Multilingual NLP

The most exciting topic in NLP

Guillaume Wisniewski

guillaume.wisniewski@u-paris.fr

UPCité Master in Computational Linguistics

September 23rd, 2025

Some Technical Details

Lecture Website

- `https://pages.llf-paris.fr/~gwisniewski/?teaching/multi` or `https://lc.cx/T0k1OS`
- look regularly: I will post deadlines ⊕ lecture cancellation on this page

Evaluation

- labs (≃ 5 or 6 labs)
- final exam (short questions with short answers)

Outline of the Lecture (tentative)

1. Linguistic Typology & NLP / Data Analysis
2. Multilingual Models
3. Developping NLP Models for all Languages
4. Machine Translation

Motivation for this lecture

There are X languages in this world...

Part I

Why Multilinguality is Important?

Babel Unbound: Discovering the Wonders of Many Languages

- The most beautiful way to describe translation (IMHO my opinion is, of course, not biased).
- How fascinating linguistic differences can be

↪ e.g. how to translate "meet on the 3rd floor" into Chinese?

- More fundamentally: the best way to check you've understood a language or a sentence ⟹ re-express it in another way.

Because I find the example fascinating...

Physical building level	Chinese	French	British English	American English
"Street Level"	一楼 (yī lóu) = 1楼	Rez-de-chaussée	Ground floor	1st floor
1 level above	二楼 (èr lóu) = 2楼	1er étage	1st floor	2nd floor
2 level above	三楼 (sān lóu) = 3楼	2e étage	2nd floor	3rd floor
3 level above	四楼 (sì lóu) = 4楼	3e étage	3rd floor	4th floor

⇒ having a machine that can translate would be awsome!

Machine Translation

- Machine Translation was "invented" before NLP

↪ a fascinating history

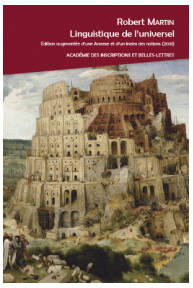
- Modern NLP methods were first developed to solve MT:
  - Early statistical models: IBM Models (≃90s)
  - Attention (pre-Transformers!)
  - Transformers
  - Subword tokenisation (e.g. BPE)
  - Synthetic data (back-translation)
  - Large-scale training

⇒ still a hot-topic to test the limits/linguistics capacities of LLMs

More Fundamentally: Why Focusing on More Than One Language Matters in Linguistics

- Broader Theoretical Insights Studying multiple languages prevents theories from being biased by a single linguistic system and helps reveal universal vs. language-specific features. (Evans and Levinson 2009)
- Better Understanding of Language Diversity Comparative work across languages uncovers typological patterns and challenges assumptions based on one language alone. (Haspelmath 2010)
- Improved Methodological Rigor Analysing different languages strengthens cross-linguistic validity and reduces ethnocentric perspectives. (Dryer and Haspelmath 2013)

Linguistic Universals



- one of the most fascinating ideas in linguistics (personnal opinion)
- Linguistic universals: features or principles shared by all human languages, revealing the common structures and constraints underlying human communication (Greenberg 1963).

↪ help identify the shared features of human languages, offering insights into cognition, communication, and the limits of linguistic variation.

Computational Linguistic Typology: Databases for Cross-Linguistic Comparison

Databases for Cross-Linguistic Comparison

- `WALS`: World Atlas of Language Structures (typological features)
- `APiCS`: Atlas of Pidgin and Creole Structures
- `PHOIBLE`: Phonetics and phoneme inventories
- `Glottolog`: Classification and bibliographic database
- `Lexibank` / `IDS` / `ASJP`: Cross-linguistic lexical datasets

Why these databases matter

- Enable large-scale, systematic comparison across the world's languages
- Provide typological, phonological, lexical and sociolinguistic features
- Support research in language typology, historical linguistics, NLP, and endangered language documentation

Example of Work

(Skirgård et al. 2023)

What is Grambank?

- A large-scale comparative database of grammatical features across the world's languages.
- Coverage: 2,467 varieties, 215 families, 101 isolates.
- 195 core features (mostly binary): clause, nominal, pronominal, verbal domains.
- Much higher per-language completeness than WALS (24% vs. 84% missing data in comparison reported by authors).
- Purpose: enable robust tests of genealogy vs. geography, constraints on grammar, identification of unusual languages, and impact of language loss.

