Why Multilingual NLP?

– Because we want to **understand** and **model** the meaning of texts in...

[Image from: epthinktank.eu]

– ...without manual (i.e., human) input and without perfect MT!

– **How many different languages are there in the world?**
  - How many have more than 10M speakers?

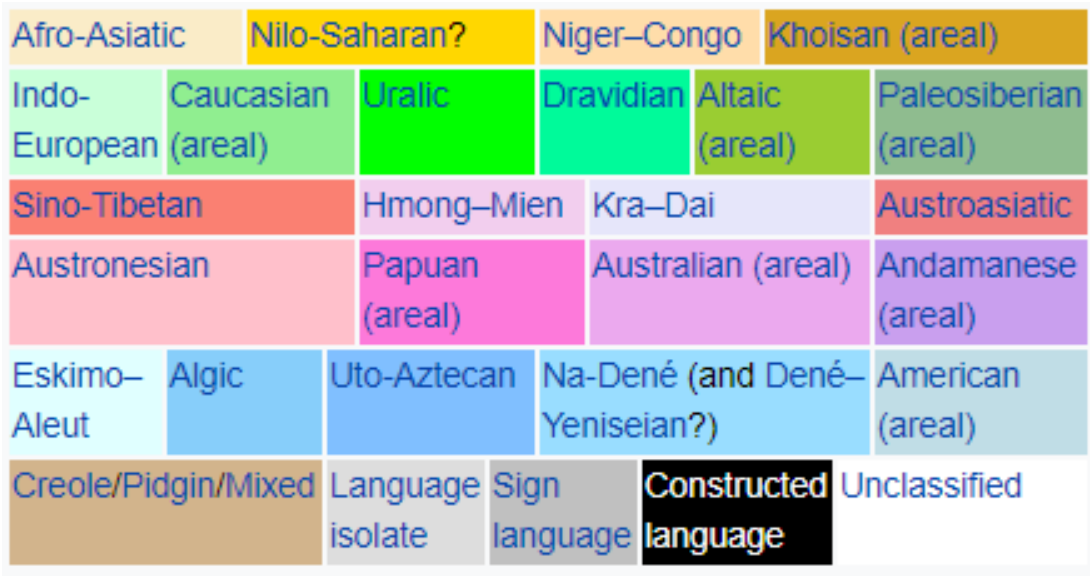
Why Multilingual NLP?

– According to Ethnologue (2020) there are **7,117 living languages**

70	Hejazi Arabic	14.5	0.188%	Afroasiatic	Semitic
71	Nigerian Fulfulde	14.5	0.188%	Niger–Congo	Senegambian
72	Bavarian	14.1	0.183%	Indo-European	Germanic
73	South Azerbaijani	13.8	0.179%	Turkic	Oghuz
74	Greek	13.1	0.170%	Indo-European	Hellenic
75	Chittagonian	13.0	0.169%	Indo-European	Indo-Aryan
76	Kazakh	12.9	0.168%	Turkic	Kipchak
77	Deccan	12.8	0.166%	Indo-European	Indo-Aryan
78	Hungarian	12.6	0.164%	Uralic	Ugric
79	Kinyarwanda	12.1	0.157%	Niger–Congo	Bantu
80	Zulu	12.1	0.157%	Niger–Congo	Bantu
81	South Levantine Arabic	11.6	0.151%	Afroasiatic	Semitic
82	Tunisian Arabic	11.6	0.151%	Afroasiatic	Semitic
83	Sanaani Spoken Arabic	11.4	0.148%	Afroasiatic	Semitic
84	Min Bei Chinese	11.0	0.143%	Sino-Tibetan	Sinitic
85	Southern Pashto	10.9	0.142%	Indo-European	Iranian
86	Rundi	10.8	0.140%	Niger–Congo	Bantu
87	Czech	10.7	0.139%	Indo-European	Balto-Slavic
88	Ta'izzi-Adeni Arabic	10.5	0.136%	Afroasiatic	Semitic
89	Uyghur	10.4	0.135%	Turkic	Karluk
90	Min Dong Chinese	10.3	0.134%	Sino-Tibetan	Sinitic
91	Sylheti	10.3	0.134%	Indo-European	Indo-Aryan

