Multilingual NLP

1. Language Diversity

(+ Course Organization)

Prof. Dr. Goran Glavaš
Center for AI and Data Science (CAIDAS), Uni Würzburg

Image: Alexander Mikhalchyk

After this lecture, you'll...

- Comprehend the linguistic diversity of the world
- Learn about language variation and linguistic universals
- Understand why we need multilingual NLP

- Know what this course is about and...
- ...which topics we'll cover
- Know what's your part of the job to earn the credits

Content

- **Why Multilingual NLP?**
- Linguistic Diversity and Universals
- Language Identification

—

- About the mNLP Course
- Topics and Schedule
- Organization

Why Multilingual NLP?

- The world is **still** massively multilingual
- Q: How many people speak English (as first or foreign lang.)?
- Q: How many people speak any foreign language?
  - 43% people are bilingual
  - **50+% monolingual**!

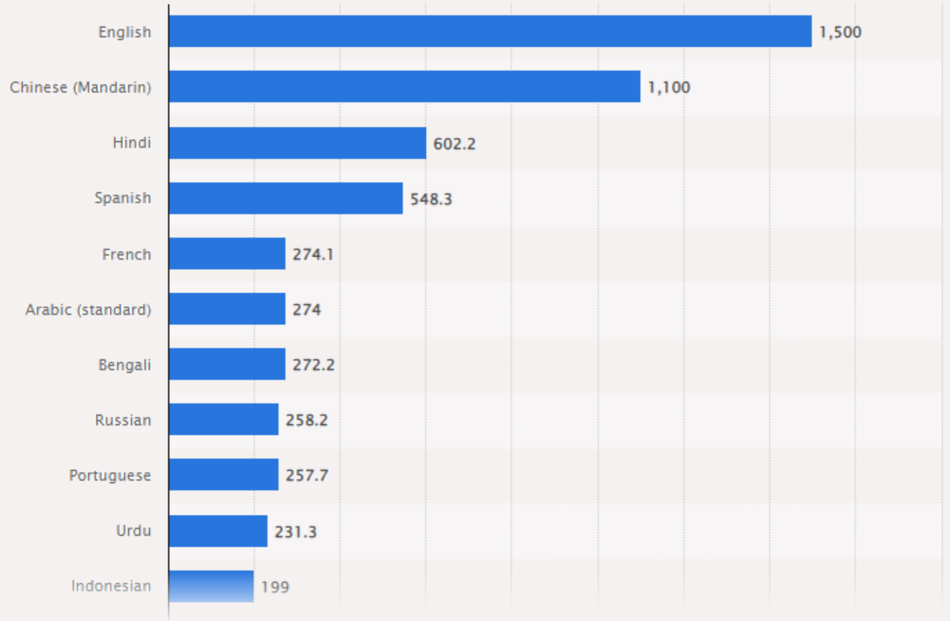


Image from: https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/

Why Multilingual NLP?

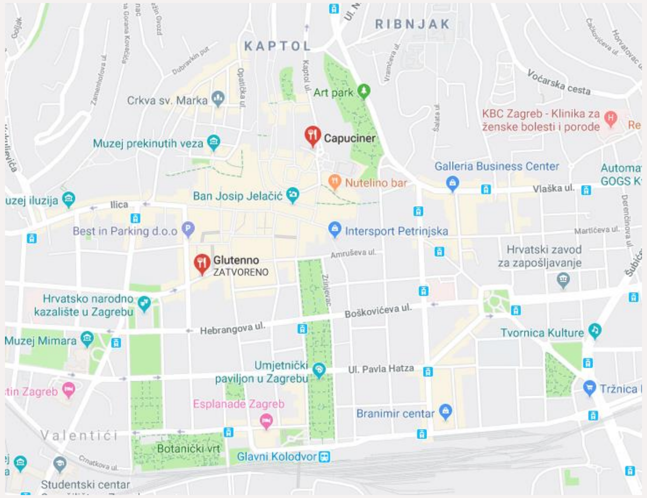
- Profound **democractic** reasons
  - Inclusion
  - Digital equity
  - Mitigating cross-cultural biases

- **Digital language divide**
  http://labs.theguardian.com/digital-language-divide
  - Without multilingual NLP, and especially in a digital world, the **limits of my language(s)** define the **limits of my world**!
  - The language(s) you speak shape your experience of the internet!

