Multilingual NLP and resources

Linguistic Resources for Natural Language Processing
LM Language Technologies and Digital Humanities
2024-25

Cristina Bosco

Overview

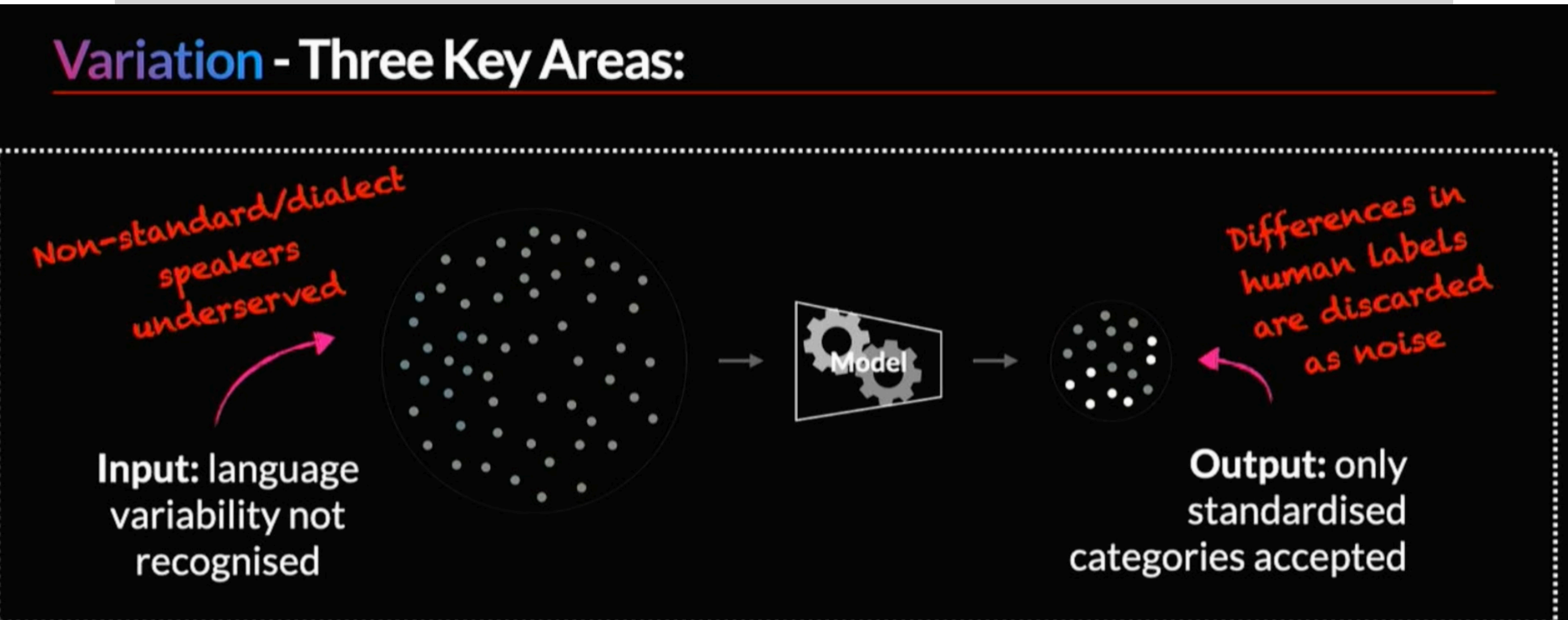
- Language and variation
- Language and languages
- Language and dialects
- Why we need multilingual NLP?
- Challenges for multilingual NLP
- Solutions for multilingual NLP

Human language is inherently featured by variation

Variation occurs in several different forms and according to several dimensions.

Language variation

Variation - Three Key Areas:



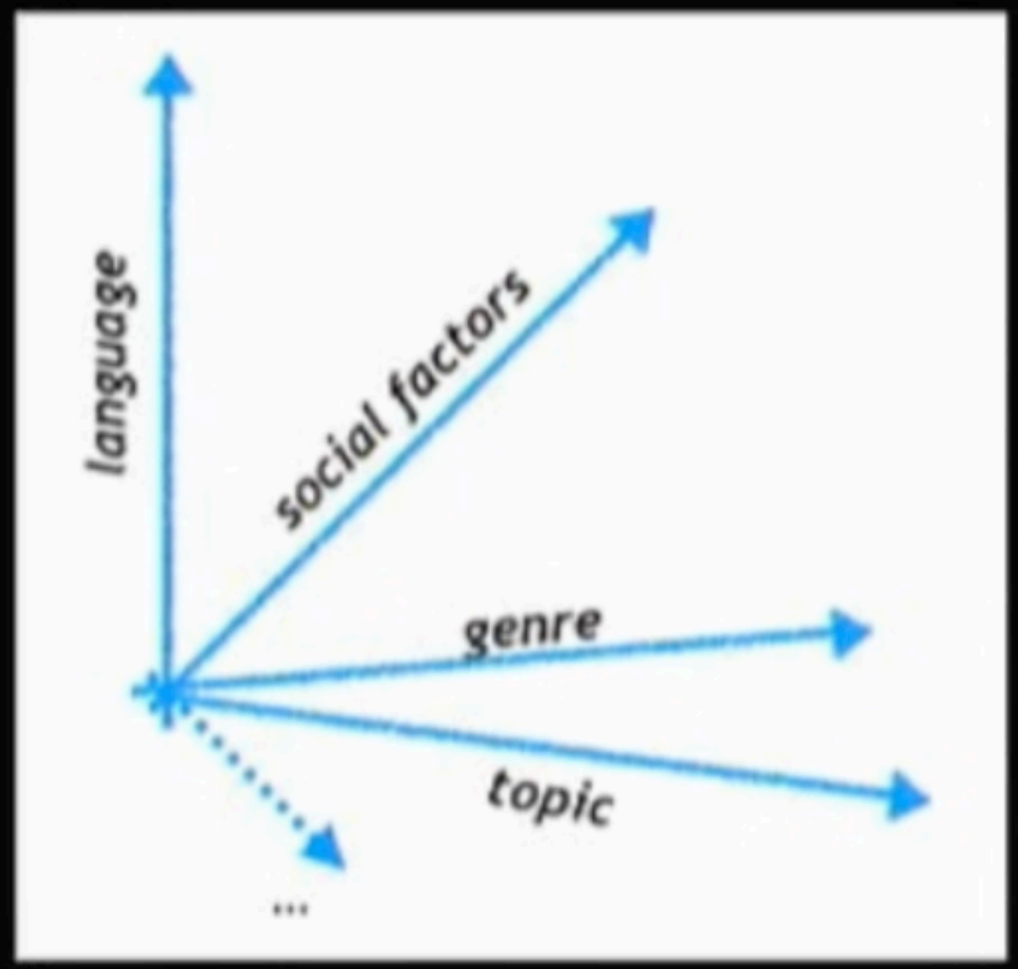
Research: Human-centric perspective & research diversity

Embracing Variation Holistically for Trustworthy NLP

38

Credits: B. Plank - ACL24

Variety Space.
Plank (2016)



What is variation?