Methods ("high level")

- Data curation: merge dialects to one variety per language; binarise multistate features; crop high-missing items; impute residual missingness.
- Spatio-phylogenetic modelling: approximate Bayesian inference for latent Gaussian models to estimate separate effects of phylogeny vs. space on each feature.
- Dimensionality reduction: PCA on the binary feature matrix to obtain a *design space* of grammars and interpret principal components.

Results I: Genealogy, design space

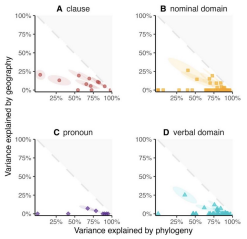


Fig. 1. Variance explained by phylogeny and geography. Each point is a Grambank feature. The panels represent different domains of grammar that the features are associated with: (**A**) clausal, (**B**) nominal domain, (**C**) pronominal domain, and (**D**) verbal domain. A high value indicates that a large part of the variance is explained by either space (*y* axis) or phylogeny (*x* axis). The ellipses represent the standard deviation of the joint posterior, tilted for the covariance.

- **Genealogy dominates** geography in explaining grammatical variation across features (mean phylogeny ≈ 0.72 vs. mean space ≈ 0.03).
- No reliable differences across grammatical domains in phylogenetic/spatial effects.
- PCA: optimal **19 components** explain ~49% of variance; first three only ~21% ⇒ grammars are neither tightly constrained nor unconstrained.
- Visualising similarity with PCs shows broad areal/family patterning in grammar.

Results I: Genealogy, design space

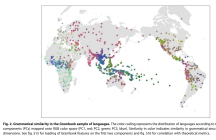
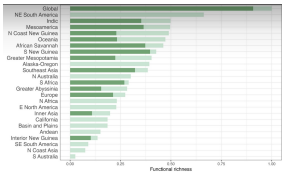


Fig. 2. Grammatical similarity in the Grambank sample of languages. The color coding represents the distribution of languages according to the first three principal components (PC) mapped onto RGB color space (PC1, red; PC2, green; PC3, blue). Similarity in color indicates similarity in grammatical structure via the first three dimensions. See Fig. S15 for loading of Grambank features on the first two components and Fig. S4 for correlation with (feature) p-models.

- **Genealogy dominates** geography in explaining grammatical variation across features (mean phylogeny ≈ 0.72 vs. mean space ≈ 0.03).
- No reliable differences across grammatical domains in phylogenetic/spatial effects.
- PCA: optimal **19 components** explain ~49% of variance; first three only ~21% ⇒ grammars are neither tightly constrained nor unconstrained.
- Visualising similarity with PCs shows broad areal/family patterning in grammar.

Results II: Unusualness & language loss



Bars representing functional richness relative to the current diversity of the world's languages are shown in light green, and functional richness of nonthreatened languages in the same areas are shown in dark green. Functional richness declines in all areas, with some regions showing dramatic decreases.

- Unusualness: isolates are ~ 4% of sample but ~ 19% of the most unusual languages; unusualness shows regional patterning (e.g., higher in N. Africa/Europe).
- Uniqueness: only five pairs of languages share the exact same Grambank profile — each language encodes a largely *unique* grammatical signature.
- Functional richness & endangerment: globally moderate decline if threatened languages vanish, but highly uneven regionally; some areas (e.g., NE South America, Alaska–Oregon, N. Australia) would drop to zero functional richness.

Working with Corpora

`https://universaldependencies.org/`

- ≥ 200 treebanks, ≥ 150 languages, ≥ 600 contributors
- same annotation guides (PoS tagging ⊕ dependencies) for all languages

↪ allow for cross-linguistic comparisons

1st example

(brosa-rodriguez-kahane-2024-new)

- Original Universal 14 (Greenberg, 1963):

↪ In conditional clauses, the protasis ("if ...") tends to precede the apodosis ("... then ...").

- Aim of the paper: quantitatively test and reformulate this universal across many languages.
- Data: Universal Dependencies v2.11 (150+ languages, 200+ treebanks).
- Method:
  - Identify "if"-like markers using PanLex.
  - Use Grew-Match to detect clause order (protasis vs. apodosis).
  - Define a dominant order when >66.6% of occurrences follow one pattern.
- Also extended to adverbial clauses for comparison.

Main Results & Implications

- About 60% of languages match the original "if-before-then" rule strictly.
- Under the new quantitative formulation, **all languages** show a higher proportion of "if-before-then" than the reverse.
- Adverbial clauses show no universal trend – more variation across languages.
- SOV languages tend to place conditionals initially more rigidly.
- Implications:
  - The original statement is too rigid – frequency-based universals fit better.
  - Conditional clauses may reflect stronger processing or pragmatic constraints than adverbials.
  - Methodology can extend to other proposed universals and less-resourced languages.