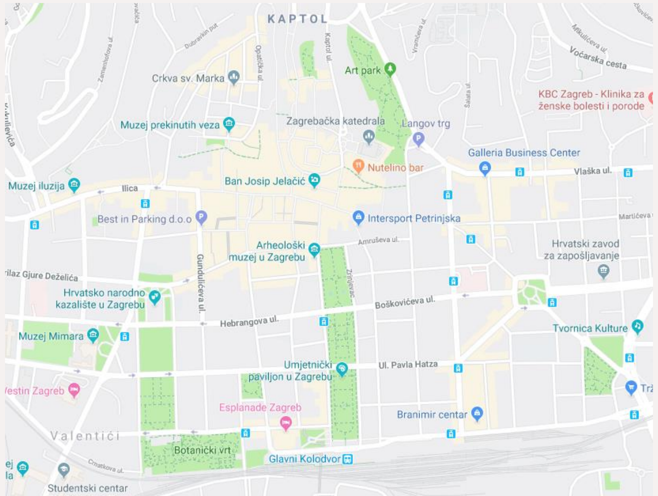
Why Multilingual NLP?

- **Inequality of information** and representation can affect how we understand places, events, processes, …

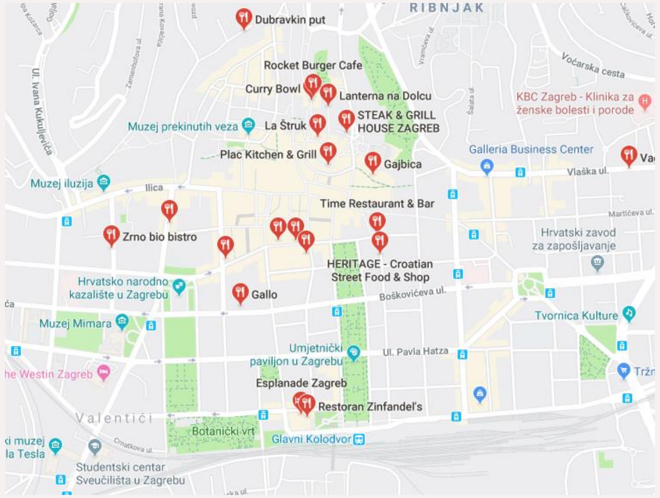
We're in Zagreb (Croatia) searching for…



…éttermek (HU)



…jatetxea (EU)



…restaurants (EN)

Why Multilingual NLP?

- **Inequality** of information: language technology not nearly equally good across languages

Model: Default (GPT-3.5)

Jack is a college professor and Jill is a college student. Jill is Jack's biological mother. Who is older, Jack or Jill?

Jill must be older than Jack since she is his biological mother.

Model: Default (GPT-3.5)

Jack ist ein College-Professor und Jill ist eine College-Studentin. Jill ist Jacks leibliche Mutter. Wer ist älter, Jack oder Jill?

Es ist unmöglich, dass Jill die biologische Mutter von Jack ist, wenn Jack ein College-Professor ist und Jill eine College-Studentin ist. Es scheint sich um eine unlogische oder humorvolle Aussage zu handeln. Ohne weitere Informationen ist es schwierig zu bestimmen, wer älter ist.

Why Multilingual NLP?

- Despite all recent ML/NLP advances (e.g., LLMs with zero-shot /few-shot in-context learning abilities), **supervised learning** still leads to best performance
- Supervised learning requires labeled training data for concrete NLP tasks (e.g., named entity recognition, sentiment classification)
- Labeling data is tedious, time-consuming, and expensive
  - For most tasks and applications, labeled data exists only in English and perhaps a handful of major world languages

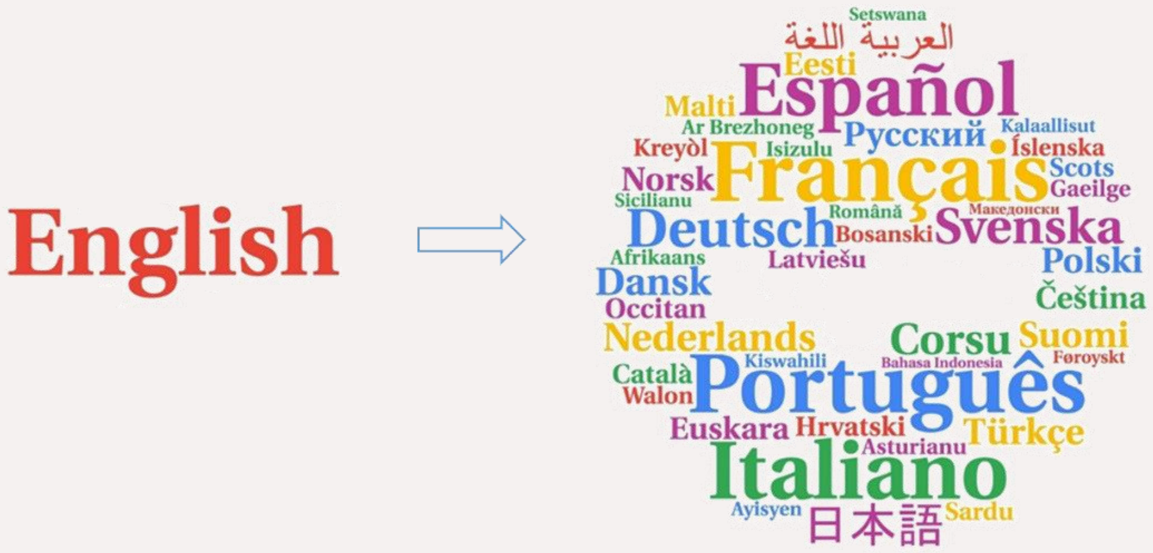
Why Multilingual NLP?