Several languages exist, which can be **written**, **spoken** or **signed**.

They can be classified as *national languages* or *dialects* or in other ways, but in many cases this is a political rather than a linguistic distinction. From a computational perspective, the things that matter are **similarity** and **variation**, that occur among different languages or in a single language.

But the evaluation of similarity and variation has to be reconsidered in a non-English centred perspective.

What is variation?

All the dimensions of variation pose challenges for Natural NLP.

Variation is intrinsic to human language and it is manifested in four main ways:

- ***diaphasic*** variation is related to the setting or the medium of communication. For example, different styles and registers, oral versus written versus signed language, newspapers versus social media, "the medium is the message" (McLuhan).

What is variation?

- ***diastratic*** variation is related to language variation in different social groups. For example, age and gender introduces variations in the usage of the same language that can be also automatically detected by *author profiling* tools.

What is variation?

- ***diachronic*** variation is related to language variation across time. For example, neologisms can be introduced in a language, or some term meaning shift can occur. This dimension of variation is well attested in historical linguistics and for studying it monitor corpora are collected.

What is variation?

- ***diatopic*** variation is language variation in space such as different dialects or national varieties of the same languages. For example, British English and American English, Canadian French and French from France.

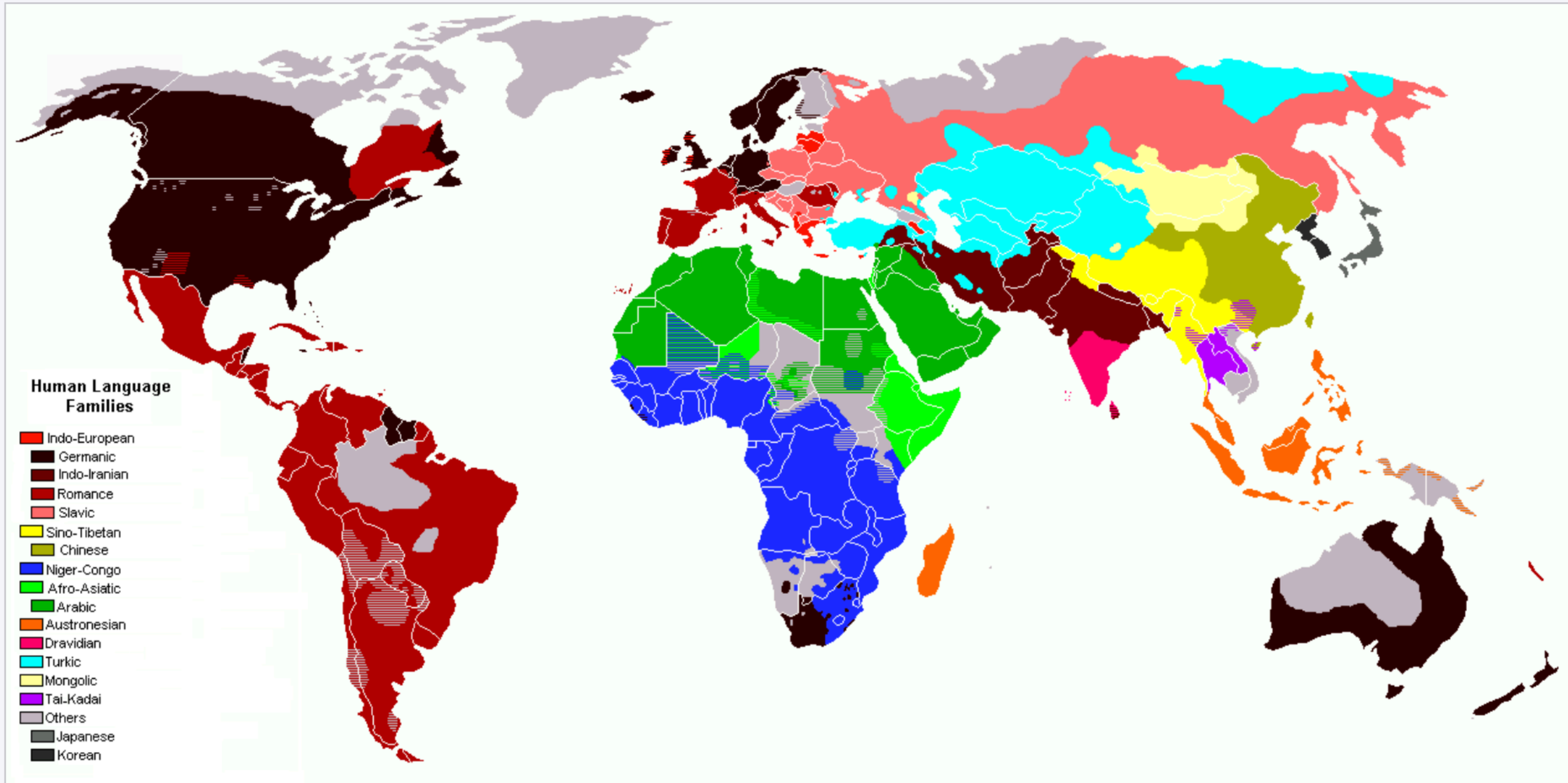
In the last few years several studied about NLP were mostly focused on diatopic variation.

English and the others

From its infancy, for historical and economic motivations, NLP has been focused on English, but several other languages exist that must be dealt by NLP techniques.

