order



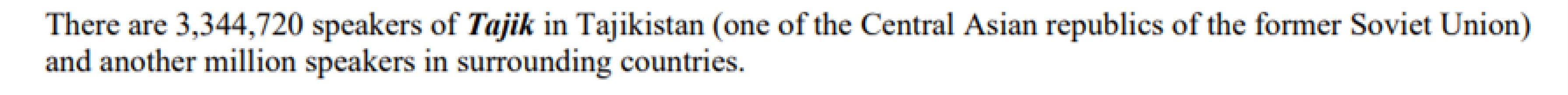
English: *He wrote a letter to a friend* ← SVO (verb-medial)

Japanese: *tomodachi ni tegami-o kaita* ← SOV (verb-final)
friend to letter wrote

Arabic: *katab risāla li ṡadq* ← VSO (verb-initial)
wrote letter to friend

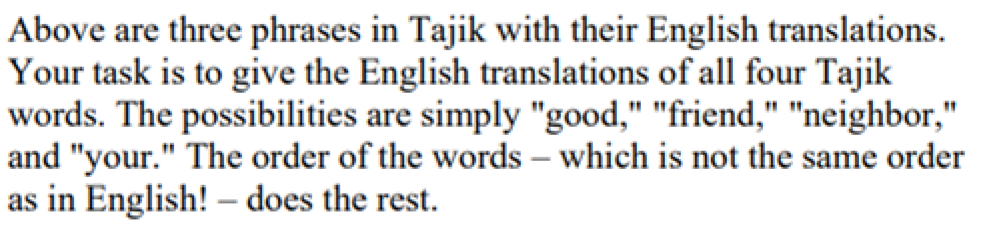
Slide adapted from Lori Levin, Jurafksy & Martin

Exercise: Tajik



There are 3,344,720 speakers of ***Tajik*** in Tajikistan (one of the Central Asian republics of the former Soviet Union) and another million speakers in surrounding countries.

дуусти хуби ҳамсояй сумо	a good friend of your neighbor
ҳамсояй дуусти хуби сумо	a neighbor of your good friend
ҳамсояй хуби дуусти сумо	a good neighbor of your friend



Above are three phrases in Tajik with their English translations. Your task is to give the English translations of all four Tajik words. The possibilities are simply "good," "friend," "neighbor," and "your." The order of the words – which is not the same order as in English! – does the rest.

дуусти ________

ҳамсояй ________

хуби ________

сумо ________

Problem by Adriana Solovyova
Slide adapted from Lori Levin

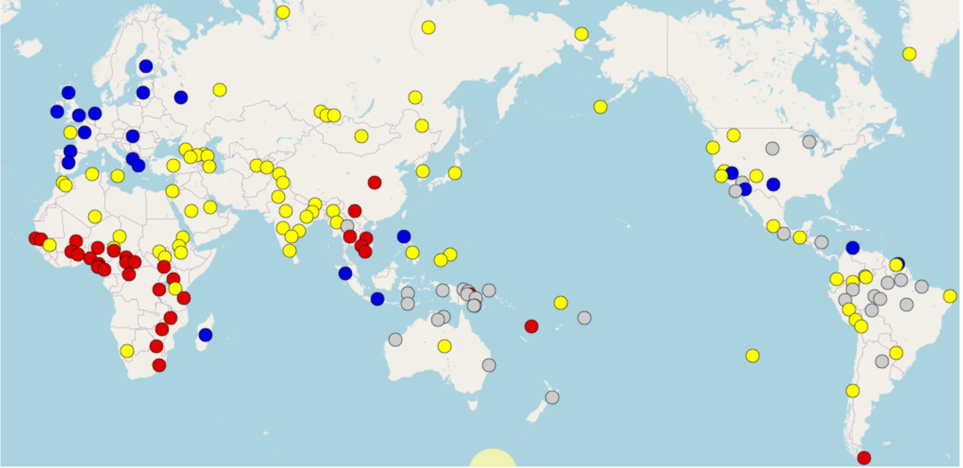
What is difficult about translation?

- People in NLP and MT have reduced “language divergences” to six major word order features from WALS, or seven lexical features
- But language typology is a system of “morphosyntactic strategies”, of which there are 1000s

THE WORLD ATLAS OF LANGUAGE STRUCTURES ONLINE

Feature 121A: Comparative Constructions

Yellow: X is big from Y, or X is big to Y
Red: X is big, exceeds Y
Grey: X is big, Y is small
Blue: X is big than Y



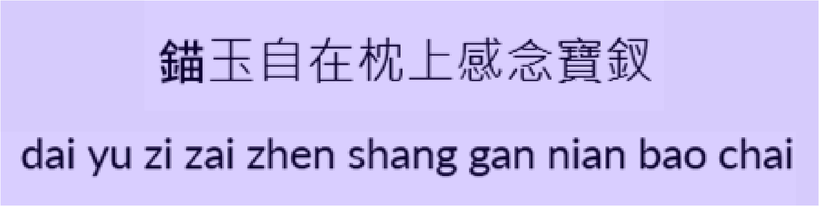
Slide adapted from Lori Levin

But the picture is not so gloomy

- MT researchers have made much progress on handling language divergence
- Use data from typologically similar languages
- Use a multilingual model trained on many typologically different languages

