CS11-737: Multilingual Natural Language Processing

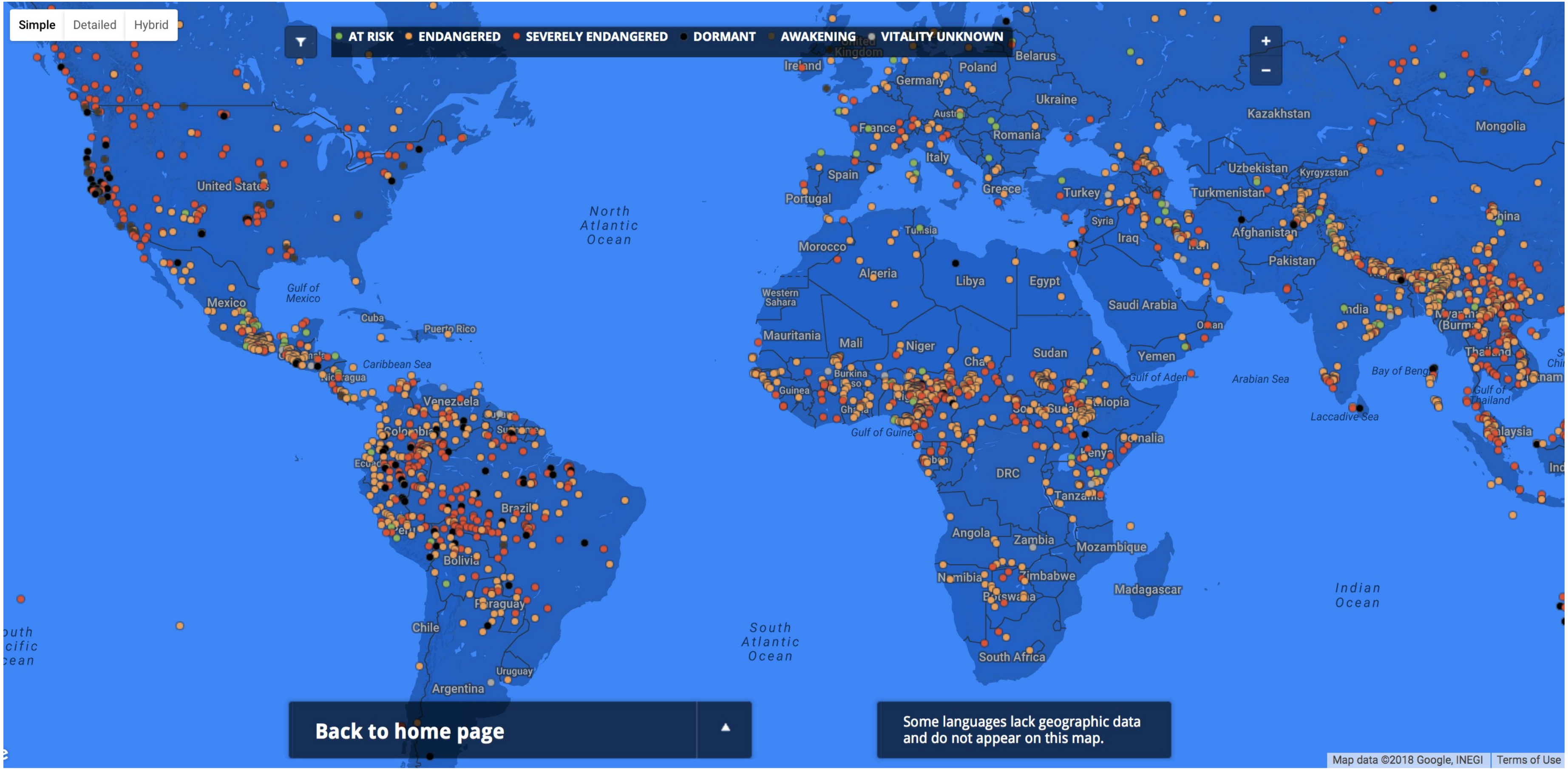
Typology: The Space of Languages

Alan W Black

Carnegie Mellon University
Language Technologies Institute

Many slides by
Yulia Tsvetkov

Linguistic diversity: ~7000 languages

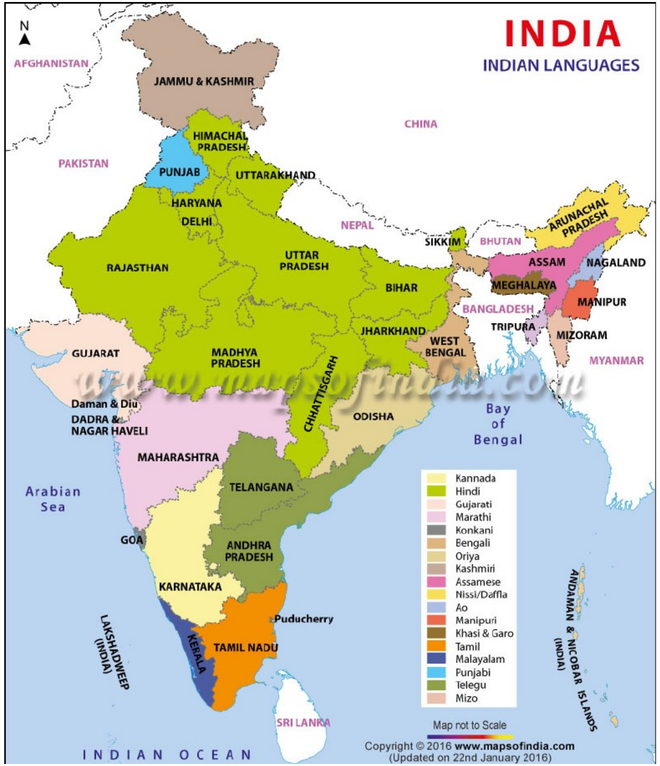


Some languages lack geographic data and do not appear on this map.