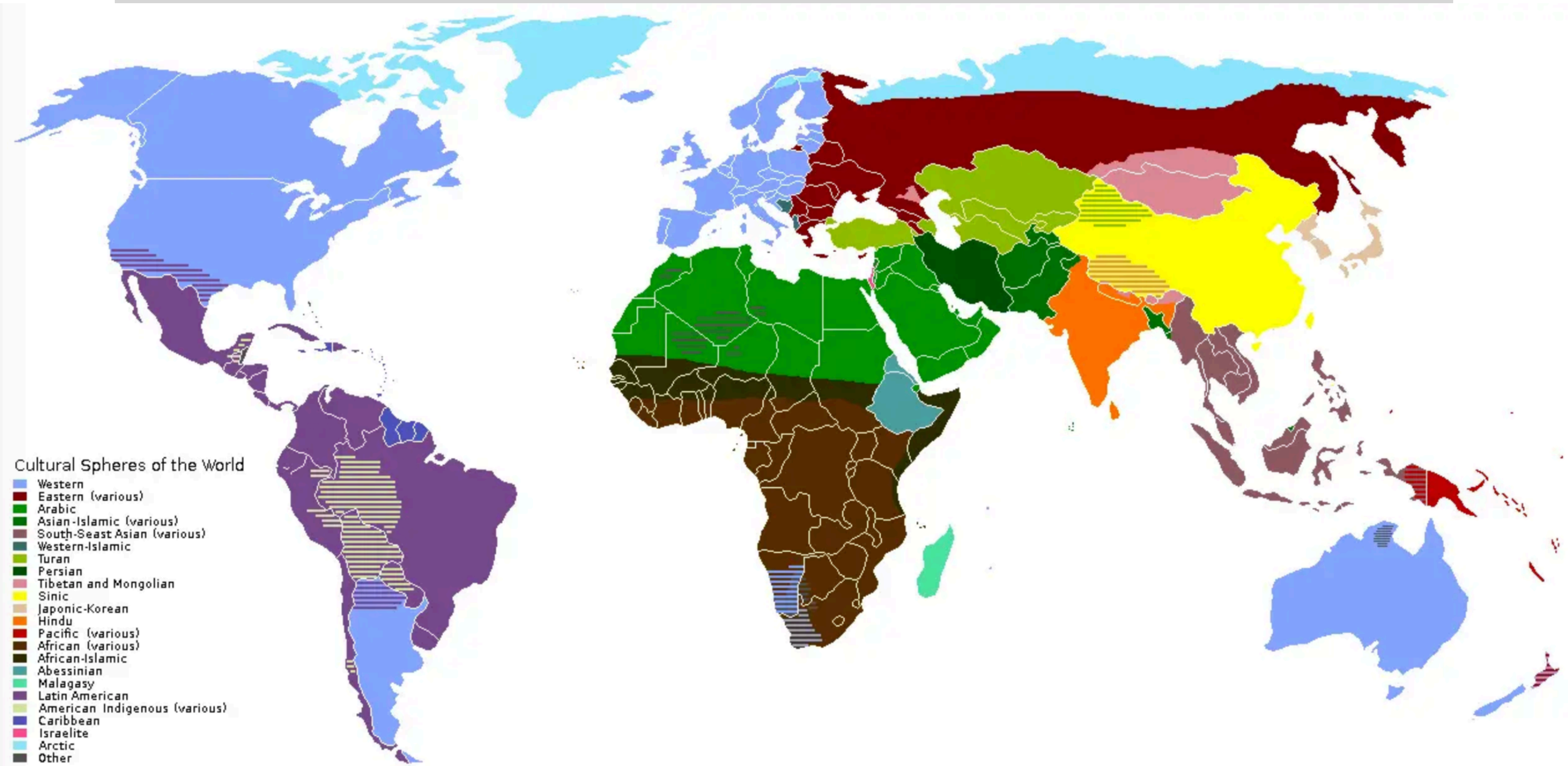
What are the ***other languages*** ?

Language and languages



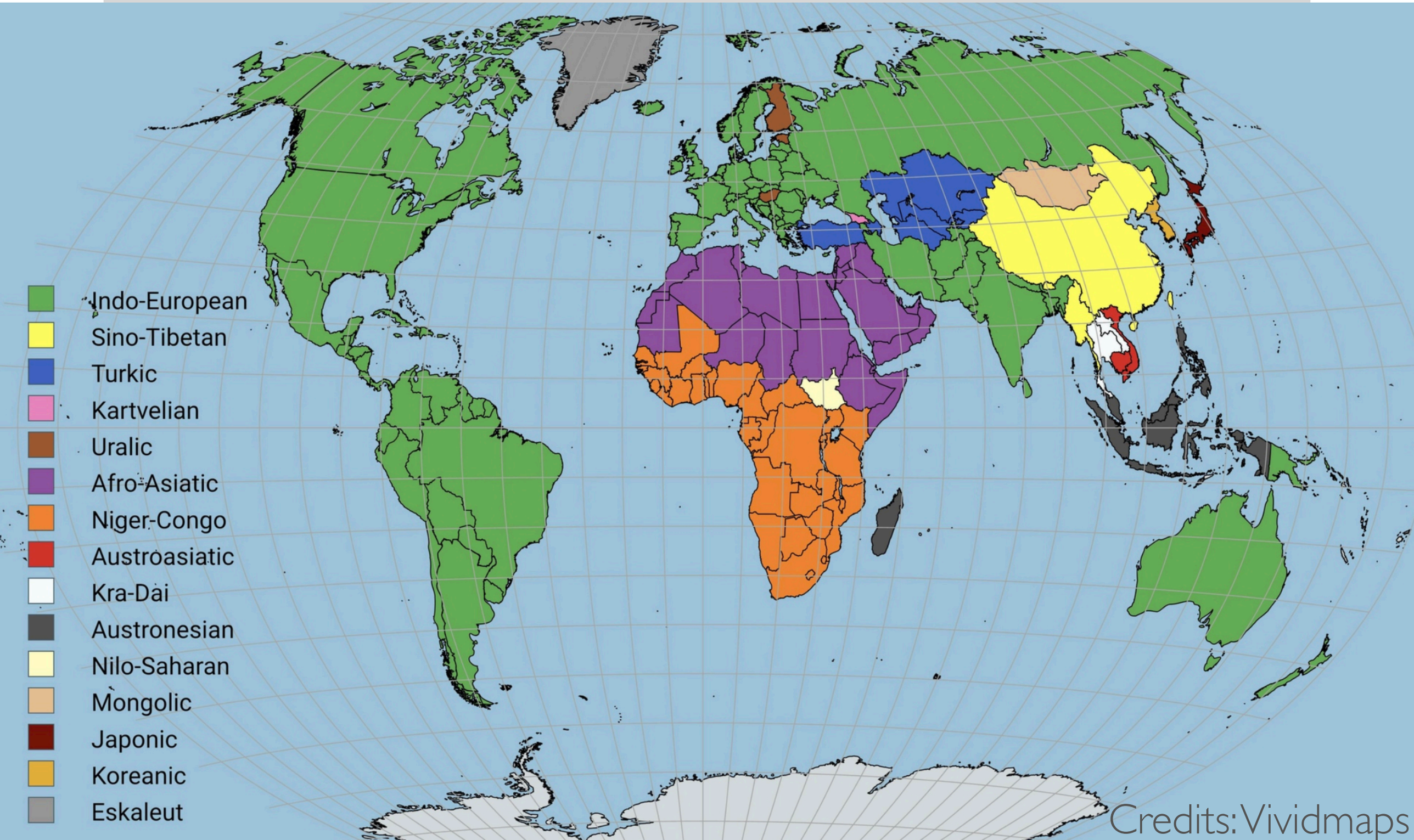
Credits: Vividmaps

Language and languages



Credits: Vividmaps

Language and languages



Credits: Vividmaps

Language and languages

There are over **7,000 languages** in the world:
it's extremely challenging to develop NLP for them all!