Slide adapted from Lori Levin

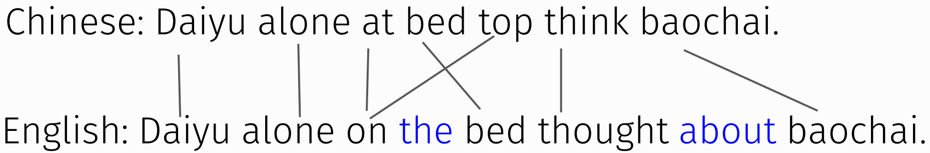
Why is translation difficult? Style and genre



鍿玉自在枕上感念寶釵

dai yu zi zai zhen shang gan nian bao chai

From "Dream of the Red Chamber", Cao Xue Qin (1792)

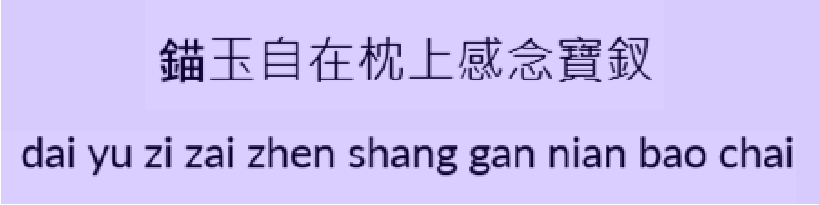


Chinese: Daiyu alone at bed top think baochai.

English: Daiyu alone on the bed thought about baochai.

Slide adapted from Patrick Fernandez, Graham Neubig, Xinyi Wang, Lori Levin

Why is translation difficult? Style and genre



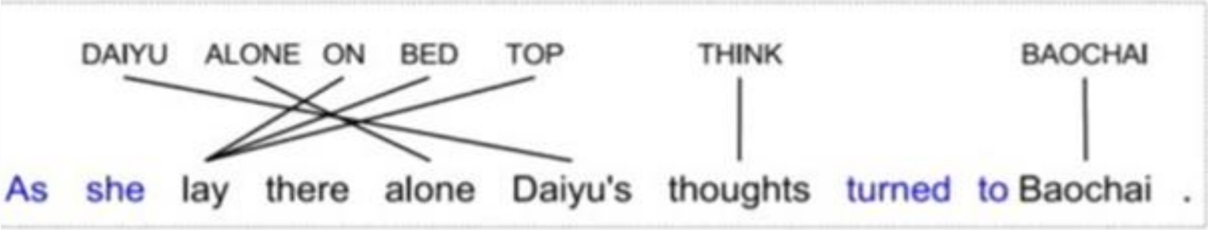
錨玉自在枕上感念寶釵

dai yu zi zai zhen shang gan nian bao chai

From “Dream of the Red Chamber”, Cao Xue Qin (1792)

Chinese:

English:



Parallel data is more likely to match styles (like literary style) than be an “exact” translation