- Q: How many different languages are there in the world?
- According to Ethnologue there are/were **7,164** living languages in 2024
- Q: How many languages with more than 10M native speakers?
- Q: Obtain labeled data for each language and task?

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

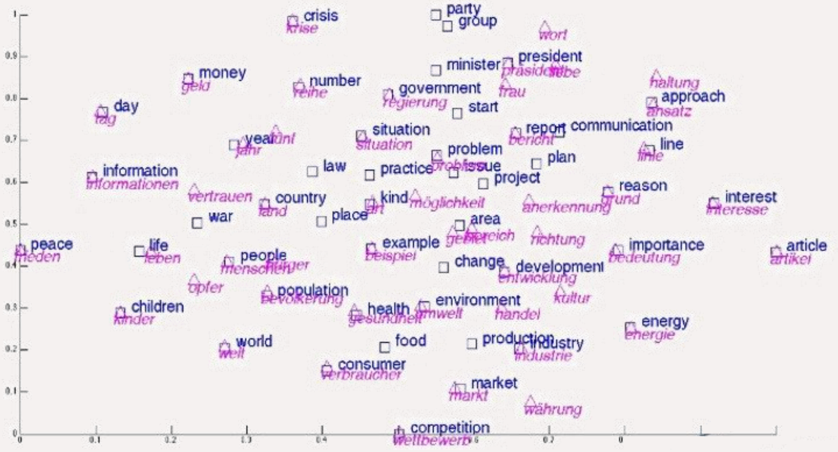
Why Multilingual NLP?

- **Cross-Lingual transfer**: transfer supervised models for NLP tasks
  - Models trained on labeled data in high-resource <u>source</u> language…
  - …make predictions on texts in low-resource <u>target</u> languages with little or no labeled data



Why Multilingual NLP?

- **Multilingual representation spaces** necessary for cross-lingual transfer
  - Words/sentences/texts that have the **same/similar meaning**, get same/similar vectors…
    - …whether from the same language or different languages



- In this course
  - Cross-lingual word embeddings
  - Multilingual language models

Content

- Why Multilingual NLP?
- **Linguistic Diversity and Universals**
- Language Identification

—

- About the mNLP Course
- Topics and Schedule
- Organization

Language Diversity

- Languages are mutually related, originate from shared ancestors
  → genealogy of languages
- Languages (genealogically related or not) may share structural (syntactic) and functional (semantic) properties
  → linguistic typology
- Languages (genealogically related or not) interact with each other and borrow concepts (and words for those concepts)
  → etymology

Genealogy of Languages

- **Language families** (and subfamilies)
  - Groups of languages that have originated from the same ancestor
  - Hierarchical organization
    - Languages from same families often (but not necessarily) also typologically similar
  - A lot of discussions and disputes among (comparative) linguists
    - [Altaic]() language family?