NLP innovations so far have a strict **focus on** a single language:
the vast majority of technological advances have been achieved in **English-based NLP** systems.

Language and languages

The vast majority of NLP development refers only to some languages:
English (369), Chinese (921), Urdu (69), Arabic (274), French (79) and Spanish (471).

This set of languages does not include some of the most used, such as Hindi (342), Portuguese (232), Bengali (228), Russian (153) and Japanese (123).

[All these numbers refer to million of speakers]

Language and languages

There is a digital divide in the field of NLP between **high resource** and **low resource languages**.

High resource languages constitute a short list starting with English, (Mandarin) Chinese, Arabic, French, German, Portuguese, Spanish and Finnish.
These languages have large, accessible collections of digitized text, large collections of recorded speech (these are all spoken, not signed languages) much of which has been transcribed, as well as annotated resources such as treebanks and evaluation sets for a large number of NLP tasks and NLP tools.

Language and languages

As of August 2019, the **Language Resources Evaluation** (LRE) **Map** lists:
- 961 resources for English and 121 for American English
- 216 for German
- 180 for French
- 130 for Spanish
- 103 for Mandarin Chinese
- 103 for Japanese.

The only other languages with more than 50 resources listed are: Portuguese, Italian, Dutch, Standard Arabic, and Czech.

The remainder of the world's ~7,000 languages have far fewer resources!

Language and languages

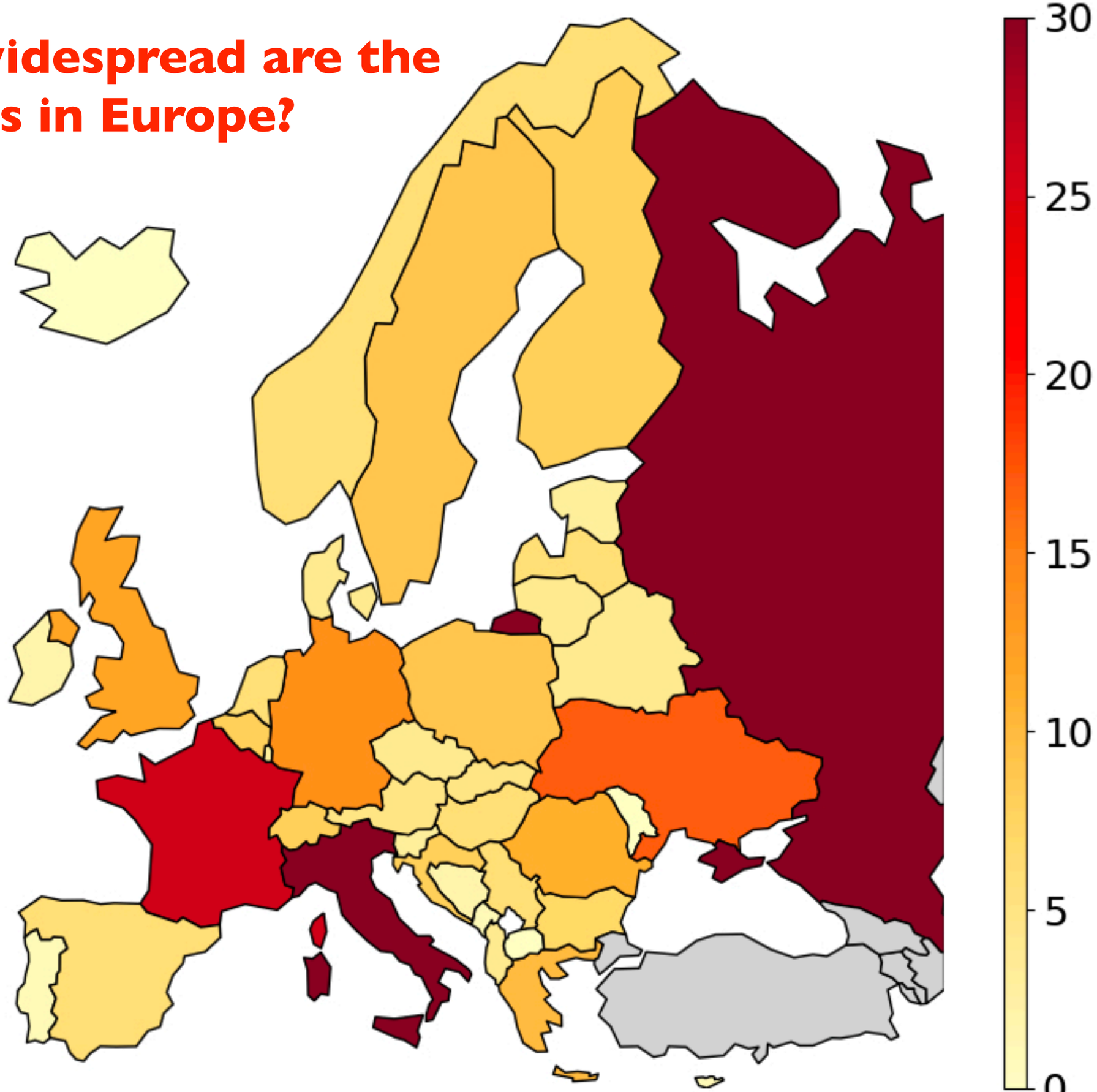
Among the most used languages there are those derived from a diatopic variation process of English, Spanish, Portuguese and French.

They are **national varieties** of *pluricentric* languages, such as Canadian French or Brasilian Portuguese, with **multiple interacting standard forms** in different countries.

Language and languages

There has been a lot of recent interest in the NLP community in the computational processing of **dialects**, with the aim to improve the performance of applications such as machine translation, speech recognition, and dialogue systems, but also to preserve them for the future generations.

