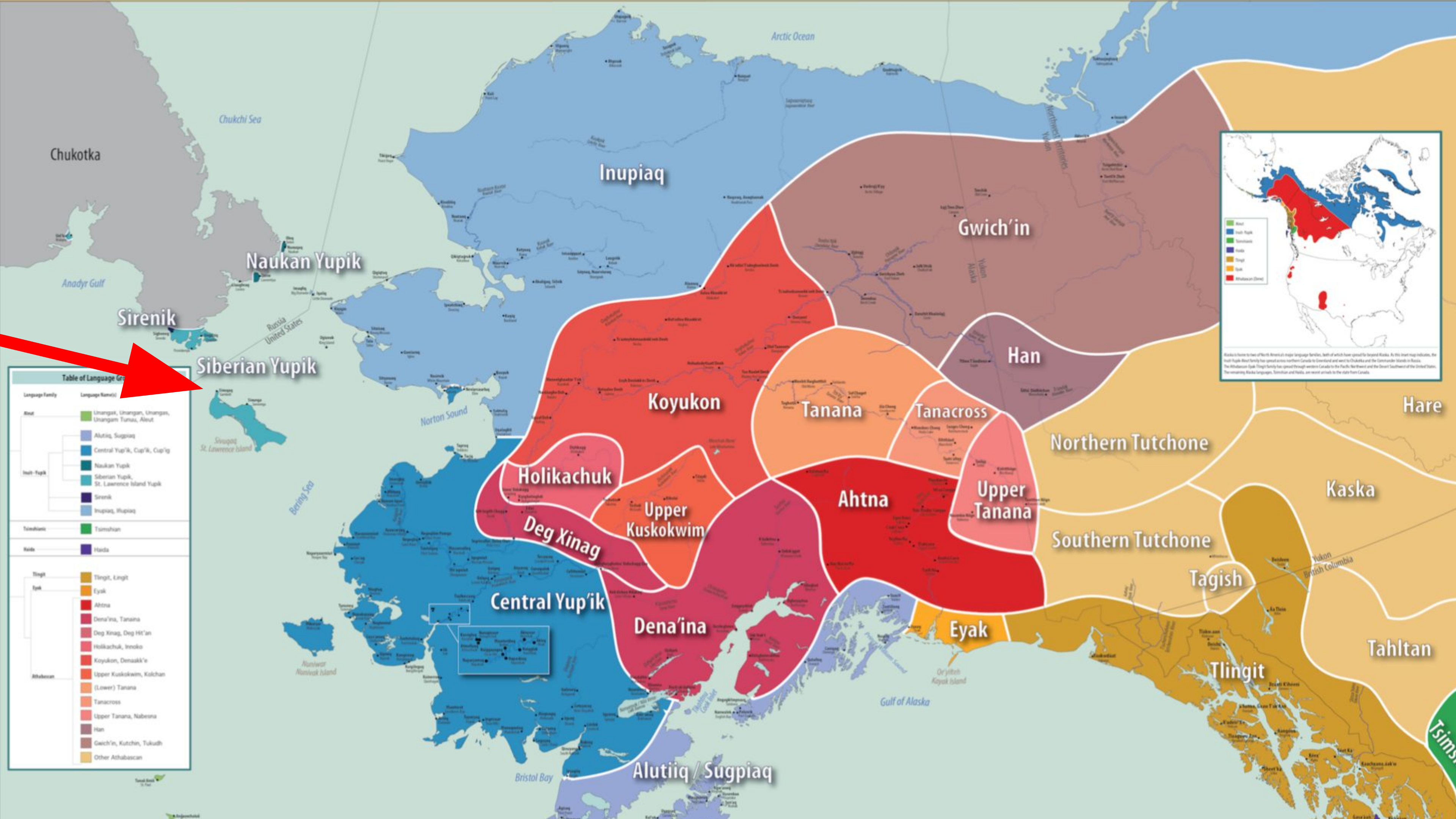
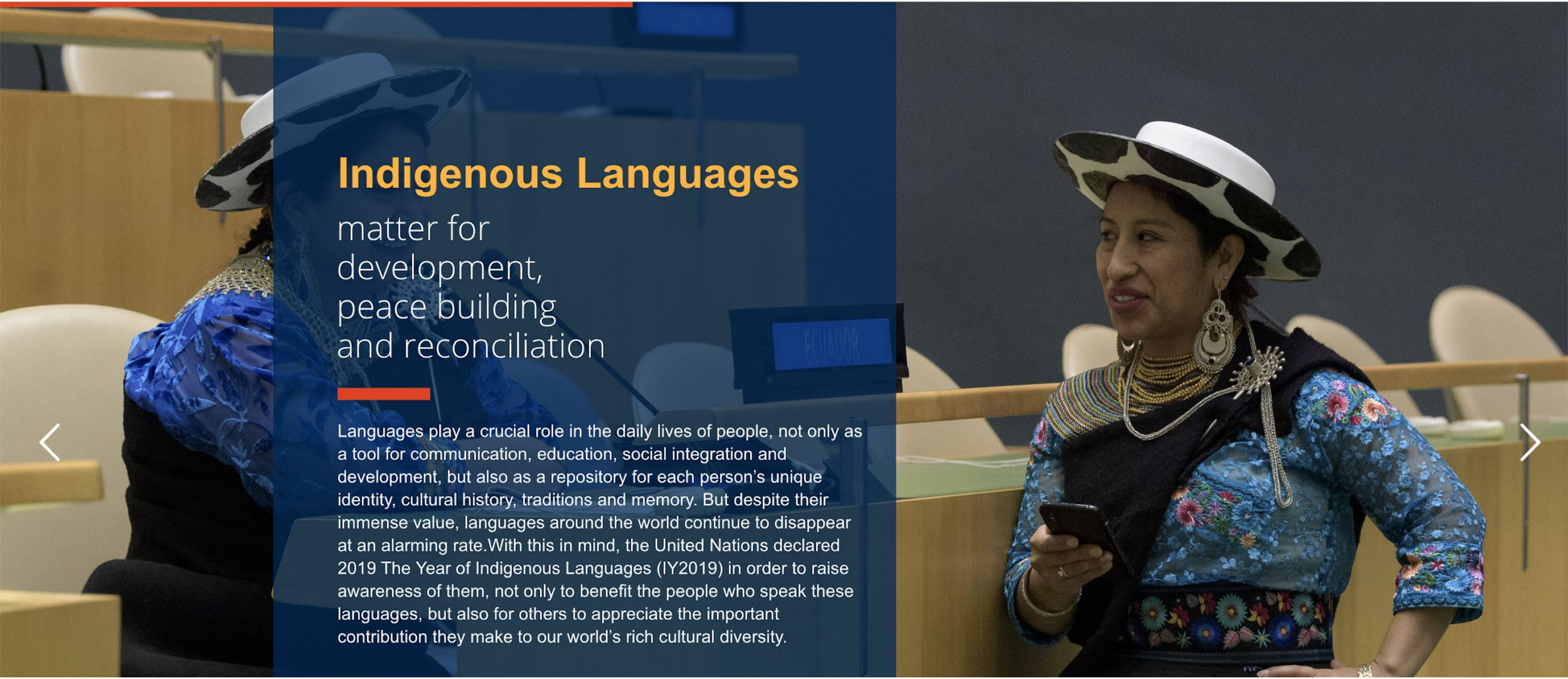
Neural Polysynthetic Language Modelling

JSALT 2019 Opening Presentation



Arctic Ocean
Chukchi Sea
Chukotka
Anadyr Gulf
Inupiaq
Naukan Yupik
Sirenik
Russia
United States
Siberian Yupik
Sivuqaq
St. Lawrence Island
Norton Sound
Bering Sea
Gwich'in
Han
Koyukon
Tanana
Tanacross
Holikachuk
Upper Kuskokwim
Deg Xinag
Ahtna
Upper Tanana
Hare
Northern Tutchone
Kaska
Southern Tutchone
Tagish
Central Yup'ik
Dena'ina
Eyak
Tahltan
Tlingit
Alutiiq / Sugpiaq
Nunivar
Nunivak Island
Bristol Bay
Gulf of Alaska
Kayak Island
Yukon
British Columbia
Table of Language Gr
Language Family
Language Name(s)
Aleut
Unangax, Unangan, Unangas, Unangam Tunuu, Aleut
Inuit-Yupik
Alutiiq, Sugpiaq
Central Yup'ik, Cup'ik, Cup'ig
Naukan Yupik
Siberian Yupik, St. Lawrence Island Yupik
Sirenik
Inupiaq, Iñupiaq
Tsimshianic
Tsimshian
Haida
Haida
Tlingit
Tlingit, Lingit
Eyak
Eyak
Athabascan
Ahtna
Dena'ina, Tanaina
Deg Xinag, Deg Hit'an
Holikachuk, Innoko
Koyukon, Denaakk'e
Upper Kuskokwim, Kolchan
(Lower) Tanana
Tanacross
Upper Tanana, Nabesna
Han
Gwich'in, Kutchin, Tukudh
Other Athabascan