Linguistic Diversity

There are about 460 languages in India.

1.38 billion people

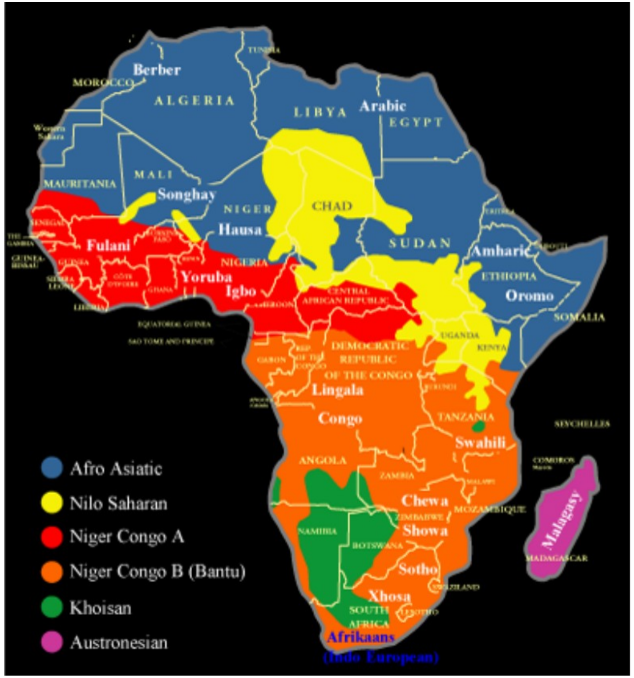


Copyright © 2016 www.mapsofindia.com
(Updated on 22nd January 2016)

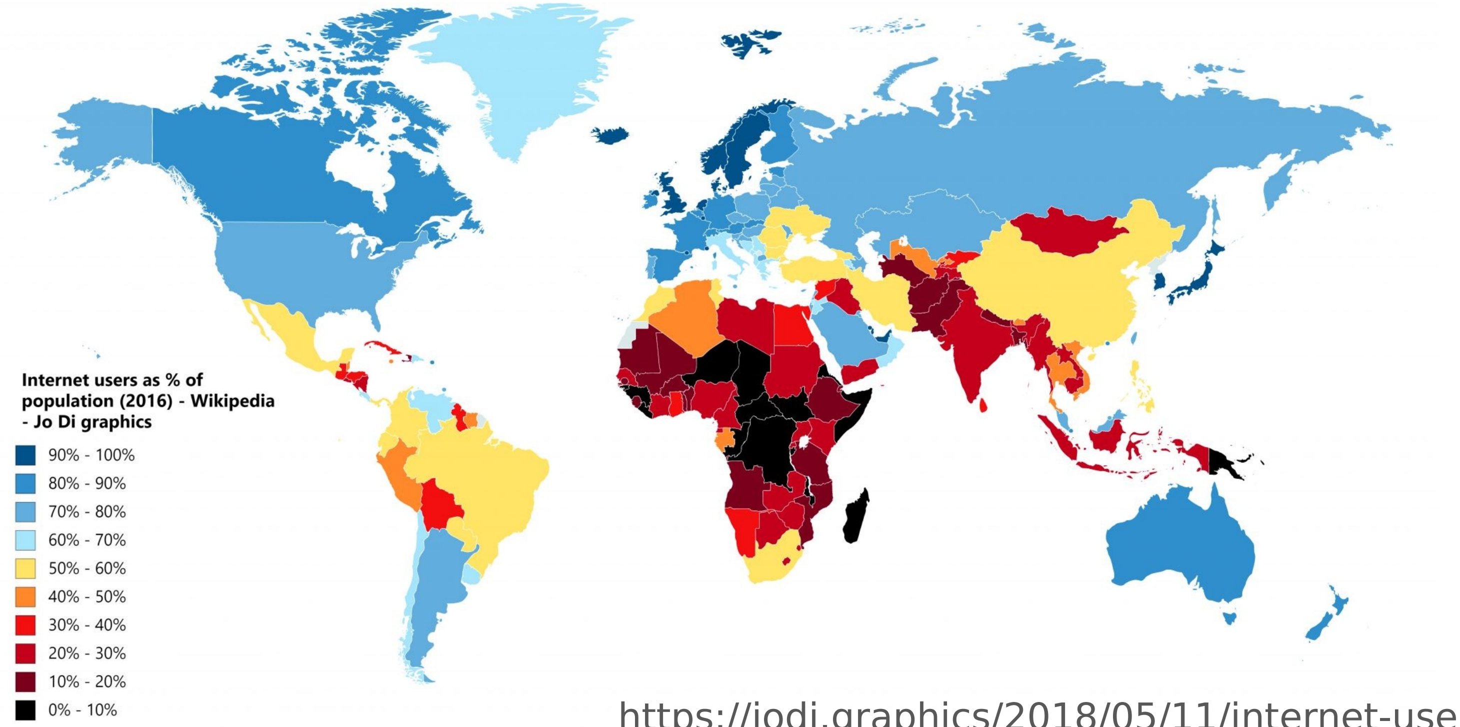
Linguistic Diversity

Africa is a continent with a very high linguistic diversity:
there are an estimated 1.5-2K African languages from 6 language families.