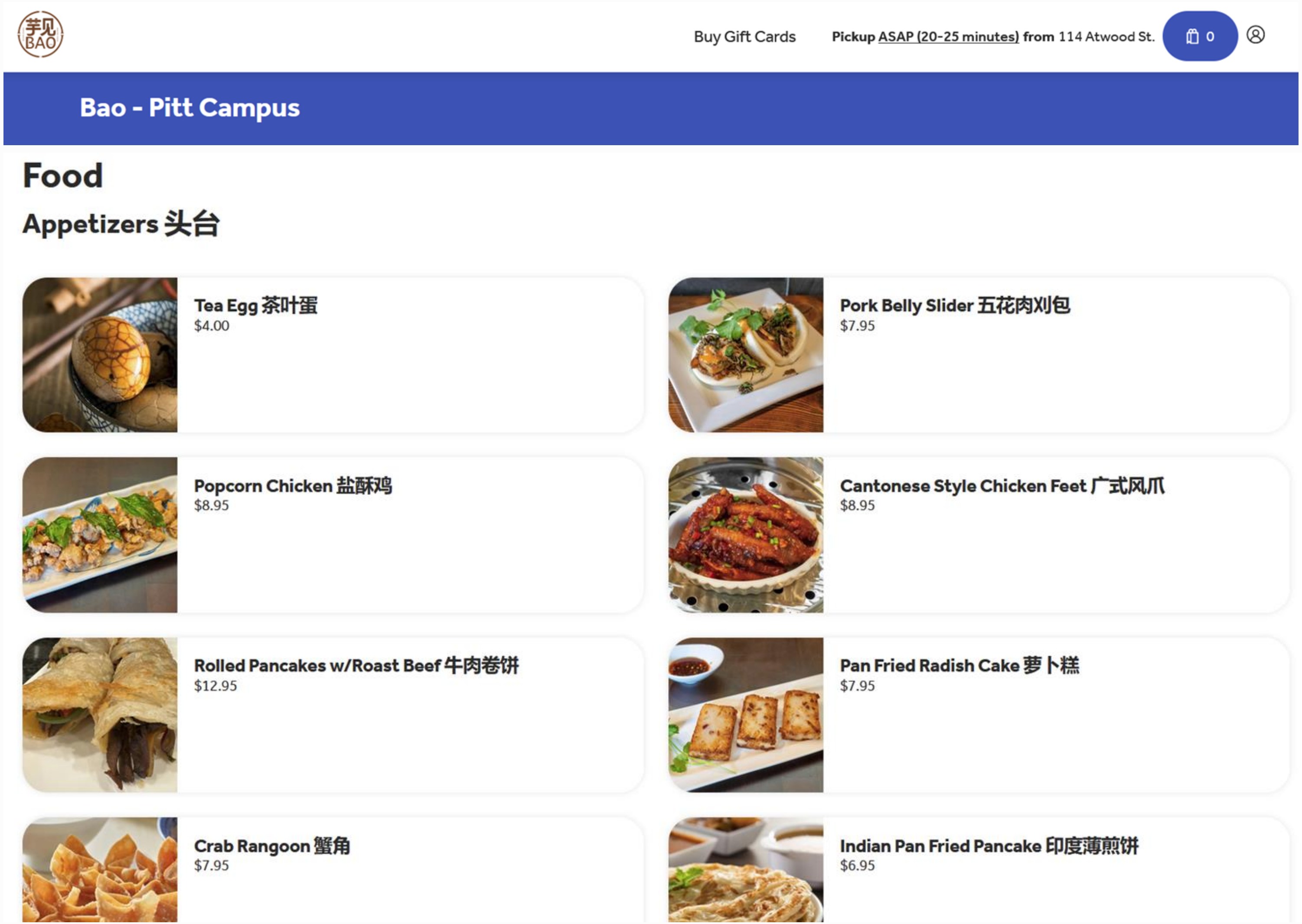
Slide adapted from Patrick Fernandez, Graham Neubig, Xinyi Wang, Lori Levin

Preparing for machine translation

1. Collect a parallel corpus
2. Align sentences
3. Tokenization
   - Split words into sub-word units, e.g., using BPE (Byte Pair Encoding)

Slide adapted from Lori Levin

Parallel corpora



Buy Gift Cards

Pickup ASAP (20-25 minutes) **from** 114 Atwood St.

0

Bao - Pitt Campus

Food

Appetizers 头台

Tea Egg 茶叶蛋	$4.00
Pork Belly Slider 五花肉刈包	$7.95
Popcorn Chicken 盐酥鸡	$8.95
Cantonese Style Chicken Feet 广式风爪	$8.95
Rolled Pancakes w/Roast Beef 牛肉卷饼	$12.95
Pan Fried Radish Cake 萝卜糕	$7.95
Crab Rangoon 蟹角	$7.95
Indian Pan Fried Pancake 印度薄煎饼	$6.95

Parallel corpora examples

- Europarl: Proceedings of the European Parliament; 21 languages; up to 2 million sentences
- United Nations Parallel Corpus: 10 million sentences in Arabic, Chinese, English, French, Russian, Spanish
- OpenSubtitles: movie and TV subtitles
- ParaCrawl: 223 million sentences in 23 EU languages

Slide adapted from Lori Levin, Jurafsky & Martin

What about parallel corpora for the other 7000 languages?

- For many languages, the only parallel text is the Christian Bible.
- Low-resource MT is a large area of research
  - How to leverage monolingual texts (backtranslation)
  - Humans in the loop
  - Leverage multilingual models

Slide adapted from Lori Levin

The “Bender Rule” [Bender 2019]

- When doing NLP work, please **name** the languages you are working with
  - “Always name the language(s) you’re working on”
- Don’t just assume the “default” language is English and work on other languages is “language specific”
- English has particularities
  - Massive amounts of training data available
  - Relatively fixed word order
  - Few inflectional forms per word (not much morphology)
  - Orthography: words indicated by whitespace, roughly phone-based

Discussion: the Bender Rule

- Korean: agglutination (affixes more important than word order), honorifics changing whole sentences, particles (Geonyeong)
- Japanese writing system (Yushui)
- Hindi idioms, OCR challenges with Devanagari (Anveshika)
- Tulu mostly an oral language (Kiran)
- Persian omitting subject pronouns when verb form makes subject clear (Rojin)
- Potential loss of cultural meanings and ways of seeing the world if language technologies are made for only English (Jack)
- Language tech won't recognize cultural values in non-English languages (Yifang)
- Cycle of data and models being created only for English and other high-resource languages (Jerry, Shijia)
- Economic incentives and research communities are English-centric (Hugh, Jiyang, Maanya)
- More important to have online presence than number of speakers (Akshat)
- Empirical results in NLP are English-specific (Alex)

Conclusion

- Modern machine translation methods use the neural encoder-decoder framework
- MT is often used in conjunction with human translators
- Language divergences (in word meaning, syntax structure, etc) make MT difficult
- Parallel corpora are used for training MT systems

Questions?