How widespread are the dialects in Europe?



Language and dialects

In Europe several languages and dialects are at risk of disappearing, in particular in Italy.

While most Italian dialects have less than 1 million speakers and are definitely or severely endangered, some are still used even by younger generations in informal settings as a way to signal their social identities, i.e., language varieties spoken in the South and North-East areas of the Italian peninsula.
Just like most languages of the world, local languages and dialects of Italy are primarily used in spoken contexts, and only a fraction of them have a recently established written form.

Language and dialects

"Italy holds especial treasures for linguists. There is probably no other area in Europe in which such a profusion of linguistic variation is concentrated into so small a geographical area." (Maiden and Parry, 1997)

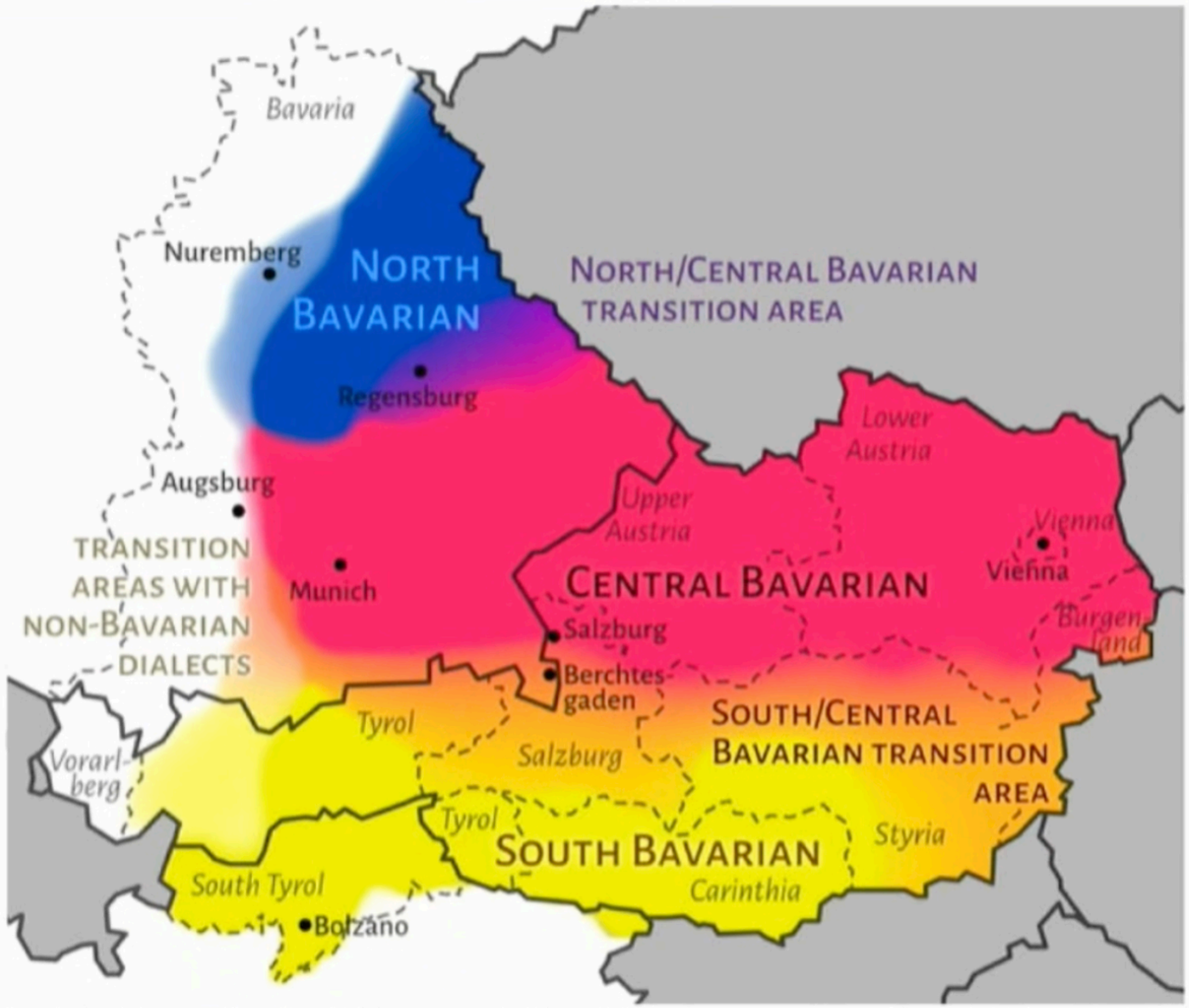
Italy is characterized by a one-of-a-kind linguistic diversity landscape in Europe, which implicitly encodes local knowledge, cultural traditions, artistic expression, and history of its speakers. However, according to UNESCO reports, over 30 language varieties in Italy are at risk of disappearing within few generations.

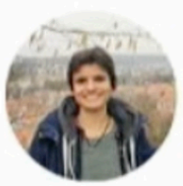
Language and dialects

Language Variation Within A Language Variety

- Away from Modeling Languages as Monoliths:
  - Example: German
  - Bavarian: More Variation!