1.33 billion people



Low-resource/multilingual NLP



https://jodi.graphics/2018/05/11/internet-users-as-of-population/

40% of world’s population: South Asia - 1.75 billion, Africa - 1.3 billion, etc.

How to define similarity across languages?

- Word overlap and sub-word overlap
  - Russian – Русский
  - Ukraininan – Українська
  - Chinese – 中文
  - Korean – 한국어
  - Vietnamese – Tiếng Việt
  - Georgian – ქართული
  - Japanese – 日本人
  - Turkish – Türk
  - Hebrew – עִבְרִית
  - Arabic –عربى
  - Hindi – हिन्दी
  - Xhosa – isiXhosa
- Areal similarity www.glottolog.org
- Demographic similarity

Genealogical similarity

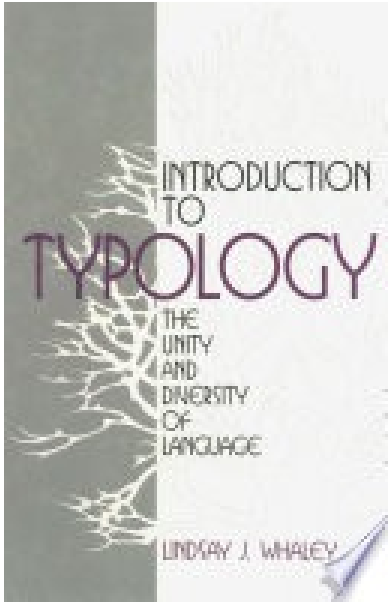
1. Niger–Congo (1,542 languages) (21.7%)
2. Austronesian (1,257 languages) (17.7%)
3. Trans–New Guinea (482 languages) (6.8%)
4. Sino-Tibetan (455 languages) (6.4%)
5. Indo-European (448 languages) (6.3%)
6. Australian [dubious] (381 languages) (5.4%)
7. Afro-Asiatic (377 languages) (5.3%)
8. Nilo-Saharan [dubious] (206 languages) (2.9%)
9. Oto-Manguean (178 languages) (2.5%)
10. Austroasiatic (167 languages) (2.3%)
11. Tai–Kadai (91 languages) (1.3%)
12. Dravidian (86 languages) (1.2%)
13. Tupian (76 languages) (1.1%)

www.ethnologue.com

Typological similarity

- Linguistic typology: classification of languages according to their functional and structural properties
    - explains common properties across languages
    - explains structural diversity across languages

“The classification of languages or components of languages based on shared formal characteristics.”