Indigenous Languages
matter for development, peace building and reconciliation
Languages play a crucial role in the daily lives of people, not only as a tool for communication, education, social integration and development, but also as a repository for each person's unique identity, cultural history, traditions and memory. But despite their immense value, languages around the world continue to disappear at an alarming rate.With this in mind, the United Nations declared 2019 The Year of Indigenous Languages (IY2019) in order to raise awareness of them, not only to benefit the people who speak these languages, but also for others to appreciate the important contribution they make to our world's rich cultural diversity.
ECUADOR

2019 | INTERNATIONAL YEAR OF

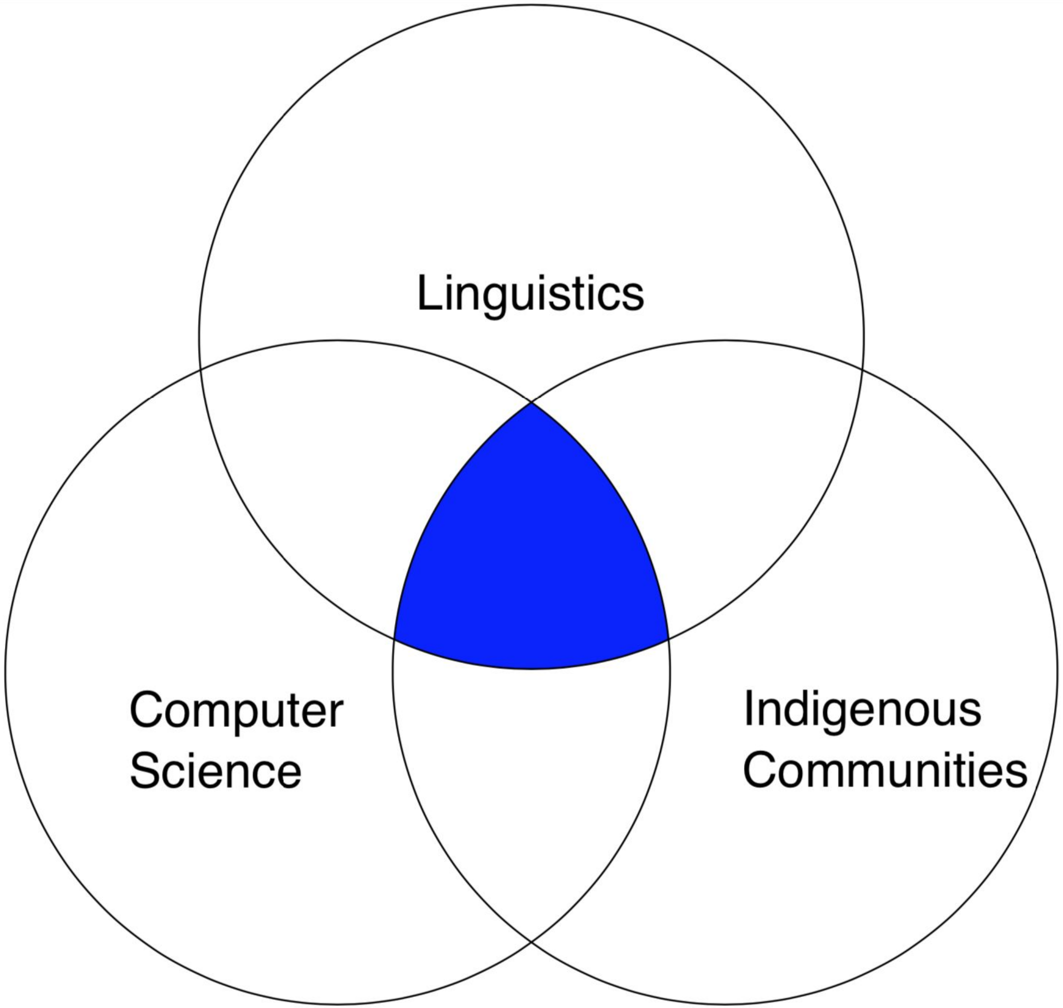
Indigenous Languages

7 thousand	370 million	90 countries	5 thousand	2680 languages
Languages spoken worldwide	Indigenous people in the world	With indigenous communities	Different indigenous cultures	In danger

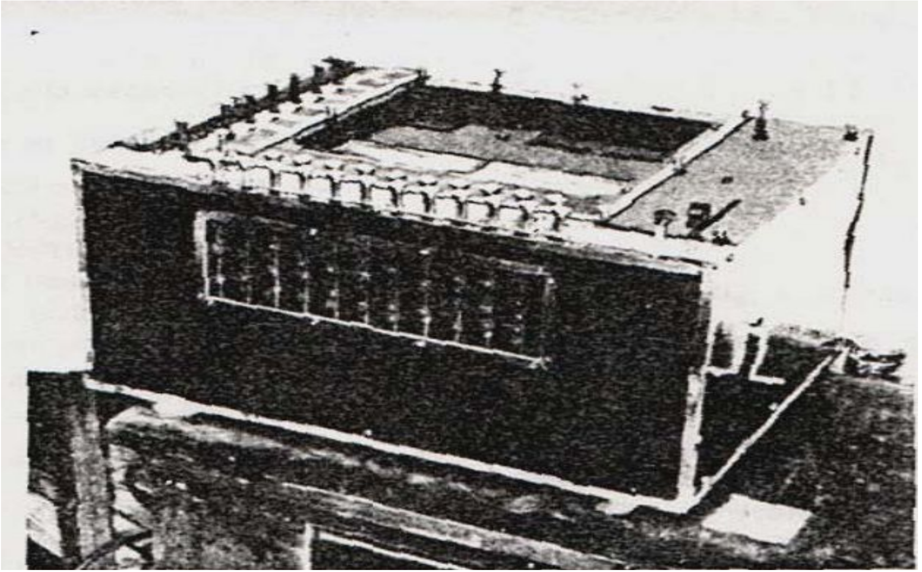
2019 | INTERNATIONAL YEAR OF

Indigenous Languages

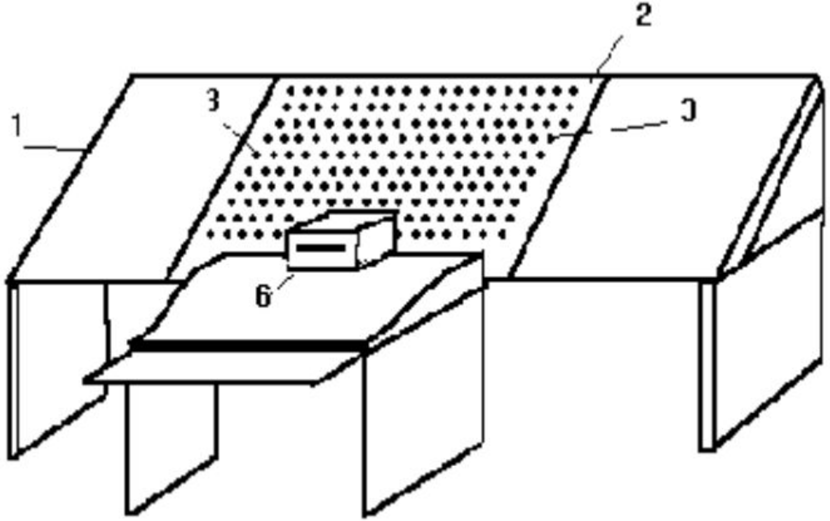
- Increasing understanding, reconciliation and international cooperation.
- Creation of favorable conditions for knowledge-sharing & dissemination of good practices with regards to indigenous languages.
- Integration of indigenous languages into standard setting.
- **Empowerment through capacity building.**
- Growth and development through elaboration of new knowledge.