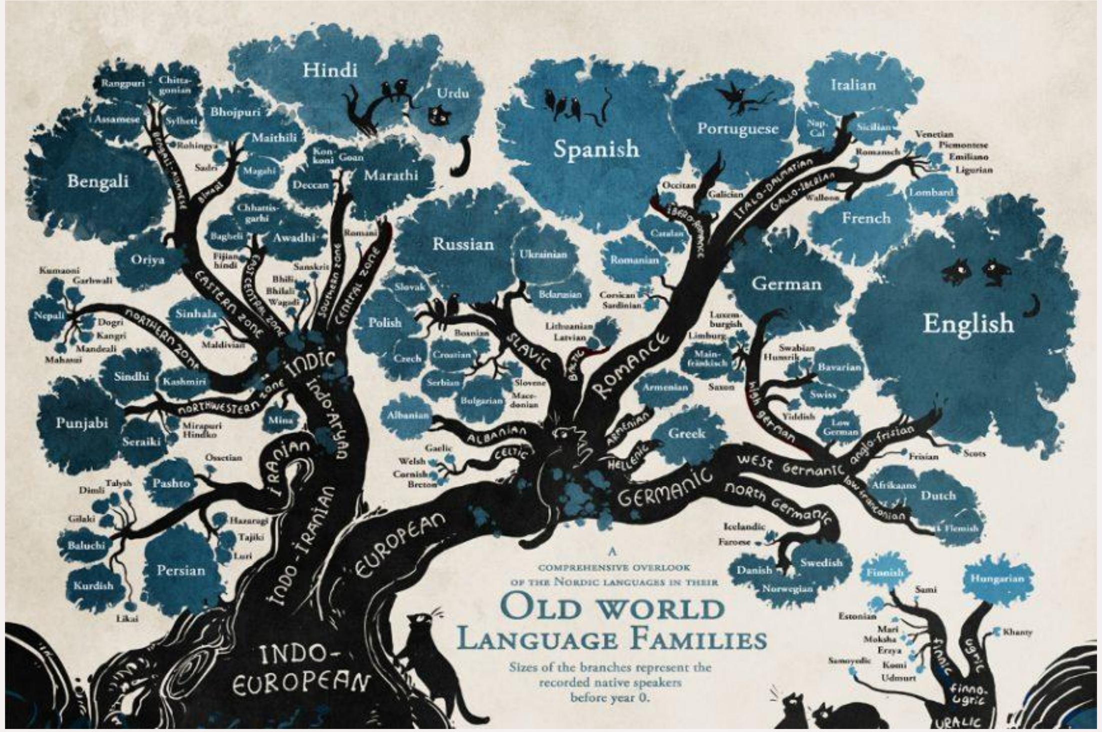


Image from:
[https://www.theguardian.com/education/gallery/2015/jan/23/a-language-family-tree-in-pictures]()

Genealogy of Languages

- Ethnologue (v. 2023) lists **13 top-level language families** (each encompassing at least 1% of 7,164 known languages)
  - Niger-Congo – 1,542 languages
  - Austronesian – 1,257 languages
  - Sino-Tibetan – 455 languages
  - Indo-European – 448 languages
  - Afro-Asiatic – 377 languages
  - …
  - Dravidian – 86 languages
  - Tupian – 76 languages
- Language technology (i.e., NLP models and tools) is <u>by far</u> most developed and effective for **Indo-European** languages
  - Q: Why?

Linguistic Typology

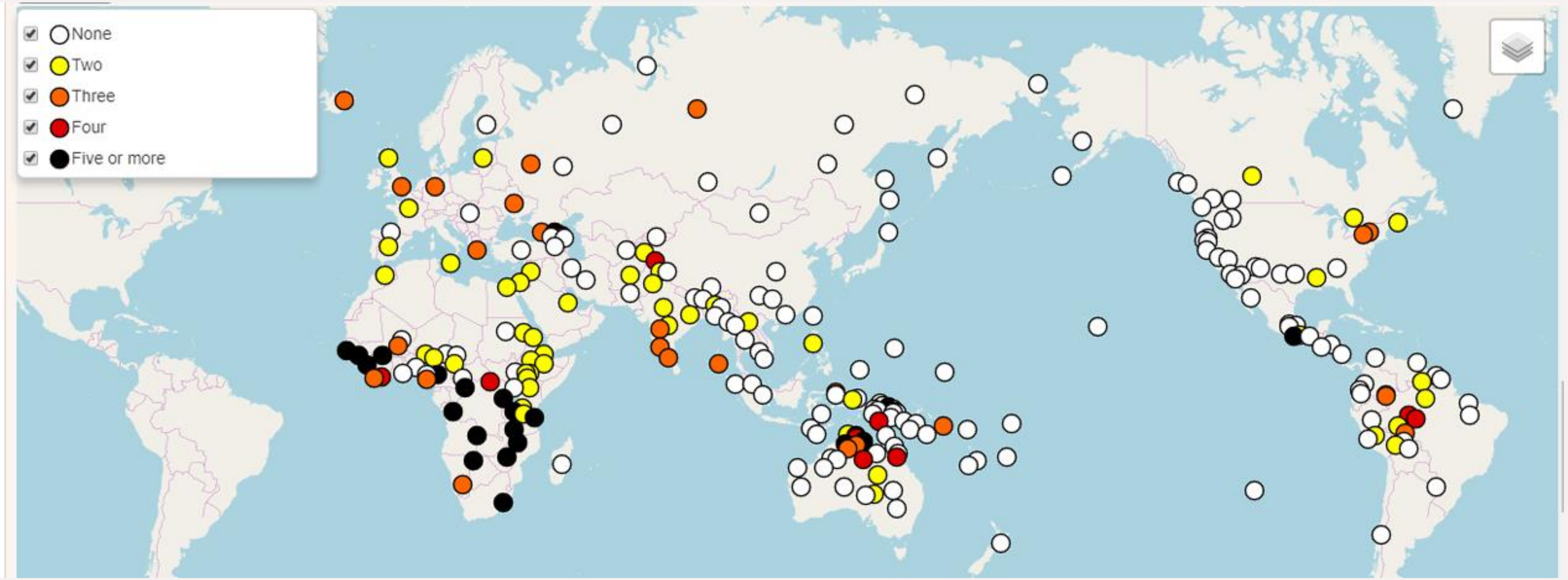
- **Linguistic Typology** = field of linguistics that classifies languages based on their functional and structural properties
  - Derived from systematic comparison between languages