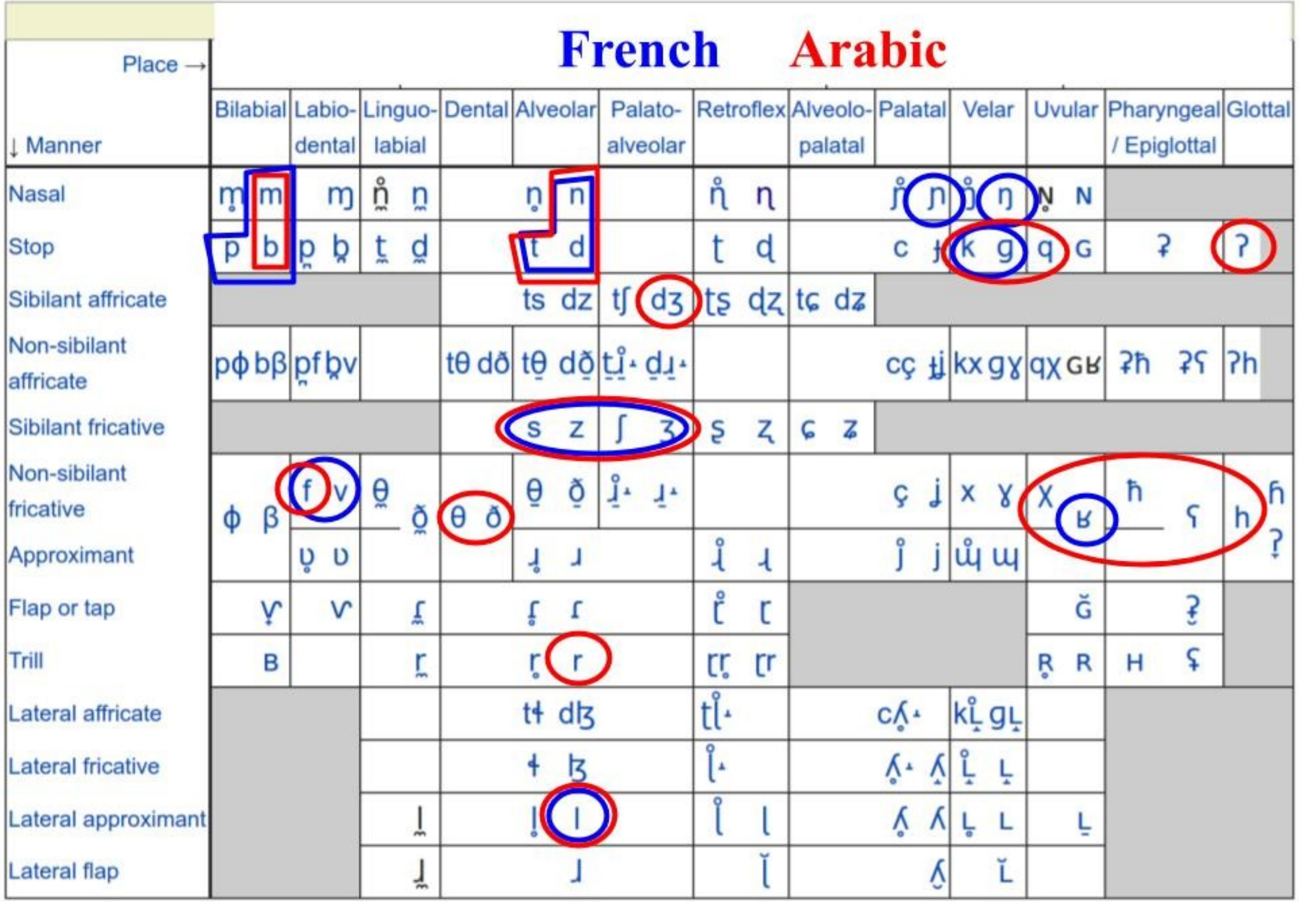


INTRODUCTION TO TYPOLOGY

THE UNITY AND DIVERSITY OF LANGUAGE

LINDSAY J. WHALEY

Linguistic typology example: phonology

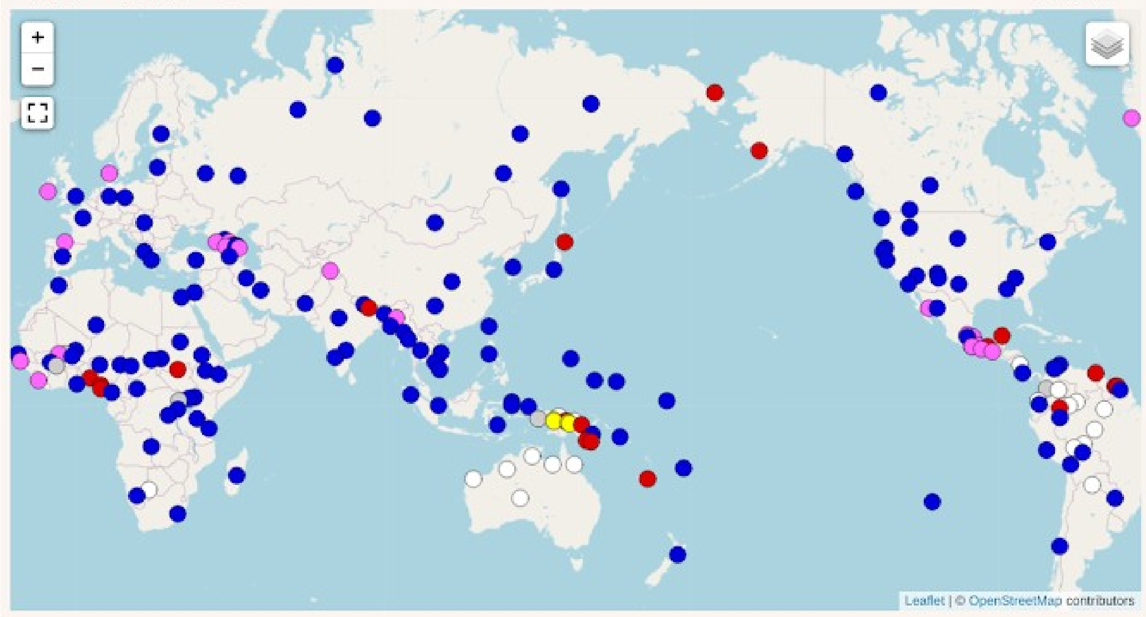


French Arabic

Place → ↓ Manner	Bilabial	Labio-dental	Linguo-labial	Dental	Alveolar	Palato-alveolar	Retroflex	Alveolo-palatal	Palatal	Velar	Uvular	Pharyngeal / Epiglottal	Glottal
Nasal	m̥ m	ɱ	n̼̊ n̼		n̥ n		ɳ̊ ɳ		ɲ̊ ɲ	ŋ̊ ŋ	ɴ̥ ɴ		
Stop	p b	p̪ b̪	t̼ d̼		t d		ʈ ɖ		c ɟ	k ɡ	q ɢ	ʡ	ʔ
Sibilant affricate					ts dz	tʃ dʒ	ʈʂ ɖʐ	tɕ dʑ					
Non-sibilant affricate	pɸ bβ	p̪f b̪v		tθ dð	tθ̠ dð̠	t̠ɹ̠̊˔ d̠ɹ̠˔			cç ɟʝ	kx ɡɣ	qχ ɢʁ	ʡʜ ʡʢ	ʔh
Sibilant fricative					s z	ʃ ʒ	ʂ ʐ	ɕ ʑ					
Non-sibilant fricative	ɸ β	f v	θ̼ ð̼	θ ð	θ̠ ð̠	ɹ̠̊˔ ɹ̠˔			ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Approximant		ʋ̥ ʋ			ɹ̥ ɹ		ɻ̊ ɻ		j̊ j	ɰ̊ ɰ			ʔ̞
Flap or tap	ⱱ̟	ⱱ	ɾ̼		ɾ̥ ɾ		ɽ̊ ɽ				ɢ̆	ʡ̆	