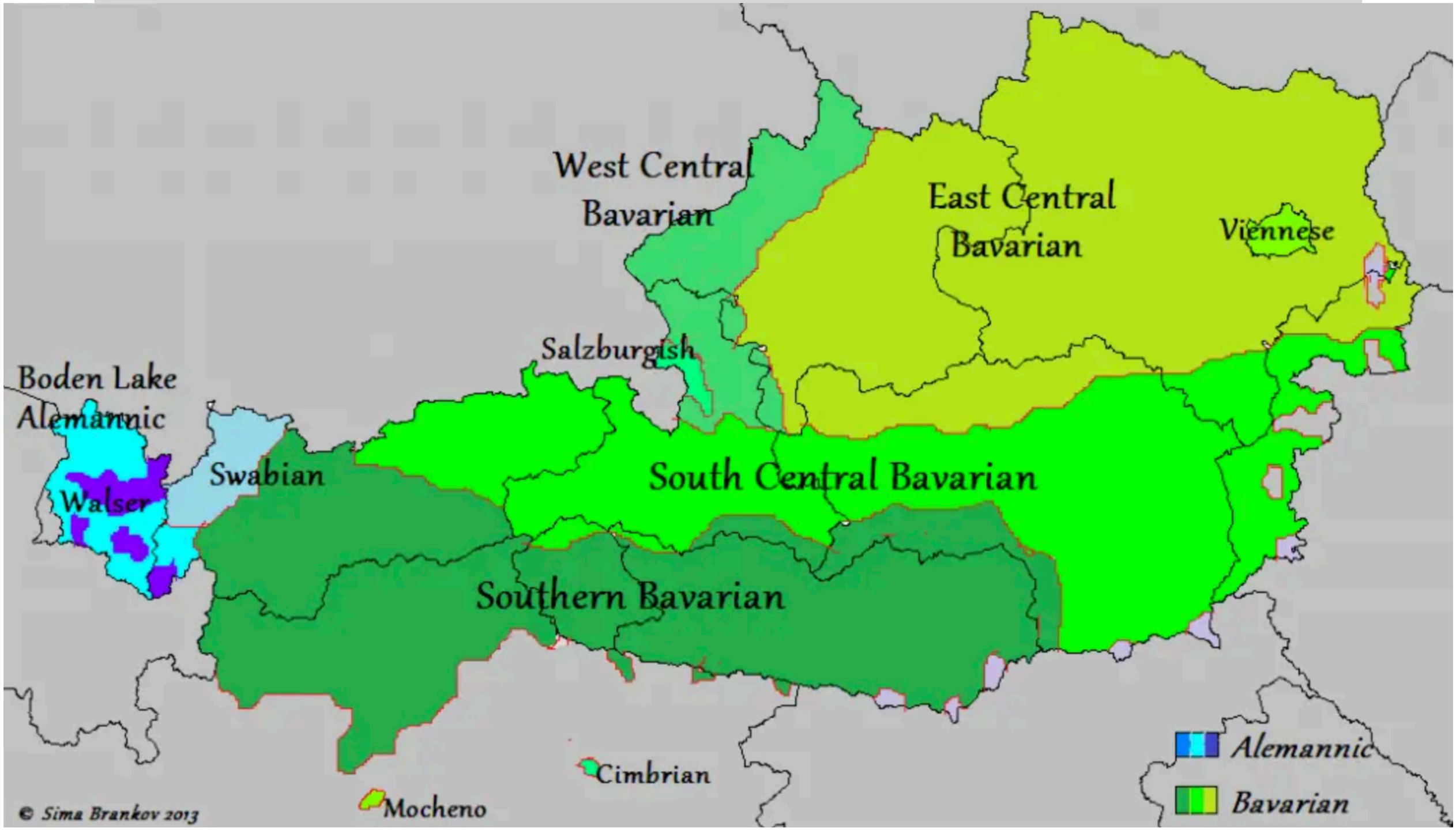
- **Opportunity**: Embrace Within-Language Variation in Multilingual NLP



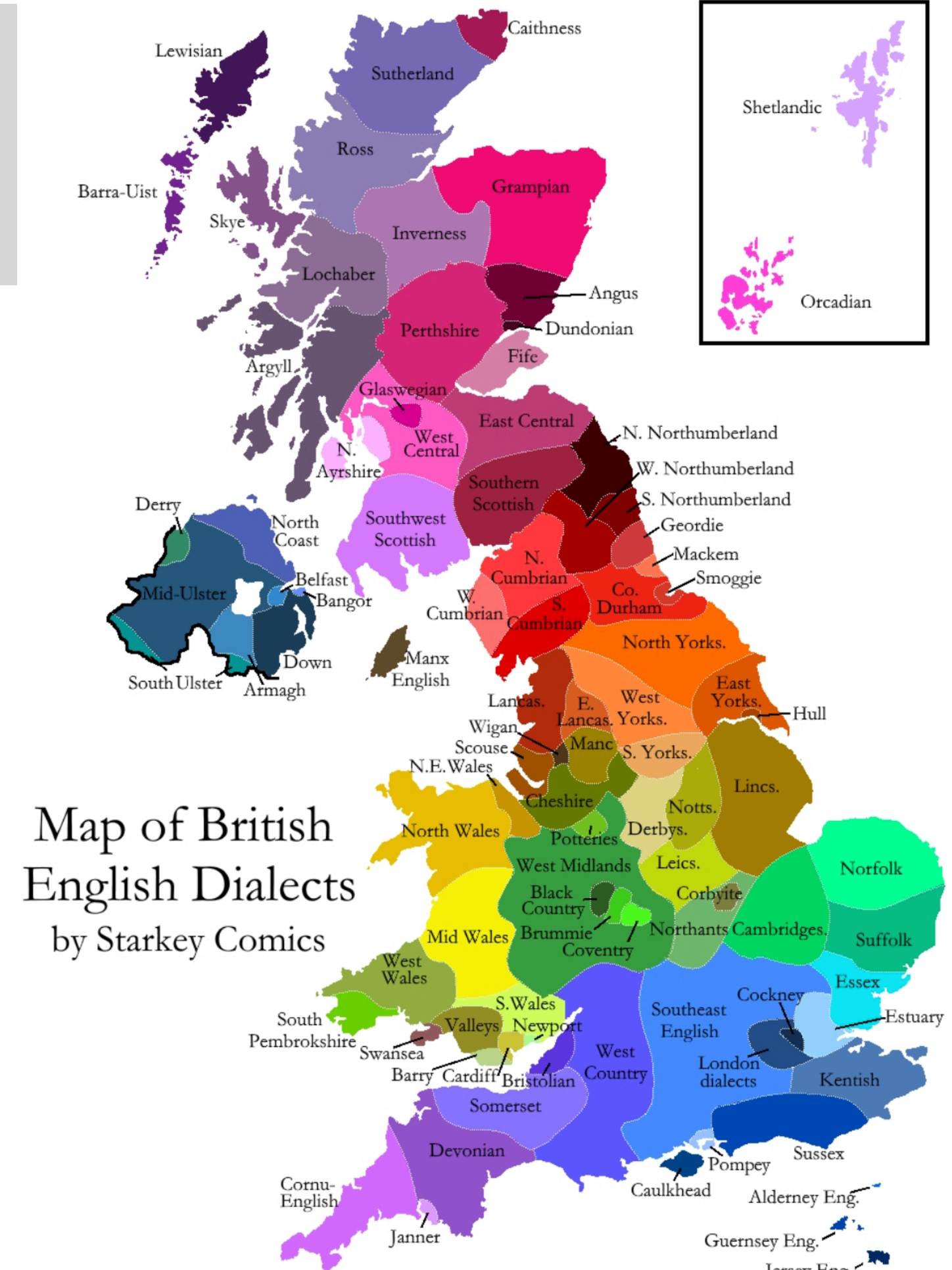


Blaschke, Kovačić, Peng, Schütze, Plank. MaiBaam: A multi-dialectal Bavarian Universal Dependency treebank. LREC-COLING 2024. 46

Language and dialects



Language and dialects



How many languages? How many speakers?