- Q: What properties?

- **Syntactic / grammatical typology**
  - Dominant word order (between Subject, Verb/Predicate and Object)
    - SVO (English, Chinese, Swahili…), SOV (Japanese, Persian, Turkish, …)
    - OSV (e.g., Tobati), OVS (e.g., Urarina)

Linguistic Typology

- **Number of grammatical genders**
  - Yimas (250 speakers) language has **11** genders!
  - Different grammatical genders for animals, plants, ...



Linguistic Typology

- **Linguistic Typology** = field of linguistics that classifies languages based on their functional and structural properties
  - Derived from systematic comparison between languages

- Q: Which properties?

- **Phonological typology**
  - Patterns in the structure and distribution of sound systems of languages
  - Has fricatives?
    - Consonants made by friction of breath in a narrow opening
    - Sounds like *th* in English
  - Has plosives?
    - Sounds like most occurrences of *k*, *g*, *b*, *d* in English

Linguistic Typology

- **Morphological typology**
  - Classifies languages based common **morphological** structures
  - **Morpheme**: smallest meaningful constituent of a linguistic expression
    - Parts of words (prefixes like *un* or *im*, stems of words, suffixes like *ing*)
  - Isolating languages: <u>morpheme per word ratio</u> close to 1, no inflectional morphology (no „declination" of nouns, no „conjugation" of verbs)
    - Chinese, Vietnamese, Yoruba
  - Fusional languages: single inflectional morpheme to denote multiple grammatical, syntactic, or semantic features
    - E.g., tense, gender, time of verbs
    - Most world languages
  - Agglutinative languages: concatenate multiple morphemes (with typically one morpheme per function)
    - E.g., Finnish, Turkish, Persian, …

Sources of Typological Knowledge

- **Typological databases**
  - Large number of typological properties coded manually by linguists for large number of natural languages

ID#	Feature Name	Category	Feature Values
1	Consonant Inventories	Phonology (19)	{1:Large, 2:Small, 3:Moderately Small, 4:Moderately Large, 5:Average}
23	Locus of Marking in the Clause	Morphology (10)	{1:Head, 2:None, 3:Dependent, 4:Double, 5:Other}
30	Number of Genders	Nominal Categories (28)	{1:Three, 2:None, 3:Two, 4:Four, 5:Five or More}
58	Obligatory Possessive Inflection	Nominal Syntax (7)	{1:Absent, 2:Exists}
66	The Perfect	Verbal Categories (16)	{1:None, 2:Other, 3:From 'finish' or 'already', 4:From Possessive}
81	Order of Subject, Object and Verb	Word Order (17)	{1:SVO, 2:SOV, 3:No Dominant Order, 4:VSO, 5:VOS, 6:OVS, 7:OSV}
121	Comparative Constructions	Simple Clauses (24)	{1:Conjoined, 2:Locational, 3:Particle, 4:Exceed}
125	Purpose Clauses	Complex Sentences (7)	{1:Balanced/deranked, 2:Deranked, 3:Balanced}
138	Tea	Lexicon (10)	{1:Other, 2:Derived from Sinitic 'cha', 3:Derived from Chinese 'te'}
140	Question Particles in Sign Languages	Sign Languages (2)	{1:None, 2:One, 3:More than one}
142	Para-Linguistic Usages of Clicks	Other (2)	{1:Logical meanings, 2:Affective meanings, 3:Other or none}

Examples

- **World Atlas of Languages**
  - 2,600+ languages
  - 192 typological properties
- **URIEL Typological Compendium**
  - 8,000+ „languages"
  - 284 typological properties
  - Accessible through a neat Python library lang2vec