Trill	ʙ		r̼		r̥ r		ɽr̥ ɽr				ʀ̥ ʀ	ʜ ʢ	
Lateral affricate					tɬ dɮ		tɭ̊˔		cʎ̝̊	kʟ̝̊ ɡʟ̝			
Lateral fricative					ɬ ɮ		ɭ̊˔		ʎ̝̊ ʎ̝	ʟ̝̊ ʟ̝			
Lateral approximant			l̼		l̥ l		ɭ̊ ɭ		ʎ̥ ʎ	ʟ̥ ʟ	ʟ̠		
Lateral flap			ɺ̼		ɺ		ɭ̆		ʎ̆	ʟ̆			

Linguistic typology example: numerals

Feature 131A: Numeral Bases



Values

●	Decimal	125
●	Hybrid vigesimal-decimal	22
●	Pure vigesimal	20
●	Other base	5
●	Extended body-part system	4
○	Restricted	20

THE WORLD ATLAS
OF LANGUAGE STRUCTURES
ONLINE

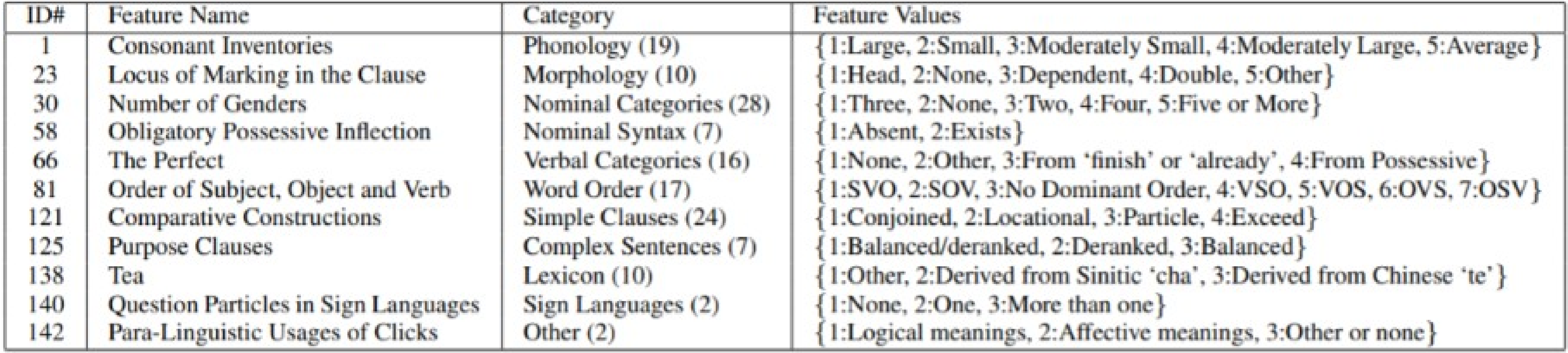
wals.info/chapter/131

WALS

THE WORLD ATLAS OF LANGUAGE STRUCTURES ONLINE

wals.info

- 2,676 languages, 192 attributes



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}

Example from Georgi, Xia and Lewis (2010)

Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013.
The World Atlas of Language Structures Online.
Leipzig: Max Planck Institute for Evolutionary Anthropology.