Linguistics
Computer Science
Indigenous Communities

Since 1933, NLP technology has overwhelmingly focused on languages & methodologies in which the word is the primary meaning-bearing unit

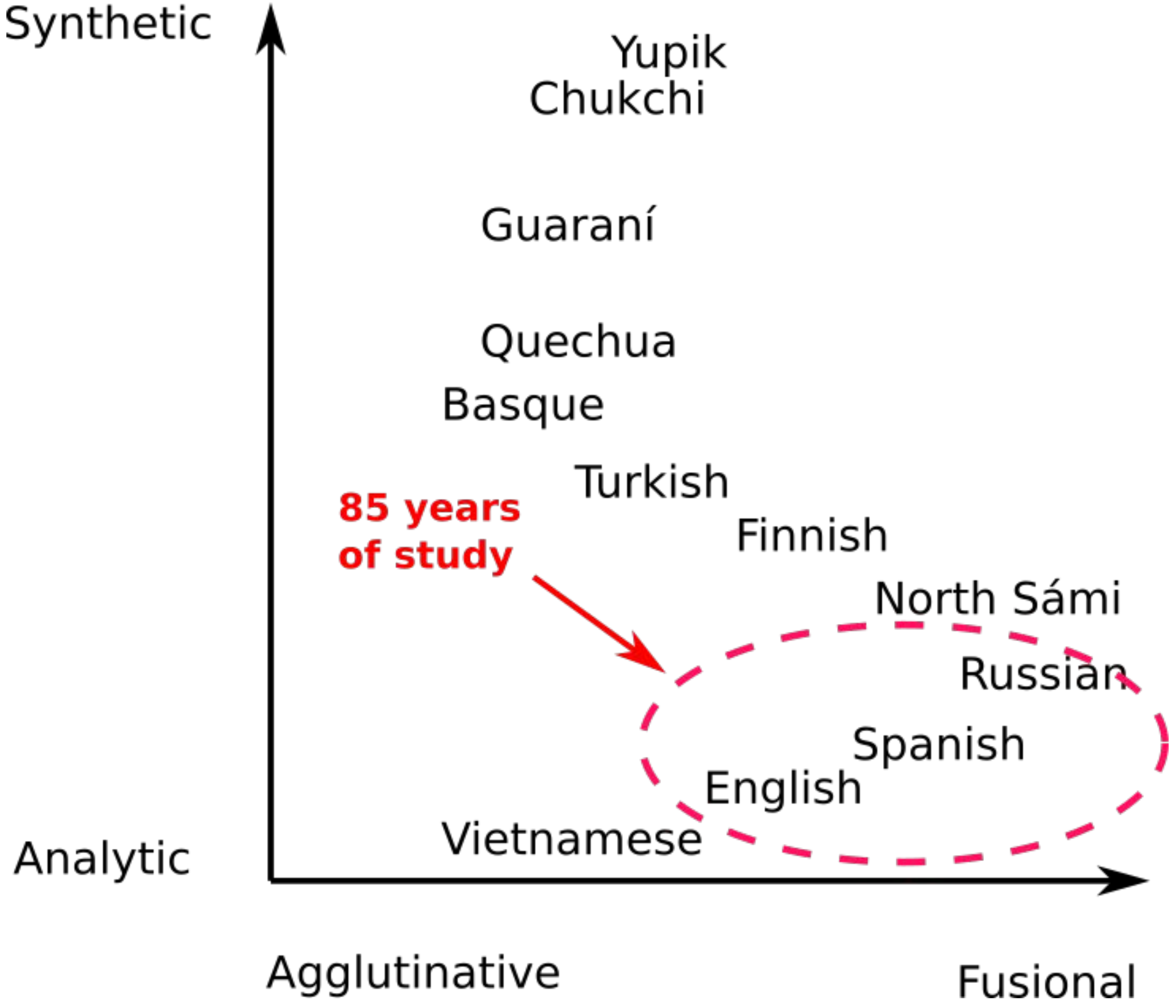


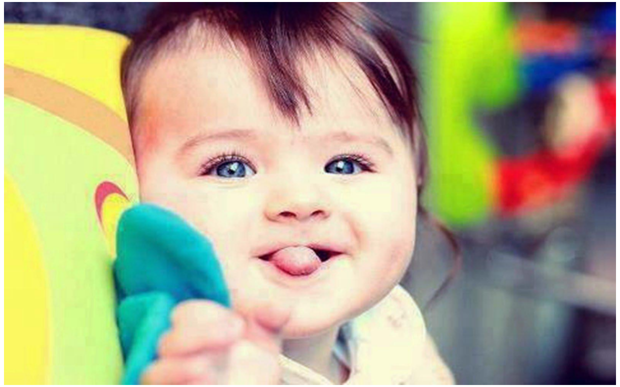
Arstrouni (1933, Paris)