Language variety

– **Language family**: group of languages that originate from the same *ancestral/parental* language (proto-language)



[Image from: Wikipedia]

– **Language isolates**: no known/demonstrable genealogical relationship with any other language:

- *Basque, Korean*
- Indo-European language isolates: *Albanian*, *Armenian*, *Greek*

Why Cross-Lingual NLP?

– Because we want to transfer supervised models for NLP tasks...
  - Trained on **annotated datasets** we have in **resource-rich languages**
  - Make predictions in resource-lean target languages

English

Language Transfer

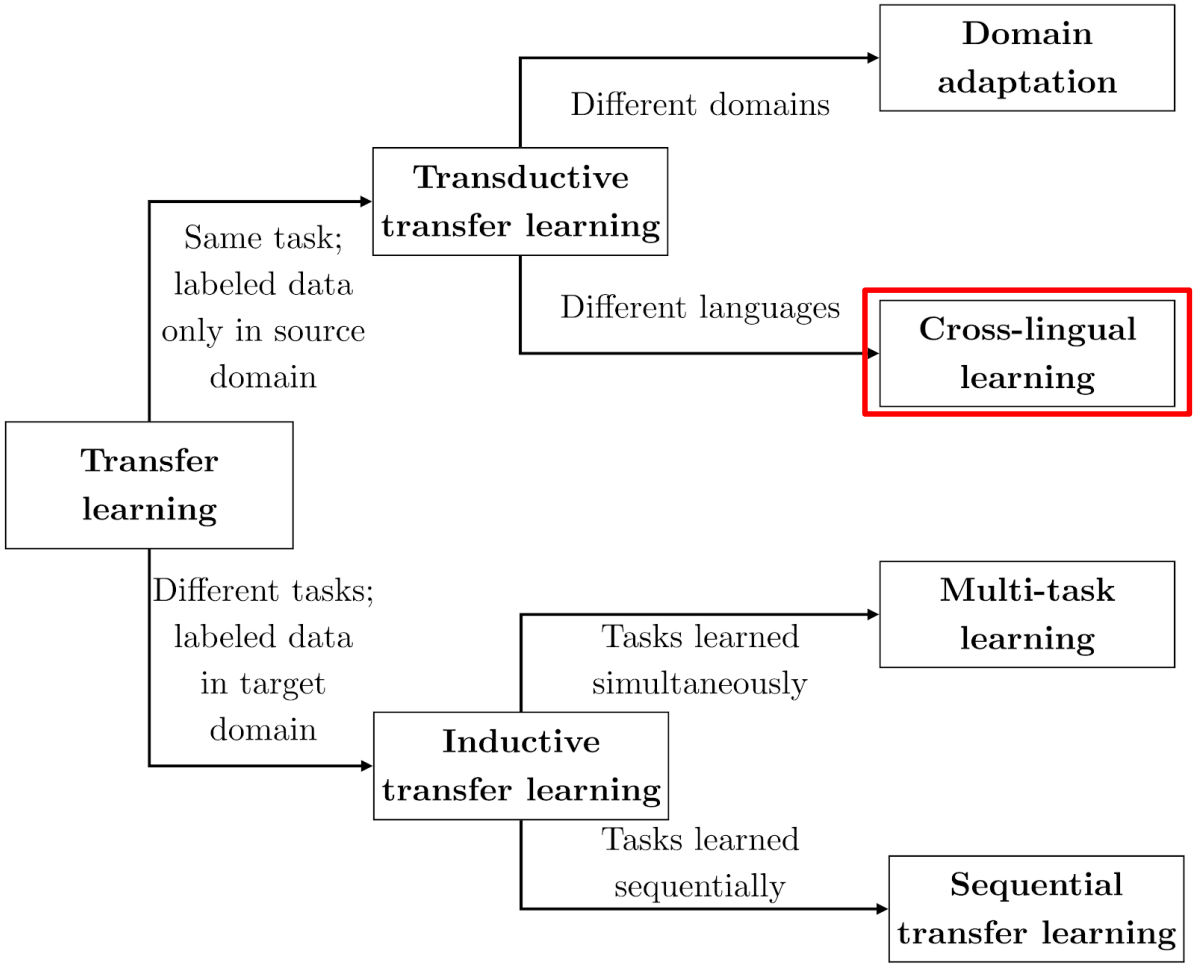
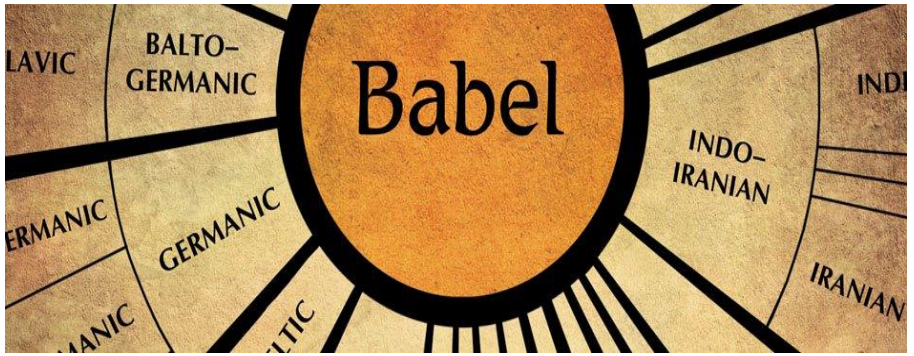