Etymology

- **Etymology** studies the origin and evolution of meaning of words including its constituent parts (i.e., morphemes)

- Vocabularies of languages develop (in part) via mutual interactions
  - These are often geographically (and historically) conditioned, not just genealogically
  - Geographic/historic links sometimes not obvious („second order")
    - E.g., South-Slavic languages have many words of Persian origin
  - Induction of **multilingual representation spaces** and cross-lingual transfer, especially for lexical tasks
    - Easier between languages with shared etymology

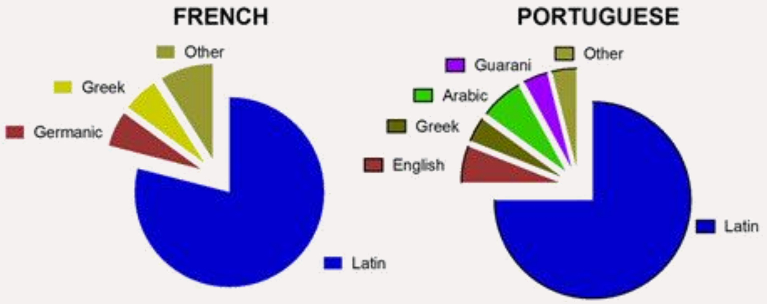


Image from: https://www.researchgate.net/figure/The-percent-distribution-of-words-in-Modern-French-and-Portuguese-as-evaluated-by_fig2_323506310

Language Universals

- **Linguistic universals** = properties that hold true for all (or almost all) languages of the world
  - All languages have vowels & consonants and distinguish between nouns & verbs
  - **Syntax**: Universal Grammar (Chomsky)? Humans (and especially children) learn any language quickly
    1. Without formal instruction
    2. From limited input/examples („poverty of stimulus")
    - Thus an universal grammar must exist in the human brain
  - **Semantics**: Irreducible semantic core (Leibniz)?
    - Meaning of all words in all languages can be derived from a finite set of core semantic concepts/components
    - Some NLP approaches subscribe to this same idea

Content

- Why Multilingual NLP?
- Linguistic Diversity and Universals
- **Language Identification**

—

- About the mNLP Course
- Topics and Schedule
- Organization

Language Identification

- Corpora for many NLP tasks and applications is crawled from the Web
  - Language of the text not given (or not reliable)

- **Language identification**: automatic prediction of the text's language
  - Document, paragraph, or sentence level
  - Code-switching: intra-sentence language changes (common in social media text)

- In principle, an <u>easy task</u>, solutions based on
  - Dictionaries
  - Character sets
  - Character frequency and distributions

  - Rule-based or machine learning solutions

Language Identification

- **Dictionary-based** language identification
  - Assuming you have a dictionary of the respective language
  - Text in language L if >X% of words found in the dictionary of L
  - Reliable for long(er) texts

- Identification based on **character sets**
  - Some characters are unique to some languages
    - „**ć**" exists only in South-Slavic languages (and Polish)
    - „Modri**ć** is a great midfielder" → Croatian?!
  - Many languages also have unique scripts

Georgian script "Mkhedruli"
Image from Wikipedia

Language Identification

- Obviously, we need to take into account (relative) **frequencies** with which characters occur in a language

1. Obtain a **character distribution** for a language p_l
   - Relative frequency (probability) of occurrence for each character
   - Computed on a large „training" corpus

2. Compute the character distribution for the text in question p_t
   - The shorter the text, the less reliable estimate of the distribution

3. Compare the two distributions with some metric
   - E.g., Kullback-Leibler divergence of p_t w.r.t. p_l

Language Identification

- **Problems** with single character distributions
  - Languages (esp. closely related) may have <u>similar distributions</u>
  - + estimates from (not too long) texts unreliable

- **Character n-gram** distributions are more reliable
  - Char n-grams: sequences of small number (2 or 3) of characters
    - E.g., higher relative freq. of ing, ent, ion, red, the?
    - E.g., higher relative freq. of sch, ung, aft, ich, ter?
  - Char n-gram distributions are much less likely to be similar across languages, even closely related ones

Language Identification

- **Machine learning** approach to language identification
  - Especially needed when distinguishing between closely related languages (e.g., Bosnian and Croatian)
- **Input features**
  - If we have dictionaries for languages in consideration:
    - Proportion of tokens found in dict. of each language
  - Overall vocabulary of characters across all langs in consideration
    - Relative frequency of each character in the text
  - Overall vocabulary of char n-grams across all langs in consideration
    - Relative frequency of each character in the text
- Need the **training dataset**: corpus of texts for which language is known