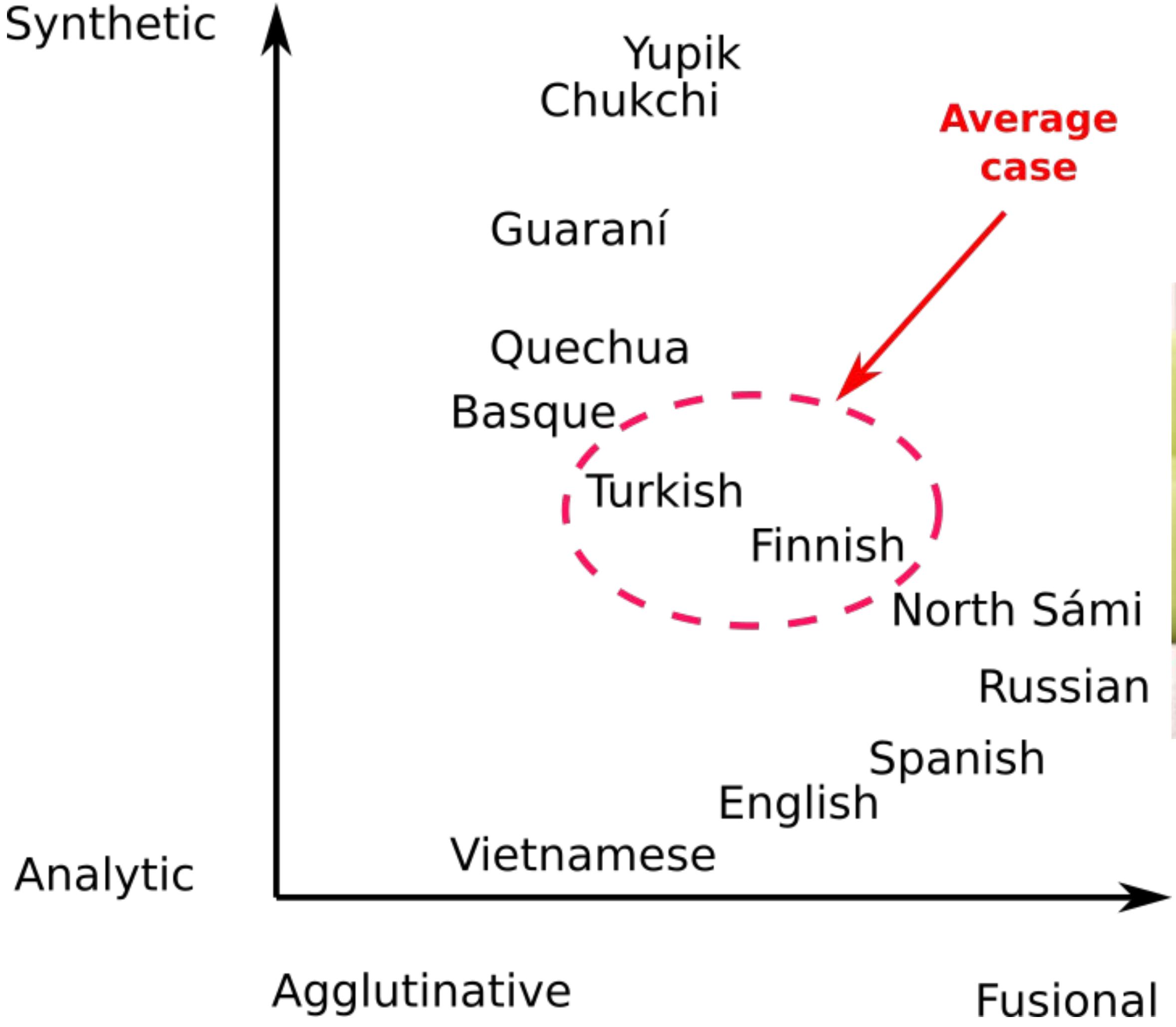


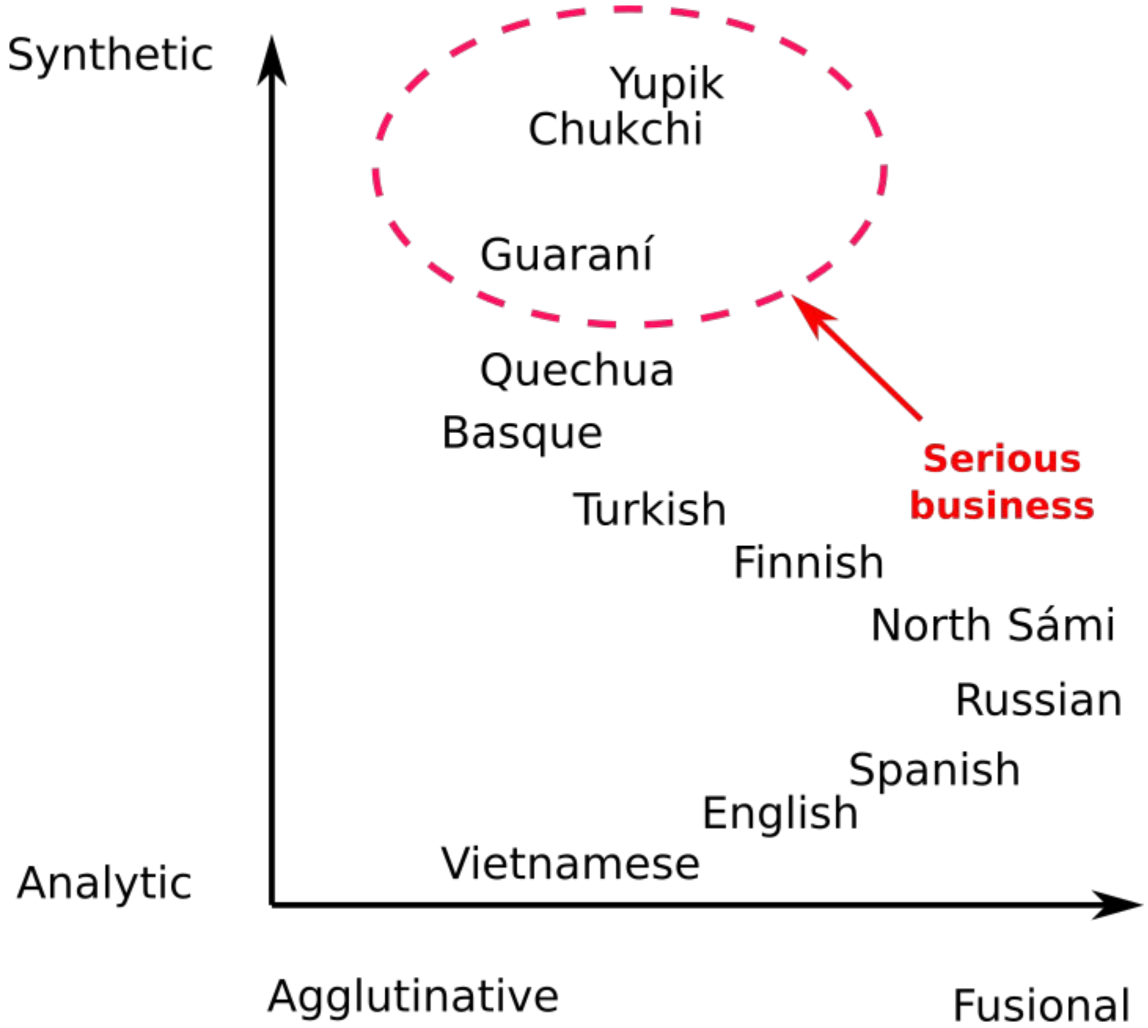
Trojanskij (1933, St. Petersburg)

For ***most*** human languages, this assumption is **fundamentally broken**


Synthetic
Yupik
Chukchi
Guaraní
Quechua
Basque
Turkish
85 years
of study
Finnish
North Sámi
Russian
Spanish
English
Vietnamese
Analytic
Agglutinative
Fusional




Synthetic
Yupik
Chukchi
Average case
Guaraní
Quechua
Basque
Turkish
Finnish
North Sámi
Russian
Spanish
English
Vietnamese
Analytic
Agglutinative
Fusional


Synthetic
Yupik
Chukchi
Guaraní
Quechua
Basque
Turkish
Finnish
North Sámi
Russian
Spanish
English
Vietnamese
Analytic
Agglutinative
Fusional
Serious business

Analytic Languages:

Little inflectional morphology

- **English** inflection
  - dog, dogs
  - walk, walks, walked, walking
- **English** derivation
  - disestablishmentarianism

Isolating Languages:

Little morphology other than compounding

- Chinese inflection
  - few affixes (prefixes and suffixes):
们:	我们,	你们,	他们,	同志们
mén:	*wǒmén,*	*nǐmén,*	*tāmén,*	*tóngzhìmén*
plural:	we,	you (pl.),	they,	comrades, LGBT people
    - “suffixes” that mark aspect: 着 *-zhě* ‘continuous aspect’
- Chinese derivation
  - 艺术**家** *yìshù**jiā*** ‘art**ist**’
- Chinese is a champion in the realm of compounding—up to 80% of Chinese words are actually compounds.

毒	+	贩	→	毒贩
dú		*fàn*		*dúfàn*
‘poison, drug’		‘vendor’		‘drug trafficker’

Synthetic Languages:

Lots of morphology relative to analytic languages

- **Czech** inflection
  - fusional suffixes:

Class	Singular	Plural
1. Nominative	město	města
2. Genitive	města	měst
3. Dative	městu	městům
4. Accusative	město	města
5. Vocative	město	města
6. Locative	městě	městech
7. Instrumental	městem	městy

Agglutinative Languages:

Verbs in **Swahili** have an average of 4-5 morphemes, http://wals.info/valuesets/22A-swa

- Words are written without hyphens or spaces between morphemes.
- **Orange** prefixes mark noun class (like gender, except **Swahili** has nine instead of two or three).
  - Verbs agree with nouns in noun class.
  - Adjectives also agree with nouns.
  - Very helpful in parsing.
- **Black** prefixes indicate tense.

Swahili	English
m-*tu* ***a***-***li***-*lal-a*	'The person slept'
m-*tu* ***a***-***ta***-*lal-a*	'The person will sleep'
wa-*tu* ***wa***-***li***-*lal-a*	'The people slept'
wa-*tu* ***wa***-***ta***-*lal-a*	'The people will sleep'

Agglutinative Languages:

Turkish

But most words have around three morphemes

uygarlaştıramadıklarımızdanmışsınızcasına

"(behaving) as if you are among those whom we were not able to civilize"

uygar	"civilized"
+laş	"become"
+tır	"cause to"
+ama	"not able"
+dık	past participle
+lar	plural
+ımız	first person plural possessive ("our")
+dan	ablative case ("from/among")
+mış	past
+sınız	second person plural ("y'all")
+casına	finite verb $\longrightarrow$ adverb ("as if")

Derivational and inflectional morphology.

From the Jurafsky and Martin textbook.

Polysynthetic Languages

- All definitions are problematic
  - “A word means a whole sentence”
- There is no one feature (e.g., noun incorporation) that is in every polysynthetic language
- The definition isn’t important
- What is important
  - There is a much larger number of distinct possible words
    - Beyond Turkish and Finnish
  - All languages have some of the properties of polysynthesis
    - Compounding, mild recursion in morphology (*oper*___ation___*aliz*___ation___), causative, applicative, passive, negative

Inuit-Yupik Language Family

Yupik
- St. Lawrence Island
- Central Alaskan

Inuit
- Inupiaq
- Innuinaqtun
- Inuktitut
- Greenlandic (Kalaallisut)

Inuktitut Morphology

And even though it's not snowing a great deal, I'm not going out.

Qanniqlaunngikkaluaqtuqlu aninngittunga.

qanniq-+-lak + uq-+-nngit-+-galuaq-+-tuq + lu

to snow a little frequently not although 3rd pers. sg. and

ani-+-nngit-+-junga

to go out not 1st pers. sg.

http://en.wikipedia.org/wiki/Inuit_grammar + Uqailaut morph analyzer

Inuktitut Morphology

qanniq +	lak+uq+nngit+galuaq +	tuq	+ lu
ROOT	LEXICAL AFFIXES*	GRAMMATICAL ENDING	(CLITIC)
Verbal or Nominal	Zero or many (*) Derivational – e.g. change v to n, n to v Semantic – adverbial, negative, tense Light verbs Certain adjectives incorporated into nouns	Subject/object of verbs (fused with mood/mode) Case Possessors of nouns (fused with case)	small set, zero or 1

Inuktitut Morphology

Some examples of flipping back and forth between noun and verb

umiaq + juaq + liuq + vik + mi

boat (n) + big (adj incorp) + make (n-v, light verb) + place-where (v-n, derivational) + locative

"in the shipyard"