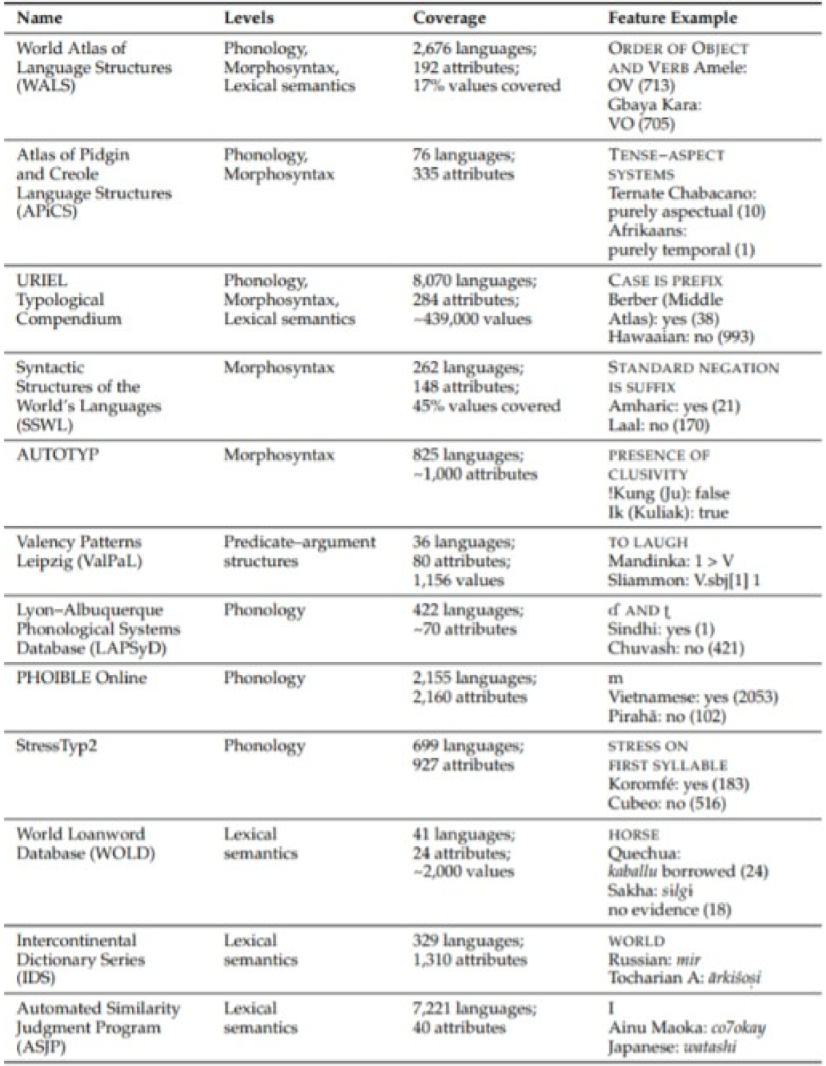
Automatic prediction of typological features

- Morphosyntactic annotation projection
  - Sentence and treebank alignments to project feature annotations from similar languages
- Unsupervised and semi-supervised feature propagation
  - Hierarchical typological clustering and majority value assignment
  - Language-family based nearest neighbor projection
  - Matrix completion
- Supervised Learning
  - Logistic regression
  - Determinant point process with neural features
- Cross-lingual distributional feature alignment

Ponti, E.M., O’horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., Shutova, E. and Korhonen, A., 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics,* 45(3), pp.559-601.

TyP-NLP Workshop at ACL 2019

Typological databases



Name	Levels	Coverage	Feature Example

World Atlas of Language Structures (WALS)	Phonology,
	Morphosyntax,
	Lexical semantics
		2,676 languages;
		192 attributes;
		17% values covered
			ORDER OF OBJECT AND VERB Amele:
			OV (713)
			Gbaya Kara:
			VO (705)

Atlas of Pidgin and Creole Language Structures (APiCS)	Phonology, Morphosyntax	76 languages; 335 attributes	TENSE–ASPECT SYSTEMS Ternate Chabacano: purely aspectual (10) Afrikaans: purely temporal (1)

URIEL Typological Compendium	Phonology,
	Morphosyntax,
	Lexical semantics
		8,070 languages;
		284 attributes;
		∼439,000 values
			CASE IS PREFIX
			Berber (Middle Atlas): yes (38)
			Hawaaian: no (993)

Syntactic Structures of the World's Languages (SSWL)	Morphosyntax	262 languages;
		148 attributes;
		45% values covered
			STANDARD NEGATION IS SUFFIX
			Amharic: yes (21)
			Laal: no (170)

AUTOTYP	Morphosyntax	825 languages; ∼1,000 attributes	PRESENCE OF CLUSIVITY !Kung (Ju): false Ik (Kuliak): true

Valency Patterns Leipzig (ValPaL)	Predicate–argument structures	36 languages;
		80 attributes;
		1,156 values
			TO LAUGH
			Mandinka: 1 > V
			Sliammon: V.sbj[1] 1

Lyon–Albuquerque Phonological Systems Database (LAPSyD)	Phonology	422 languages; ∼70 attributes	ɗ AND ţ Sindhi: yes (1) Chuvash: no (421)
PHOIBLE Online	Phonology	2,155 languages; 2,160 attributes	m Vietnamese: yes (2053) Pirahã: no (102)
StressTyp2	Phonology	699 languages; 927 attributes	STRESS ON FIRST SYLLABLE Koromfé: yes (183) Cubeo: no (516)

World Loanword Database (WOLD)	Lexical semantics	41 languages;
		24 attributes;
		∼2,000 values
			HORSE
			Quechua:
			kaballu borrowed (24)
			Sakha: *silgi*
			no evidence (18)

Intercontinental Dictionary Series (IDS)	Lexical semantics	329 languages; 1,310 attributes	WORLD Russian: *mir* Tocharian A: *ārkiśoṣi*
Automated Similarity Judgment Program (ASJP)	Lexical semantics	7,221 languages; 40 attributes	I Ainu Maoka: *co7okay* Japanese: *watashi*

Ponti, E.M., O'horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., Shutova, E. and Korhonen, A., 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics,* 45(3), pp.559-601.