2nd example: work of Futrell

Quantifying Word Order Freedom in Dependency Corpora

- **(futrell-etal-2015-quantifying)**
- languages with more word order freedom have more case marking

↪ hypothesis first stated by **sapir1921language**

- quantify word order freedom from an analysis of dependencies in UD
- large (?) correlation between quantitative word order freedom of subjects and objects and the presence of nominative-accusative case marking.

Large-scale evidence of dependency length minimization in 37 languages

- **(futrell15pnas)**
- show that speakers prefer word orders with short dependency lengths and that languages do not enforce word orders with long dependency lengths

↪ more efficient parsing and generation of natural language

And for NLP?

- We have models that perform well on English (e.g. `Transformer`)

↪ Will they work equally well on all languages?

↪ Can they capture the full range of linguistic features?

- For example: will a pre-trained speech model "work" on a click language when no such language is present in the training data?
- Do some linguistic features require specific models? ⇒ One model to rule them all? ⇒ Do we still need to develop new models?

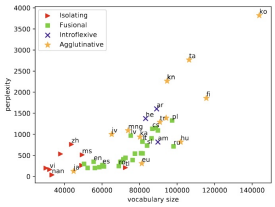
Case Study: Performance of a Language Model

- very nice work by Gerz et al. 2018: measure performance of a LM with respect to morphology complexity across a wide array of language

↪ LM ⇒ no need for annotated data

- show that morphology kind has a huge impact on performance
- ⚠⚠ this result is crap not sound ⚠⚠

↪ you can not compare perplexity across copora 🤔 this is NLP 101 🤗



- multilinguality is important from a "theoritical" point of view...
- ...but it also has "pratical" importance

The State and Fate of Linguistic Diversity and Inclusion in the NLP World, ACL 2020

(Joshi et al. 2020)

- The awakening to linguistic equity and inclusive AI
- ↪ Dominant tools risk excluding entire communities, reinforcing digital inequalities.
- Survey of languages represented in major corpora, benchmarks and ACL papers.
- Extreme concentration on a few languages: English dominates, followed by some European and Asian languages.
- Strong correlation between a country's economic/technical resources and its language's NLP presence. ⇒ Obvious, but needs precise quantification.

The/A Taxonomy of Languages (1)

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0 — The Left-Behinds	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2,191	1.2B	88.38%
1 — The Scraping-Bys	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2 — The Hopefuls	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3 — The Rising Stars	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4 — The Underdogs	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5 — The Winners	English, Spanish, German, Japanese, French	7	2.5B	0.28%

from Joshi et al. 2020

The/A Taxonomy of Languages (2)

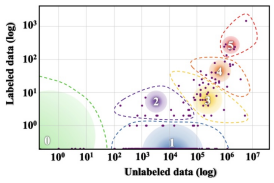


Figure 2: Language Resource Distribution: The size of the gradient circle represents the number of languages in the class. The color spectrum VIBGYOR, represents the total speaker population size from low to high. Bounding curves used to demonstrate covered points by that language class.

- The availability (or lack) of annotated and unannotated resources shapes the choice of methods.
- Three essential needs:
  1. Invest in data collection for under-represented languages.
  2. Collaborate with local language communities.
  3. Develop balanced multilingual benchmarks.

Where Are We Now?

- Increasingly, studies highlight evaluations on "linguistically diverse" datasets
- ↪ A real shift
- ↪ Not just for "ethical" reasons
- But the road ahead is still long!
- ↪ No clear definition of linguistic diversity
- ↪ Lack of benchmarks

Measuring Linguistic Diversity

- ↪ : Samardzic et al. 2024 introduces new metric to define/quantify diversity (rather than just counting the number of languages considered)
- ↪ Ploeger et al. 2024: no clear definition of "linguistic diversity"
- ↪ Estève et al. 2025: unified taxonomy of why, what on, where, and how diversity is measured in NLP

ChatGPT Performance on Major vs. Minor Languages

- Clear gap in performance between high- and low-resource languages.
- Based on Robinson et al. (2023), Lai et al. (2023), Yong et al. (2023).
- Major languages (English, Spanish, French) perform much better.
- Minor languages (African, Indigenous, under-resourced Asian) lag far behind.

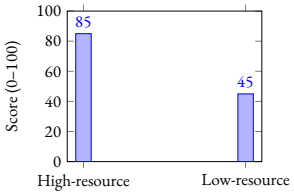


Figure: Indicative scores from recent studies.