Content

- Why Multilingual NLP?
- Linguistic Diversity and Universals
- Language Identification

—

- **About the mNLP Course**
- Topics and Schedule
- Organization

Course description

- **Why this course?** Because…
  - …NLP is a central area of AI
  - …we're going through a real AI/NLP revolution
  - …because most existing NLP courses
    - (1) either cover only basics (traditional NLP)
    - (2) or are outdated (focus on models that are no longer relevant)

- The course is not just about **multilingual** NLP
  - It's about **state-of-the-art** NLP
  - From LM-pretraining and Transformer architecture to instruction-based fine-tuning and Large Language Models (LLMs)

Course description

- **Why this course?**

Model: GPT-4

I'm doing a master-level course on multilingual natural language processing. I'm trying to convince students that such a course is extremely useful, especially at this point in time. What are the best arguments to convince them?

Multilingual natural language processing (NLP) is an increasingly important field in the era of global communication and interconnectedness. Here are some key arguments to convince students of the value of a master-level course in this area:

1. Growing need for multilingual solutions: As more people access the internet and communicate across borders, there is an increasing demand for NLP solutions that can handle multiple languages. A course on multilingual NLP can equip students with the skills and knowledge to address these needs and contribute to the development of cutting-edge technologies.
2. Economic and business opportunities: Businesses and organizations are expanding their operations globally, making multilingual NLP a valuable skill set. Students with expertise in this field will be highly sought-after, creating numerous career opportunities in industries such as technology, finance, marketing, and more.

Course description

- **Why this course?**

3. Cross-cultural understanding: Multilingual NLP can help bridge the gap between different cultures and languages, fostering better communication, collaboration, and understanding. This not only benefits individuals and organizations but also has broader societal implications.
4. Improvement in machine translation: As multilingual NLP improves, so does the quality of machine translation. Students with expertise in this area can contribute to developing more accurate and efficient translation systems, benefiting countless people worldwide.
5. Transfer learning and resource sharing: Multilingual NLP research has led to the development of transfer learning techniques, which enable the sharing of resources and knowledge across languages. This can significantly reduce the time and effort required to

Course description

- **Why this course?**

6. Linguistic diversity: By studying multilingual NLP, students will gain a deeper understanding of the world's linguistic diversity and the challenges associated with it. This knowledge can be used to develop more inclusive technologies and policies that support language preservation and revitalization.
7. Competitive advantage: As the demand for multilingual NLP expertise grows, having this skill set can provide a significant competitive advantage in the job market. Students who master this field will be better positioned to secure high-paying and rewarding roles in a variety of sectors.
8. Interdisciplinary nature: Multilingual NLP is an interdisciplinary field, drawing on knowledge from computer science, linguistics, and psychology, among others. Students will develop a diverse skill set and a well-rounded understanding of the complexities of human language and communication.

Course description

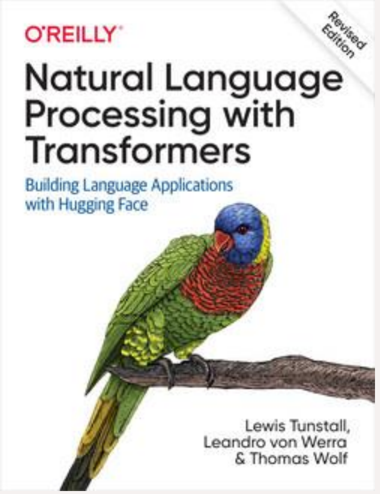
- **Target audience**: students who want to
  - Gain theoretical understanding of state-of-the-art NLP
  - Obtain practical hands-on experience implementing, training, and evaluating modern neural NLP models

- **Prerequisites** (recommended, not mandatory)
  - Linear algebra, probability theory (intermediate)
  - Machine learning (intermediate)
  - (Traditional) Natural language processing (basic)

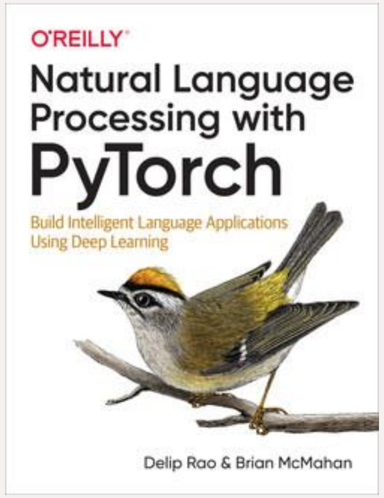
  - Python programming skills (intermediate to advanced)
    - Practical parts of this course will be in PyTorch

Textbooks

- Natural Language Processing with Transformers
  Lewis Tunstall, Leandro von Werra, Thomas Wolf