There are approximately 7,000 languages believed to be spoken around the world.

– The three largest language groups (Mandarin, Spanish, and English) have more than 1.5 billion native speakers.

– 96 per cent of all languages are spoken by only three to four per cent of all people.

– 2,000 of the world's languages have less than 1,000 native speakers.

– The Ethnologue database lists exactly 7,099 individual languages in a comprehensive geographic database.

– The highest language diversity in the world can be found in Africa and Asia, both with more than 2,000 living tongues.

– Europe with only around 250 living languages and dialects spoken.

– 50 to 90 per cent of the currently spoken languages could be extinct.

Why we need multilingual NLP?

Several languages are used all over the world, but NLP techniques are suitable for only a few of them.

Although only a few people can benefit from these techniques, they strongly condition the communication among everybody over the world.

Why we need multilingual NLP?

The most important motivations for making available NLP for virtually all existing languages are:

1. To contrast the reinforcement of social disadvantages
2. To limit normative biases in NLP
3. To preserve endangered languages

1. NLP can reinforce social disadvantage

1. The lack of NLP for several people reinforces their social disadvantage:

Technology is only accessible for you when its tools are available in your language.

For example, the lack of spell-checkers for a language impairs those who speak and write it; OCR (optical character recognition) is still limited for non-English languages, making harder and slower the collection of data for them.

1. NLP can reinforce social disadvantage

Psychological research has shown that **the language you speak modifies the way you think**.

The Internet inherently incorporates the societal norms of the *driving* languages, and in particular English, making everyday increasingly more challenging to introduce new aspects to this deeply ingrained communication mechanism.
As NLP continues to develop without bringing in a diverse range of languages, it will be more difficult to incorporate them, endangering the global variety of languages.

1. NLP can reinforce social disadvantage

The promise of language technology includes **pro-social applications** such as:
- biomedical applications (e.g. matching patients to research studies or automatically flagging patients for time-sensitive tests based on physician notes)
- machine translation of documents available on the web
- interactive tutoring for language learning and other learning scenarios
- …

These benefits should be *available to all*.

1. NLP can reinforce social disadvantage

The existence of even the most basic language technology (e.g. keyboards or input systems supporting the writing system, spell checkers, web search tools) builds up the **value of a language** which can be an important factor in self-esteem and educational outcomes for speakers of minoritized languages and can contribute to the maintenance of languages under threat of displacement by local majority languages.

1. NLP can reinforce social disadvantage

Language technology **can reinforce and amplify racism**.

Each time a task is performed based on a supervised approach, groups of annotators are enrolled who annotate data. Annotated data became in turn the reference ground truth for the task.

For example, has been observed in hate speech detection tasks for English that tweets with features of African American language were more likely to be labeled as hate speech than other varieties of English.

1. NLP can reinforce social disadvantage

Language technology **can reinforce and amplify racism**.

This problem is not solved by using LLMs and retrained models, such as word embeddings.

They are indeed based on large collections of data that can be available only for dominant languages and mirroring the meaning of words that is encoded in texts by dominant people, their biases and stereotypes against minorities.

1. NLP can reinforce social disadvantage

Some very recent studies show how LLMs are contributing to the spread of racism in the context of **healthcare and health professions.**

LLMs are rapidly being integrated into clinical practices:

- LLM-based pilot programmes are underway in hospitals for automating administrative or documentation tasks

- clinicians have begun using ChatGPT to communicate with patients and draft clinical notes, but also to find support in their diagnoses and decisions about therapies.

1. NLP can reinforce social disadvantage

The recent study "Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study" (https://www.sciencedirect.com/science/article/pii/S258975002300225X) co-authored by D. Jurafsky

shows that GPT-4 did not appropriately model the demographic diversity of medical conditions, consistently producing clinical reports based on stereotypes about demographic features of patients.

The diagnoses created by GPT-4 for clinical cases are based on stereotype about certain races, ethnicities, and genders, showing a significant association between demographic attributes and recommendations for more advanced (and expensive) procedures.

2. NLP encodes normative biases based on English

2. The availability of tools only English-centered introduces normative biases in NLP:

English and English-adjacent languages have **specific grammatical structures that are not representative** of other languages. By supporting mostly those languages, the NLP technologies are progressively treating English as the normal and default language on which all settings are based.

For example, some Part of Speech tag set or parsing approach created for English has been applied on other languages.

2. NLP encodes normative biases based on English

The field of NLP is caught in a negative feedback loop that hinders the expansion of the languages we work on.

Work on languages other than English is often considered by the most of researchers as "*language specific*" and thus reviewed and considered as less important than equivalent work on English.

2. NLP encodes normative biases based on English

On the one hand, to scientists who work on a language are selected because their are mother tongue speakers of it.

On the other hand, the possibility to work on other languages are limited with respect to those for people working on English.

2. NLP encodes normative biases based on English

As a relatively language agnostic system is trained on English, it learns not only all its grammatical norms, but also the **cultural implications** that are closely related to it.

This can have a strong impact on all the intelligent processes that depend on language and follow its interpretation. The result can be a form of globalisation.

2. NLP encodes normative biases based on English

Phenomena not attested in English have been studied less than those occurring in this language.

For example, tokenization performs very poorly on languages that feature *reduplication*, which is common to many languages (Afrikaans, Irish, Punjabi, and Armenian), but quite scarcely attested in English or Romance languages.

- Reduplication is a process in which the root/stem of a word or the whole word is repeated exactly or with a slight change (boogie-woogie, willy-nilly; tran tran, piano piano).

2. NLP encodes normative biases based on English

Emily Bender provides a quick list of ways in which **English fails to represent all languages**, that is, properties of English that are not broadly shared, even among the world's widely used languages:

- *It's a spoken language, and not a signed language*.

Right off the bat, if we take only English, we've restricted our attention away from an important class of languages.

2. NLP encodes normative biases based on English

- *It has a well-established, long-used, roughly phone-based orthographic system*.
Phone-based means that the letters correspond to individual sounds. English orthography only approximates this principle. Other languages such as Spanish have much more transparently phone-based orthographies, still others represent only consonants (e.g. Hebrew and Arabic, traditionally) or have symbols which represent syllables rather than single sounds (e.g. Malayalam, Korean, or the Japanese kana), or use logographic systems (e.g. Chinese, or the sinographs borrowed into Japanese as kanji). And, many of the world's languages are not written, or are written but don't have a long tradition of being written and/or don't have standardized orthographies. We routinely underestimate how much standardization simplifies the task of NLP for English.