URIEL

- URIEL typological compendium
  - Phonology, morphosyntax, lexical semantics
  - 8.070 languages, 284 attributes, $439,000 values
- `lang2vec` representations from URIEL
  https://pypi.org/project/lang2vec/

Littel, Patrick, David R. Mortensen, and Lori Levin. 2017. URIEL Typological database. *In* Proc. EACL

Malaviya, C., Neubig, G. and Littell, P., 2017. Learning language representations for typology prediction. *In* Proc. EMNLP

Linguistic universals

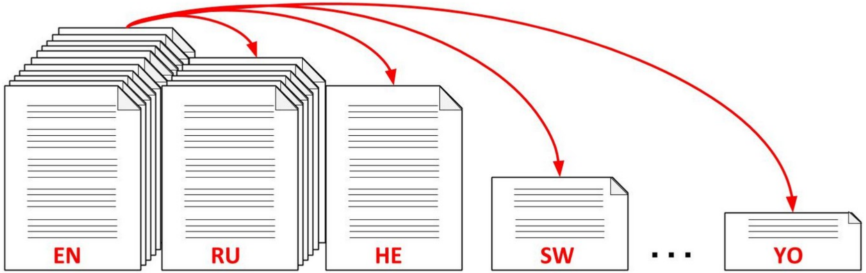
- All languages have vowels and consonants
- All (or at least nearly all) languages of the world also make a distinction between nouns and verbs

Approaches to low-resource/multilingual NLP

- Manual curation and annotation of large-scale resources for thousands of languages in infeasible or prohibitively expensive

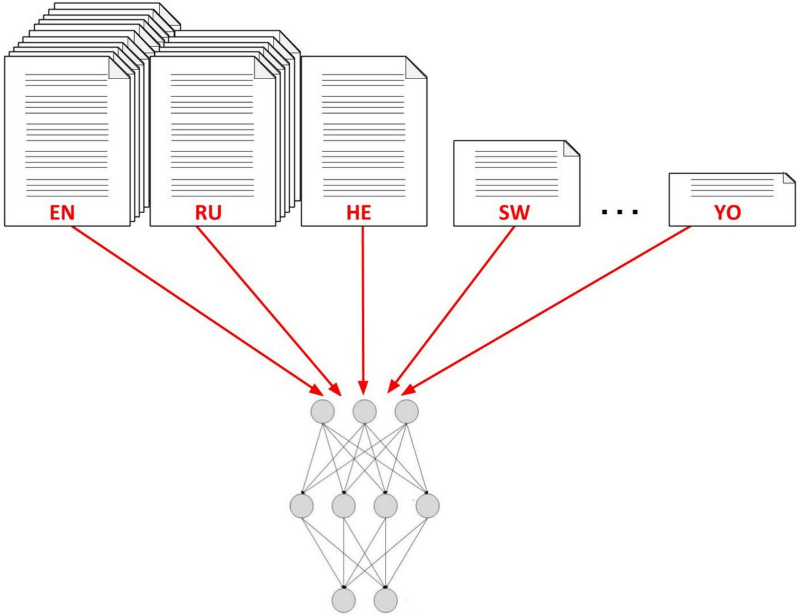
- Unsupervised learning (Snyder and Barzilay 2008; Cohen and Smith, 2009; Snyder, 2010; Vulić, De Smet, and Moens 2011; Spitkovsky et al., 2011; Goldwasser et al., 2011; Titov and Klementiev 2012; Baker et al., 2014, and many others)

Approaches to low-resource/multilingual NLP



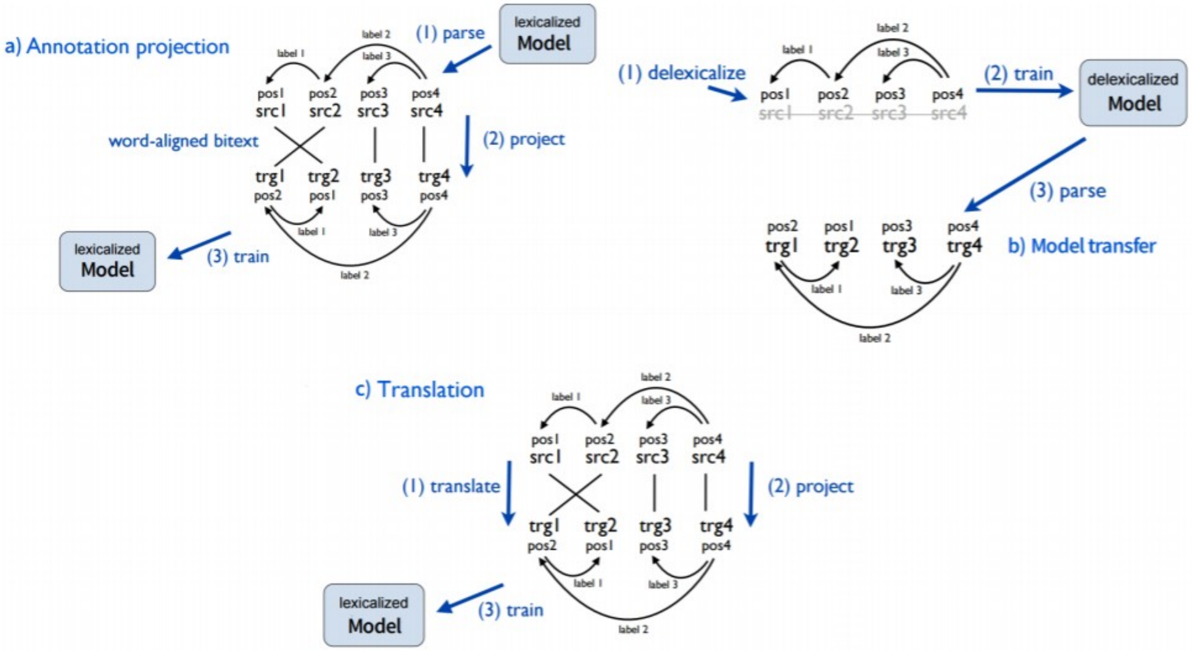
- Cross-lingual transfer learning – transfer of resources and models from resource-rich source to resource-poor target languages
  - Transfer of annotations (e.g., POS tags, syntactic or semantic features) via cross-lingual bridges (e.g., word or phrase alignments)
  - Transfer of models – train a model in a resource-rich language and adapt (e.g. fine-tune) it in a resource-poor language
- Zero-shot learning – train a model in one domain and assume it generalizes more or less out-of-the-box in a low-resource domain
- Few shot learning – train a model in one domain and use only few examples from a low-resource domain to adapt it