Image from [Ruder, 2019]

Crossing the Language Chasm

– **Old paradigm**:
  - Language-specific NLP models
  - Language-specific feature computation (i.e., preprocessing)

– **New paradigm(s)**:
  - Representation learning: semantic vectors (embeddings)
  - Multilingual / cross-lingual representation learning



Crossing the Language Chasm: symbolic approaches

1. Full-Blown MT (SMT or NMT)
   - Parallel data needed, critical for under-resourced languages
   - Translate everything from the target language to the source language
2. Multilingual KBs
   - Texts represented using entities from a multilingual KB
   - Same entity ID for same concepts across languages
   - Issues: coverage, entity linking

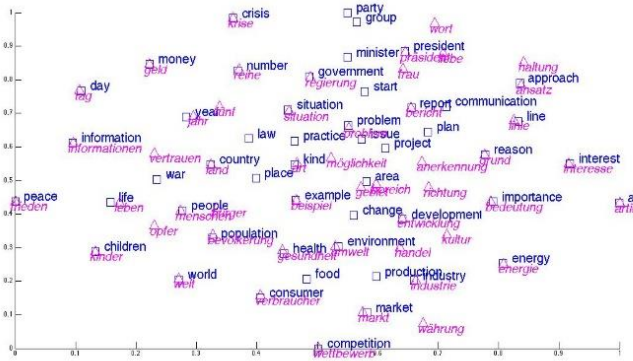
BabelNet 2.0

A very large multilingual **encyclopedic dictionary** and **ontology**

Crossing the Language Chasm: representation learning

3. Multilingual / Cross-lingual representations of meaning

- **Word-level**
  - Cross-lingual word embeddings
  - Words with similar meaning across languages have similar vectors



- **Text encoding**
  - Multilingual unsupervised pretraining
    - Multilingual BERT [Devlin et al., ‘19]
    - XLM(-R) [Conneau & Lample, ’19, Conneau et al., 2020]
    - mT5 [Xue et al., 2020]