2. NLP encodes normative biases based on English

- *The standardized orthography for English provides a standardized notion of "word" indicated by whitespace*.
This isn't true for all languages, even those with standardized orthography. Many NLP systems for Chinese, Japanese, Thai and other languages have to start with the problem of word tokenization.

- *English writing uses (mostly) only lower-ascii characters found on every computer*.
For the most part, we don't have to worry about rarer character encodings, unsupported Unicode ranges, etc. when working with English.

2. NLP encodes normative biases based on English

- *English has relatively fixed word order.*

Compared to many languages in the world, English is rigid in its word order, insisting on subject-verb-object in most circumstances, adjectives before nouns but relative clauses after, etc. Without testing on more flexible word order languages, how can we know the extent to which systems rely on this property of English?

2. NLP encodes normative biases based on English

- *English forms might 'accidentally' match database field names, ontology entries, etc.*

Many language technologies achieve task-specific goals by mapping strings in the input language or transformations of those strings into syntactic or semantic representations to external knowledge bases. When the input strings and the field names or entries in the knowledge base are in the same language, processing shortcuts become available. But for how many languages is this true?

2. NLP encodes normative biases based on English

- *English has massive amounts of training data available*

If we focus all of our attention on methodologies which rely on amounts of training data that simply aren't available for most of the world's languages, how are we going to build systems that work for those other languages? Similarly, if we only value work that uses those technologies (e.g. in conference reviewing), how can we expect to make any progress on cross-linguistically useful NLP?

NLP and preserve endangered languages

3. To preserve endangered languages

For example, the collection and development of linguistic resources for dialects and language varieties can be a way to not miss a great cultural wealth and be able to study languages and dialects that we are taking.

Challeges in multilingual NLP

Several challenges must be faced to address multiple languages in NLP.

They meaningfully depend on the features of the language addressed.

Challenges in multilingual NLP

The past decades of NLP shows that the **performance** of NLP systems significantly **degrades** when faced with language variation. Ideally applications should be trained on resources that enable us to take into account different dimensions of variation in modelling language.

It is somewhat simplistic to assume that corpora could fully represent a language without considering variation. In corpus linguistics, researchers have tried to address variation and represent it in corpora. Language technologies are doing the same.

Challenges in multilingual NLP

There has been a lot of recent interest in the NLP community in the computational processing of language varieties and dialects, with the aim to improve the performance of applications such as machine translation, speech recognition, and dialogue systems.

But the progress in the field of NLP mostly depends on the existence of language resources.

Challenges in multilingual NLP

The collections of written, spoken or signed language are often associated with careful annotations reflecting the intended output of the NLP system for the task at hand.
For example, annotated corpora are used for speech recognition systems or for dialogue systems such as Siri, Alexa or Google Home.

The availability of annotated data is crucial.

Challenges in multilingual NLP

Acquiring text corpora for **dialects** is particularly challenging as dialects are typically vastly underrepresented in written text. The typical solution consists in producing text corpora by transcribing speech, after the collection of spoken dialectal data. The **transcription** can be done automatically, for example, using Automatic Speech Recognition, which was used to produce Arabic dialectal text corpora, or **manually**, which was used to build the ArchiMob corpus for (Swiss) German dialects. Alternative approaches to dialectal data collection include **social media**, for example, Twitter, and **translations**.

Challenges in multilingual NLP

The collection and annotation of corpora which include several different **variants of a language** is crucial for allowing the automatic identification of the varieties, i.e. a step to be introduced in the NLP pipeline.

For example, the Discriminating between Similar Language Corpus Collection (**DSLCC**), which includes data from several pairs or groups of:

- similar languages, such as Bulgarian and Macedonian, Czech and Slovak, Bosnian, Croatian, and Serbian, Malay and Indonesian
- different national variations of the same language, such as Brazilian and European Portuguese, British and American English, Argentinian and Peninsular Spanish.

Challenges in multilingual NLP

The DSLCC features **journalistic texts** collected from multiple newspapers in each target country.

This allows to alleviate potential topical and stylistic biases intrinsic to any newspaper, in order to prevent NLP systems from learning a specific newspaper's writing style as opposed to learning the language variety it represents.

Challenges in multilingual NLP

As in other similar project (see e.g. CORIS for Italian) newspaper are assumed as the most accurate representation of the contemporary written standard of a language and to represent national language varieties.

Other popular data sources, used in a variety of NLP tasks, such as **Wikipedia**, are not suited to serve as training data: Wikipedia is a collaborative resource, which allows speakers of multiple language varieties and non-native speakers to contribute to the same article.

Challenges in multilingual NLP

Unsupervised, weakly supervised, semi-supervised, or distantly supervised machine learning techniques can reduce the overall dependence on labeled data for **training**.

But even with such approaches, there is a need for both sufficient labeled data to **evaluate** and **testing** system performance and typically much larger collections of unlabeled data to support the very data-hungry machine learning techniques.

Challenges in multilingual NLP

For **signed** languages, the difficulty is related to the multimodality involved in the communication process.

The translation of a signed language to a spoken or written language and viceversa can be strategic for allowing a higher degree of inclusion of the deaf community in educational, health and social contexts, in practice, in everyday life.

Challenges in multilingual NLP

For example, for the Italian Sign Language or the American Sign Language, the lack of standardized transcription methodologies for generating the input for NLP tools makes difficult the application of NLP and thwarts the efforts of researchers studying the languages of the deaf.

For **Italian Sign Language** (LIS), some effort has been done in the development of tools for translating from a written language to LIS in small domains. These projects include the development of an avatar that signs information about train schedule in a train station or the weather information in a TV news program (see http://www.crit.rai.it/eletel/2009-2/92-02.pdf)

Challenges in multilingual NLP

Another interesting and challenging phenomenon to be taken into account is **code-mixing** (also called *code-switching* or *language alternation*).
It occurs when a speaker alternates between two or more languages or language varieties in the context of a single conversation or situation.
It is often used by multilingual speakers, but also by single language speakers in particular communication contexts and dynamics, such as social media.

Challenges in multilingual NLP

For dealing with code mixing, NLP tools must be informed about the languages they have to deal with, this means that the recognition of the languages must precede other analyses. Language identification tools are therefore especially useful for the analysis of texts where code mixing occurs.