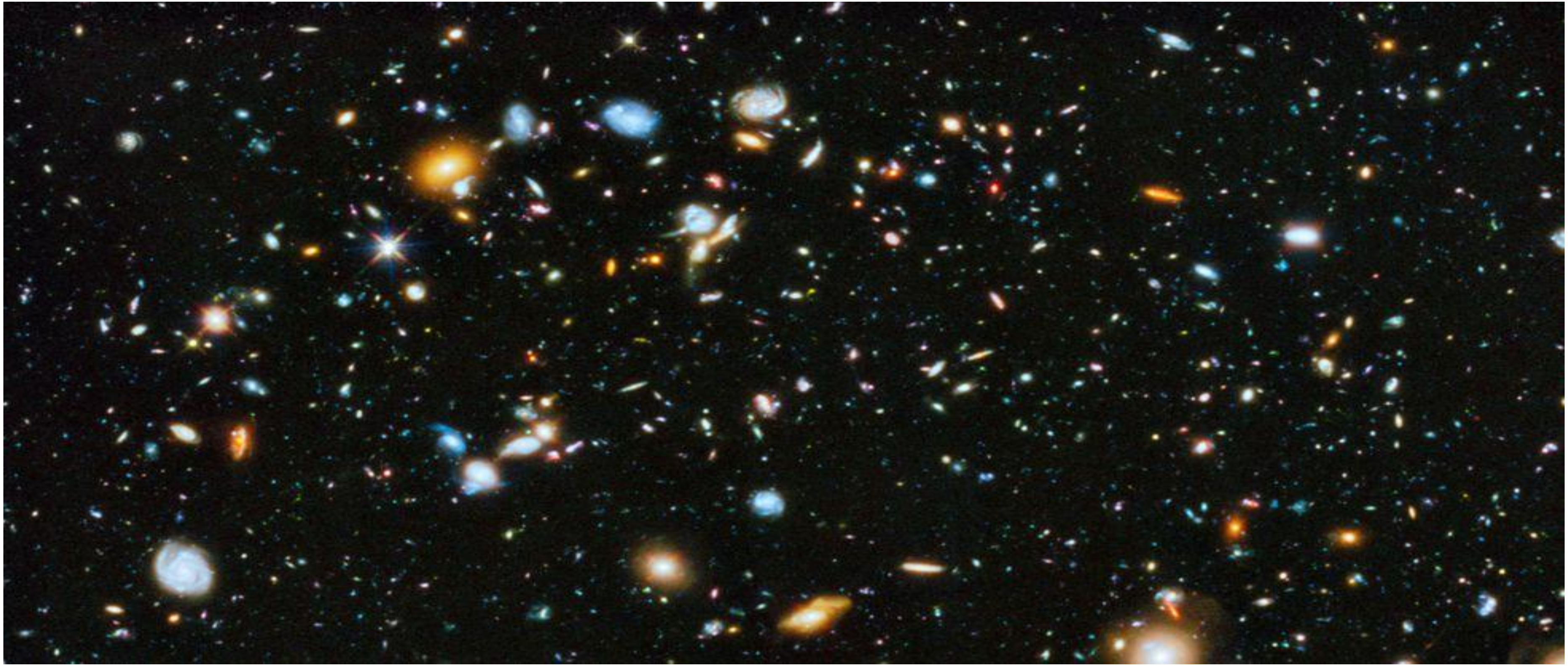
ilinniaq+ vik + siuq + junga

learn (v) + place-where (v-n, derivational) + look for (n-v, light verb) + 1-s

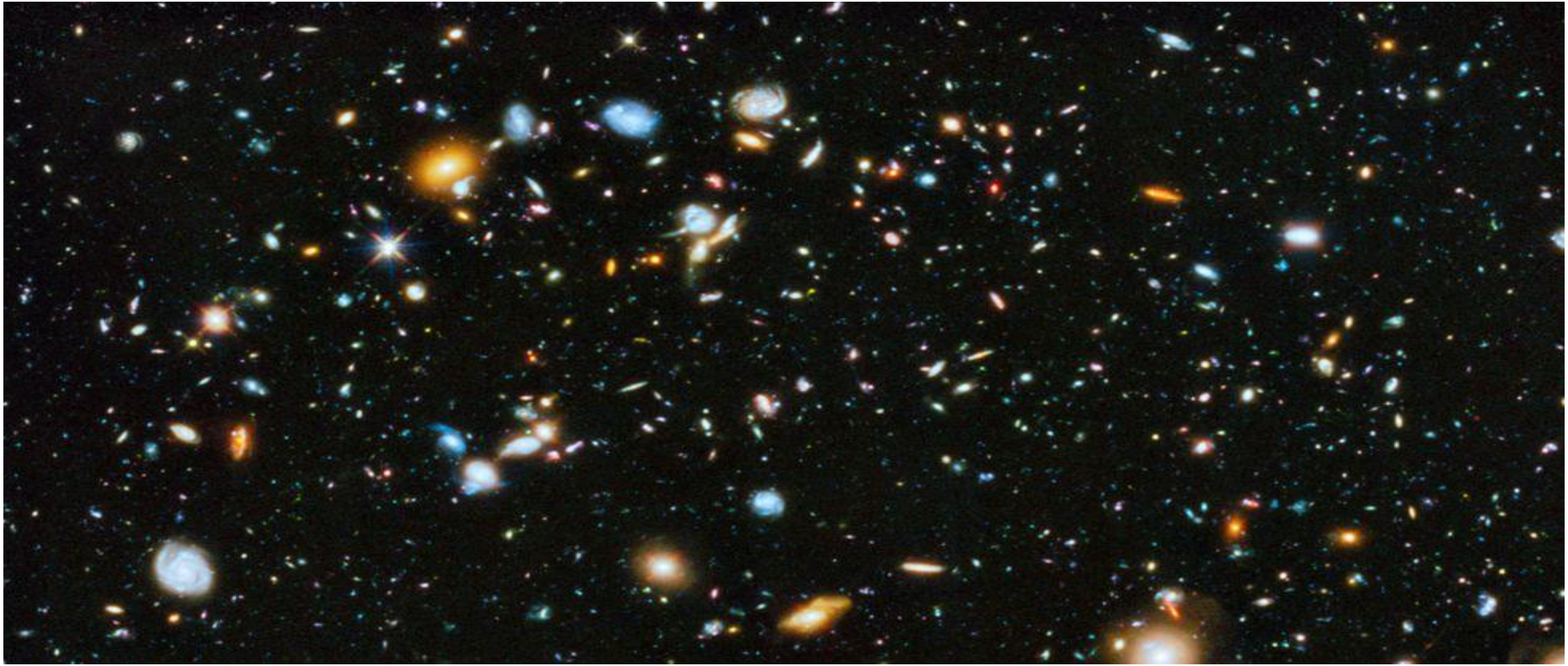
"I'm looking for a school"

http://www.inuktitutcomputing.ca/Technocrats/ILFT.php#morphology

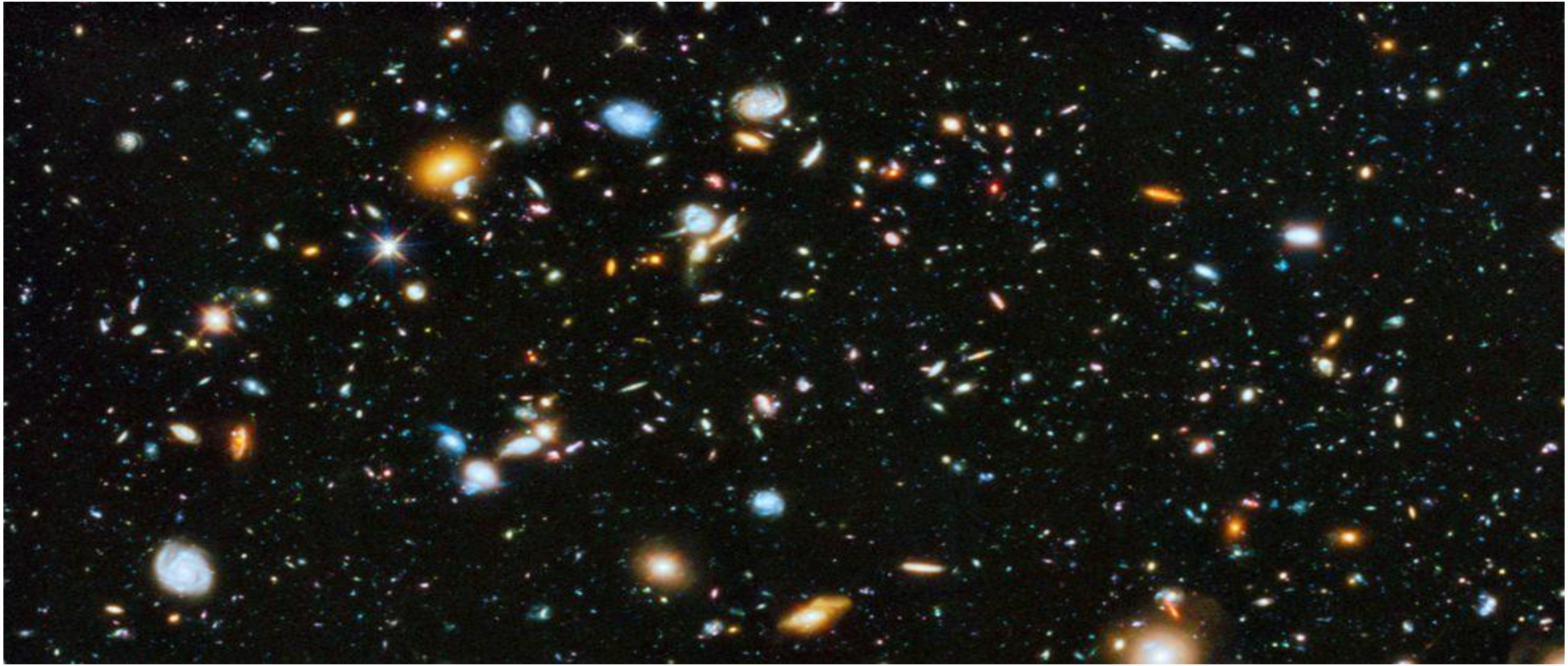
There are 1.2×10^{23} stars in the observable universe