Approaches to low-resource/multilingual NLP



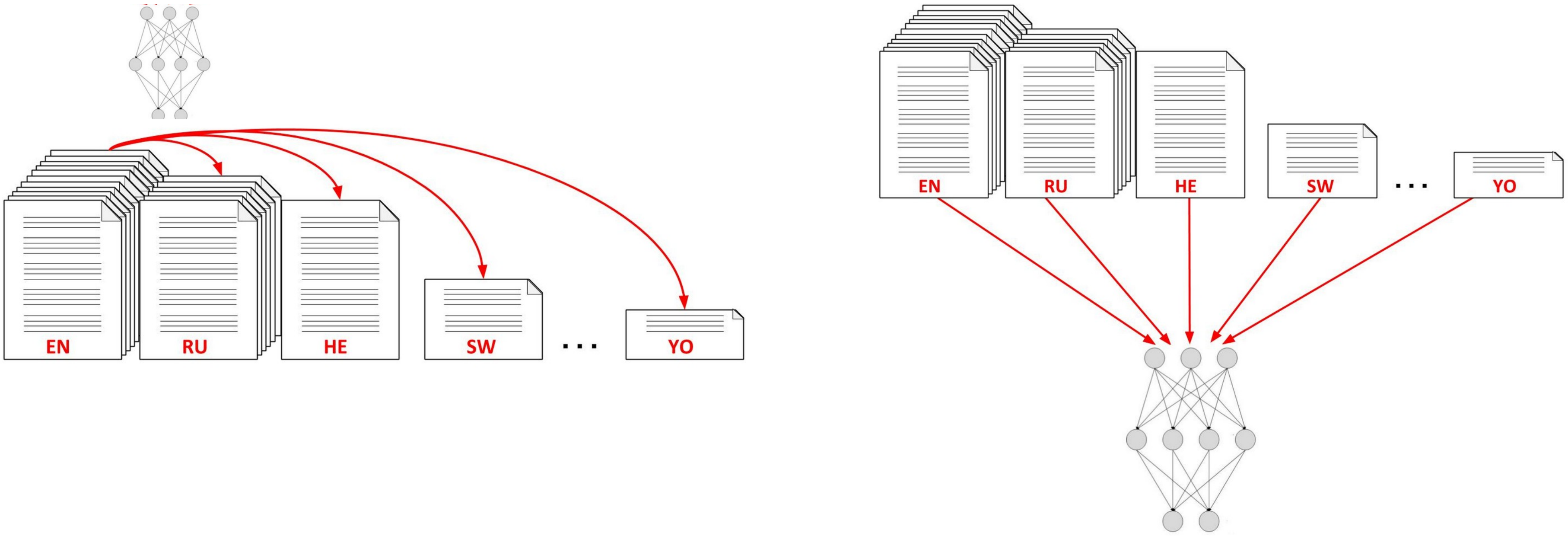
- Joint multilingual learning – train a single model on a mix of datasets in all languages, to enable data and parameter sharing where possible

Linguistic typology in NLP



Ponti, E.M., O’horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., Shutova, E. and Korhonen, A., 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics,* 45(3), pp.559-601.

Choosing transfer languages



Lin, Y.H. et al. 2019. Choosing Transfer Languages for Cross-Lingual Learning. In Proc. ACL.
https://arxiv.org/pdf/1905.12688.pdf

Open research problems

- how to extract typological features automatically from existing multilingual resources such as Universal Dependency treebank, UniMorph, Wikipedia, or Bible corpora
- how to accurately predict typological knowledge while controlling for genealogical and areal biases
- how to incorporate linguistic typology into models
- how to alleviate negative transfer and catastrophic forgetting in multilingually trained models using typological knowledge

Further readings

- Papers in tracks on morphology/phonology or multilinguality at *CL conferences
- Workshops: SIGMORPHON, SIGTYP, ComputEL, AfricaNLP, DeepLo, etc.

Class reading and discussion

- Reading
  - Ponti, E.M., O’horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., Shutova, E. and Korhonen, A., 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics,* 45(3), pp.559-601.
- Discussion question
-
- What are some unique typological features of a language that you know regarding phonology, morphology, syntax, semantics, pragmatics?