- Natural Language Processing with PyTorch
  Delip Rao, Brian McMahan



Content

- Why Multilingual NLP?
- Linguistic Diversity and Universals
- Language Identification

—

- About the mNLP Course
- **Topics and Schedule**
- Organization

Topics and Schedule

- **Lecture 01**: Language Diversity & Course Organization (Apr 17)
- **Lecture 02**: Neural Language Modeling & Tokenization (Apr 24)
  - History of Neural LMs
  - Subword Tokenization Algorithms
- **Lecture 03**: Training and Optimization (May 15)
  - Unifying NLP with Neural LMs
  - Gradient Descent and Backpropagation, Adaptive Optimization
  - Dropout

Topics and Schedule

- **Lecture 04**: Transformer Almighty (May 22)
  - Attention Mechanism
  - Dissecting the Transformer Architecture
  - Pretraining-Fine-Tuning Paradigm

- **Lecture 05**: Multilingual Word Representations (May 29, Online)
  - Joint Bi/Multilingual Embeddings
  - Projection-Based Cross-Lingual Word Embeddings (CLWEs)
  - Unsupervised CLWEs
  - Evaluating Multilingual Word Representations

Topics and Schedule

- **Lecture 06**: Multilingual Language Models (June 12)
  - Pretraining Multilingual LMs
  - Cross-Lingual Transfer with Multilingual LMs
  - Zero-Shot vs. Few-Shot Cross-Lingual Transfer
  - Evaluation (Tasks and Benchmarks) in CL Transfer

- **Lecture 07**: Modularization and Adaptation (June 19)
  - Curse of Multilinguality
  - Post-Hoc Adaptation of Language Models
  - Modularization and Parameter-Efficient Fine-Tuning

Topics and Schedule

- **Lecture 08**: Cross-Lingual Transfer for Token-Level Tasks (June 26)
  - Word Alignment Methods
  - Cross-Lingual Transfer with Label Projection

- **Lecture 09**: Neural Machine Translation (July 3)
  - Encoder-Decoder NMT
  - Decoder-Only NMT
  - Massively Multilingual (N-to-N) NMT
  - MT Evaluation

Topics and Schedule

- **Lecture 10**: Multilingual Sentence Encoders (July 10)
  - From BERT to Sentence-BERT
  - From Sentence-BERT to Multilingual Sentence BERT
  - Supervised Training of Sentence Encoders
  - Self-Supervised Training of Sentence Encoders

- **Lecture 11**: LLMs, Instruction-Tuning and Generative NLP (July 17)
  - Large Language Models
  - Prompting and In-Context Learning
  - Instruction-Tuning

Content

- Why Multilingual NLP?
- Linguistic Diversity and Universals
- Language Identification

—

- About the mNLP Course
- Topics and Schedule
- **Organization**

Exercises

- In the **first half of the course**, TAs will teach you how to implement, train, and evaluate SotA NLP models in PyTorch
    - Session #1: Crash Course in PyTorch
    - Session #2: Implement, Train, Evaluate w. PyTorch Lightning
    - Session #3: Training LMs w. HuggingFace Transformers
    - Session #4: Presentation of project topics
    - Session #5: Cross-lingual transfer with multilingual LMs

Projects

- In the **second half of the course**, you will work on a small-scale project yourself: optional, the motivation is **exam bonus**
- Projects are to be carried out in teams of (up to) 3 students
- Example topics:
  - Bilingual specialization of a multilingual LM
  - Cross-lingual transfer to a language with script unseen in pretraining
  - Parameter-efficient fine-tuning (e.g., adapters or low-rank adaptation)
- Project presentations: last exercise session (July 19)

Paper Reading Homeworks

- For lectures L7-L10, you will be provided relevant (and recent) **research papers** on the topic of the lecture and a set of questions
  - L7: Paramerer-Efficient Fine-Tuning
  - L8: Word Alignment from Parallel Data
  - L9: Massively Multilingual NMT
  - L10: Multilingual Sentence Encoders
- **Homework** (individual): submit answers to questions
  - Grading: binary – pass or fail (0 or 1 point)

Exam (Bonus)

- Increase your (passing) exam grade by one (e.g., from 2.0 to 1.7)

- Exam bonus is earned combined from reading homeworks and projects
  - Each reading homework: 1 point, so max. 4 points in total
  - Projects will be graded on a 4-grade scale: 0 to 3 points

- At least **5** (of max. 7) bonus points needed for the **exam bonus**

- **Exam format**
  - Written: 90 minutes with both practical and theoretical problems

Image: Alexander Mikhalchyk