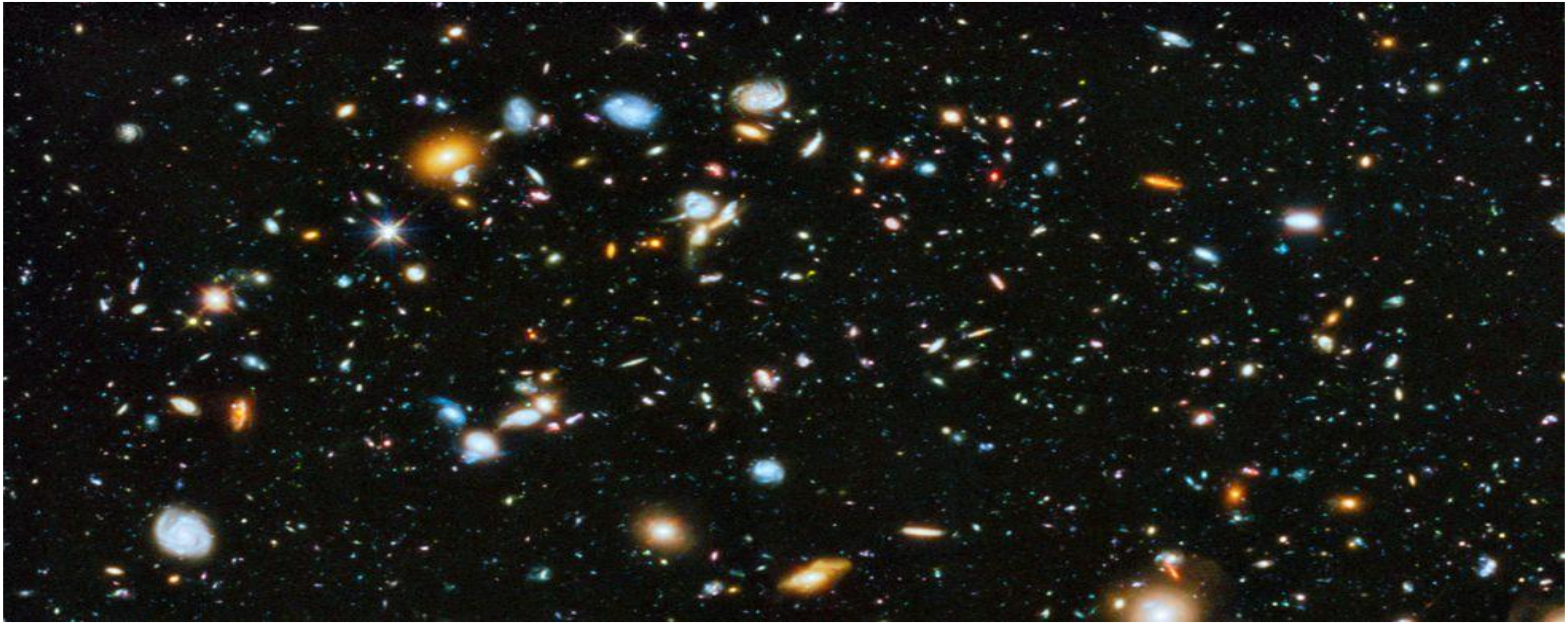
There are 1.2×10^{23} possible Yupik word forms



Big data is NOT the solution



We need a language model that sees the stars, not just the galaxies

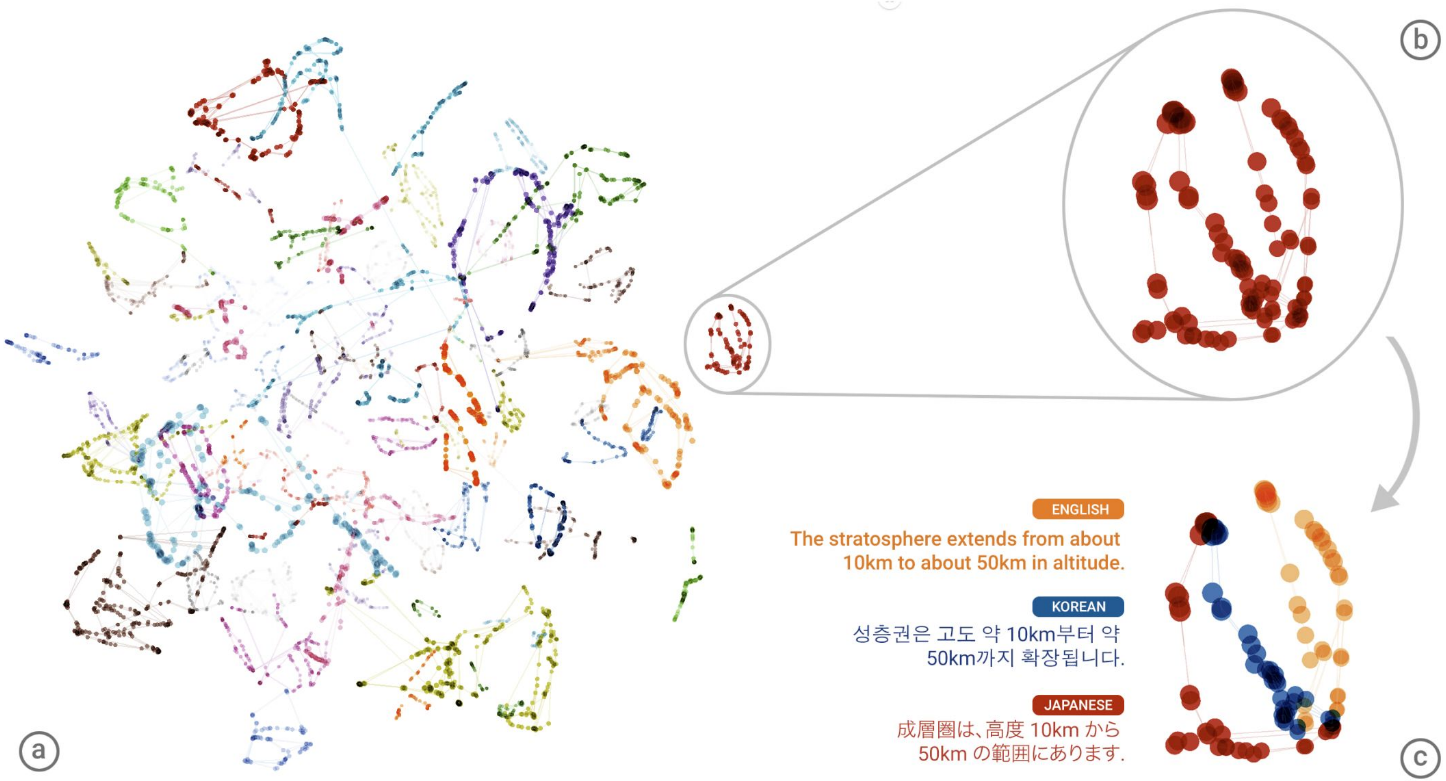


Create a language model that handles the hardest case

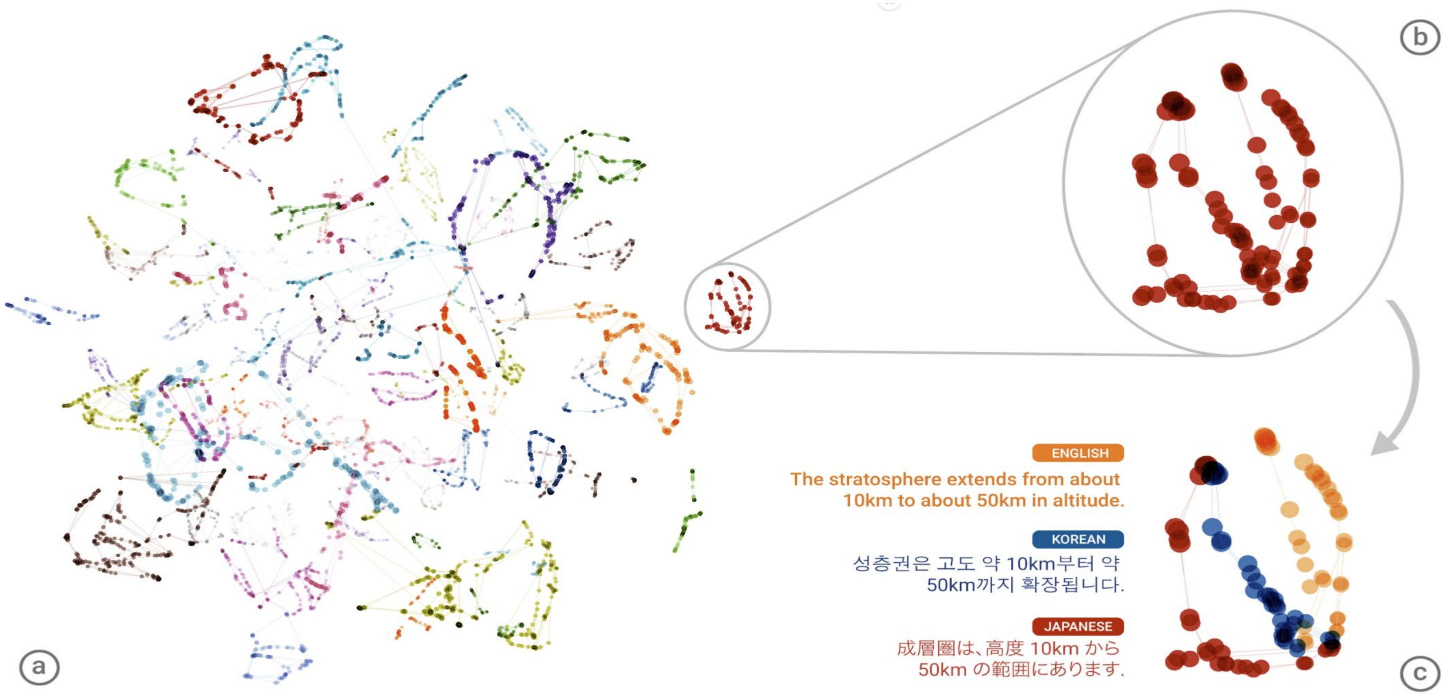
… where every other token is OOV ….

… and you'll learn lessons that will apply to the >6700 human languages that are not like English

Research questions

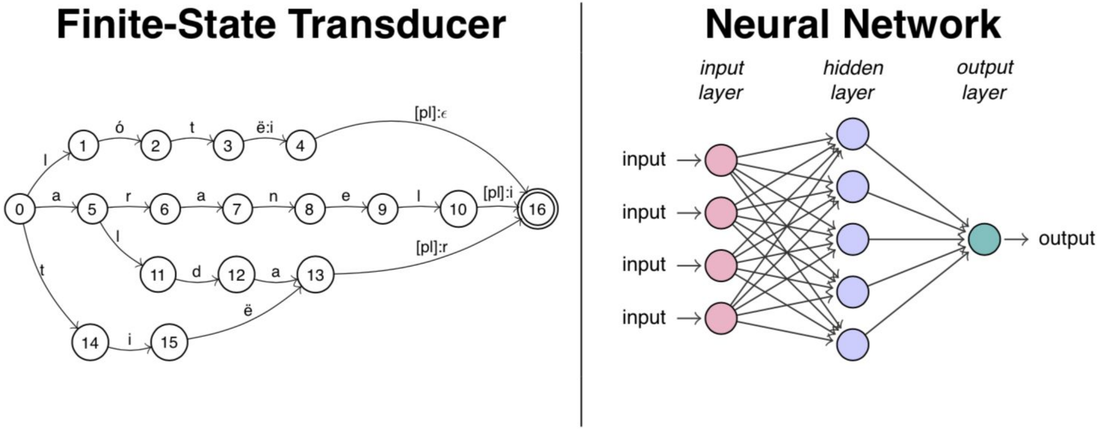


1) What embedding representations are most appropriate for languages with a high ratio of morphemes to words?



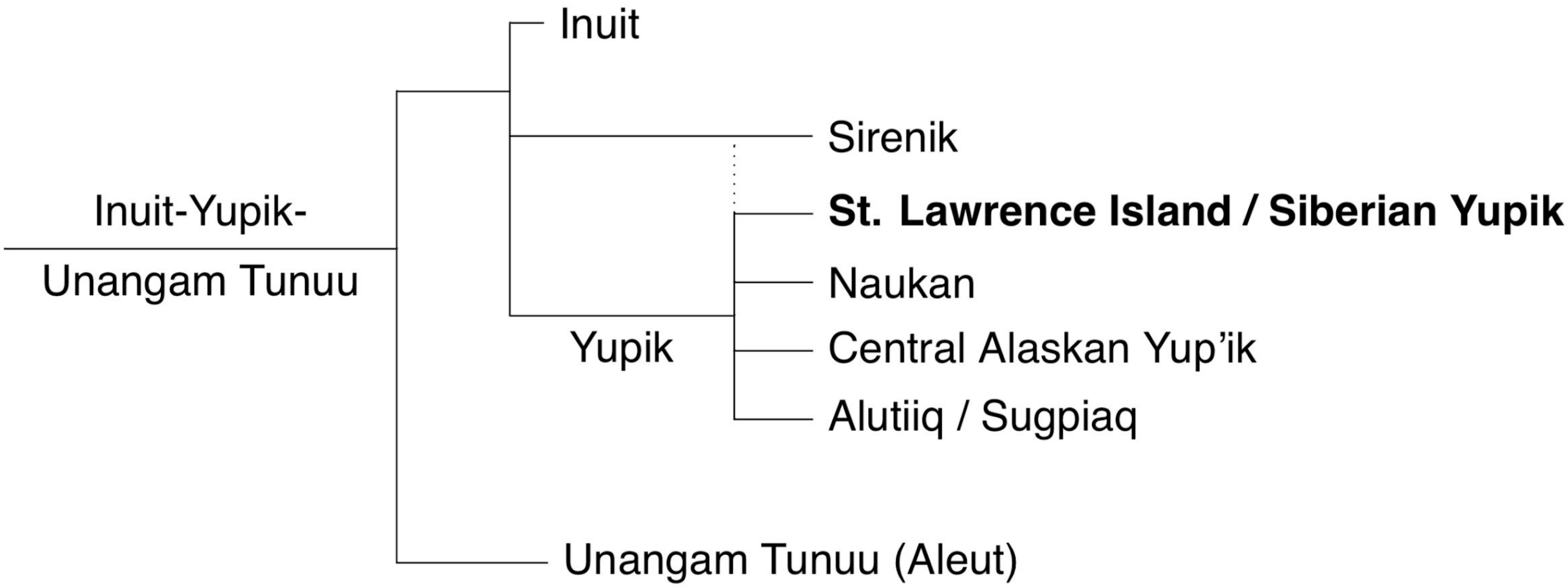
2) Can FST-based morphological analyzers effectively be used to bootstrap ***more*** effective neural LMs?

- Morphological analyzers may be implemented as a



- Neural systems require LOTS of data
  - But Yupik is a low-resource language
  - Very few surface form-lexical form pairs available

3) How can common neural models be developed for closely related languages in order to maximize the utility of sparse digitized resources?



Overview

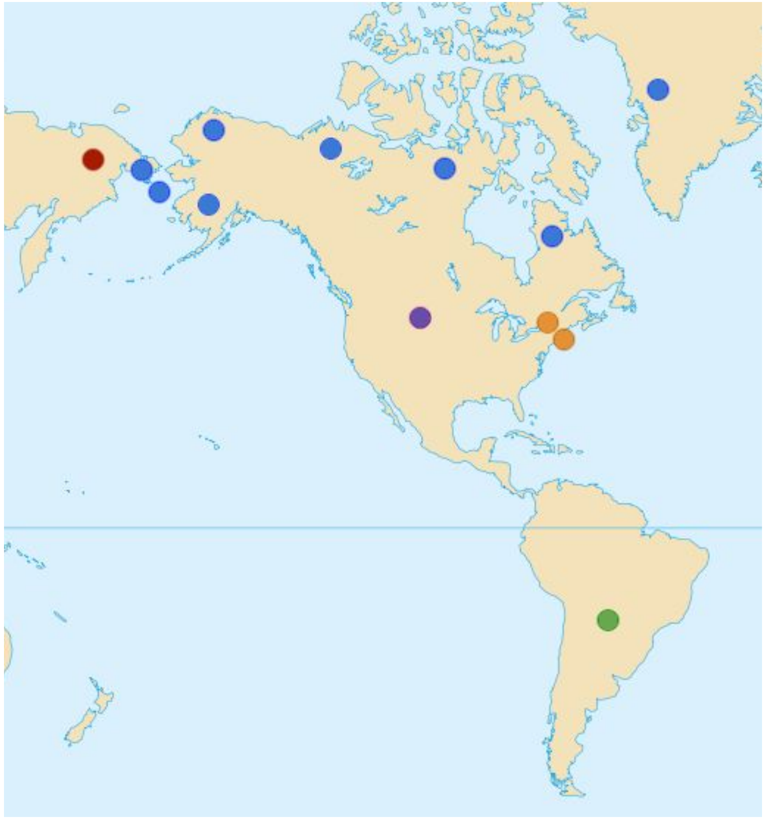
- Language selection and data collection

- Baseline systems
  - Language modelling
  - Machine translation
  - ASR

- Downstream applications
  - Spell-checkers & on-device morph-aware dictionaries
  - On-device predictive text
  - Audio transcription / search of audio archives
  - Interactive machine translation of monolingual polysynthetic texts

Language selection

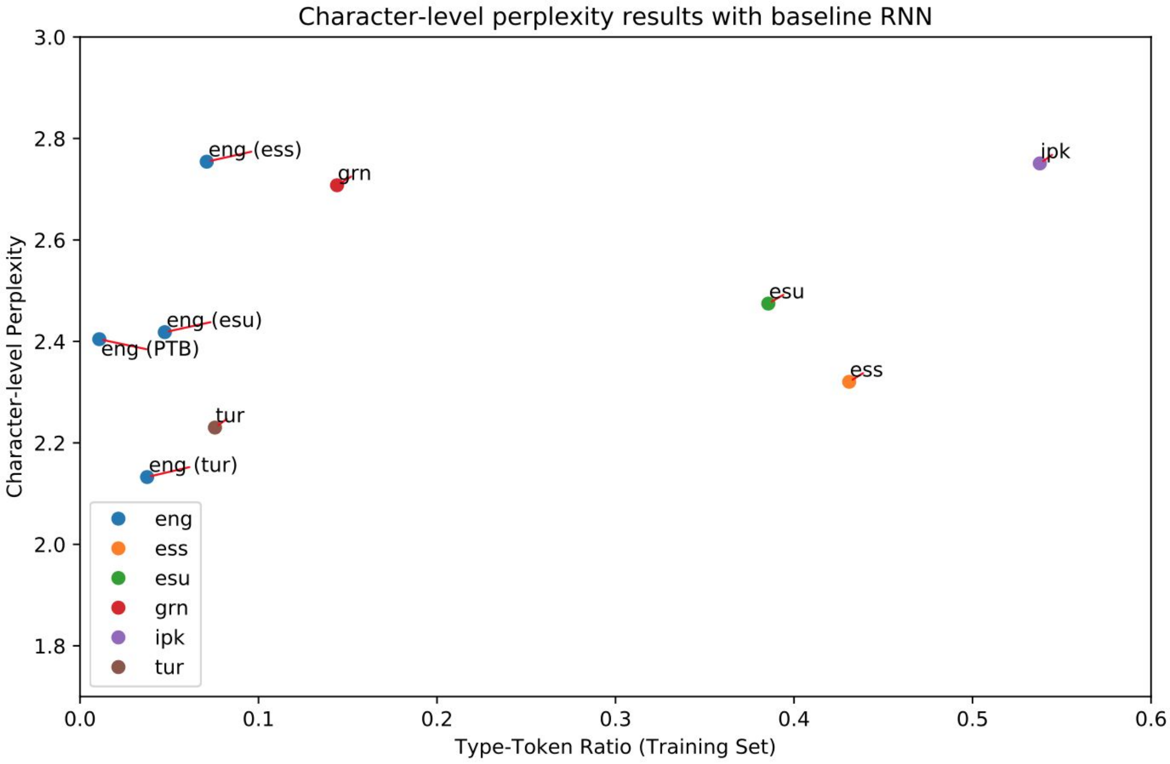
Saint Lawrence Island Yupik
Alaskan Yup’ik
North Slope Iñupiaq
Inuktitut
Kalaallisut
Chukchi
Guaraní
Seneca
Mohawk
Crow



Data collection

Language	Code	Monolingual	Parallel	FST	Audio
St. Lawrence Island Yupik	ess	24,456	8,002	✔	33h*
Alaskan Yup'ik	esu	—	45,254	—	—
North Slope Iñupiaq	esi	4,070	—	✔	—
Inuktitut	iku	1,552	1,300,159	(✔)	—
Kalaallisut	kal	5,949	—	✔	—
Chukchi	ckt	1,015	—	✔	1h30*
Guaraní	grn	—	30,078	✔	—
Seneca	see	—	—	—	12h
Mohawk	moh	—	—	—	—
Crow	cro	—	—	—	7h

Baseline system: RNN LM



Baseline system: MT

English → Inuktitut

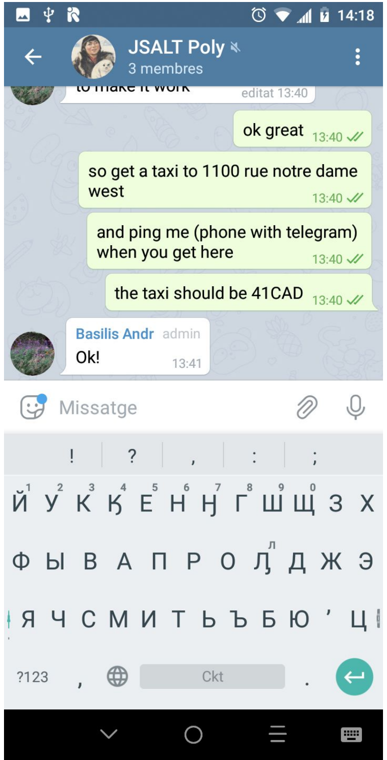
BPE	NMT	SMT
500	16.7	13.1
1000	16.8	14.8
5000	16.9	15.2
15000	17.0	15.1
30000	16.7	15.2
60000	16.8	15.5

Baseline system: ASR

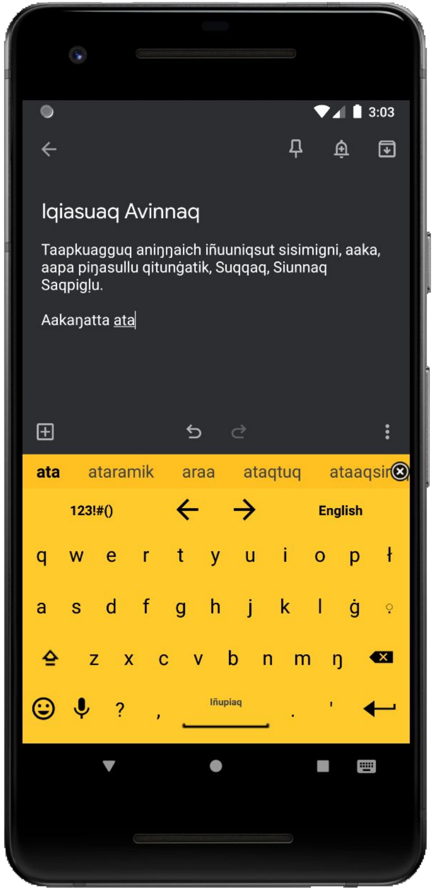
- DeepSpeech ASR baseline for Crow
- Forced aligner trained for SLI Yupik

Downstream applications:

On-device predictive text



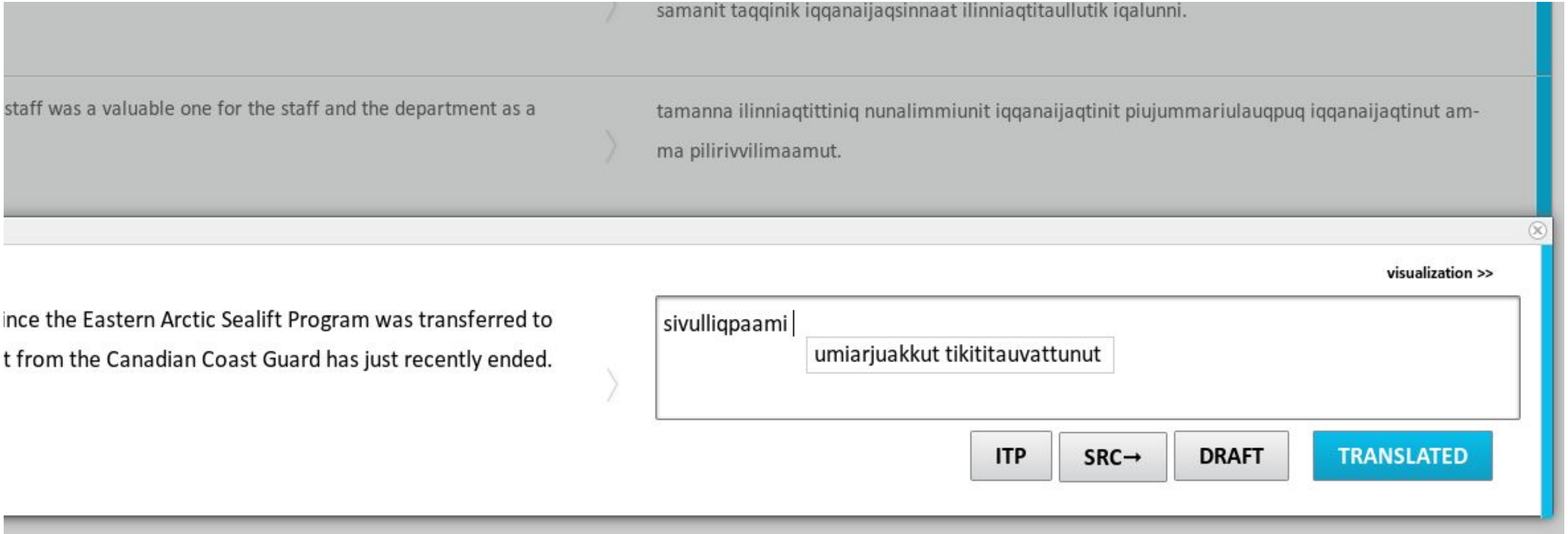
Giellatekno



Downstream applications: Spell-checkers

Giellatekno

Downstream applications: Interactive MT



Downstream applications: Audio transcription / search

Workshop goals

- What embedding representations and neural LM architectures are most effective for languages with a high ratio of morphemes per word?

- Can rule-based FSTs be used to bootstrap more robust neural models?

- To what extent can common neural models be developed for closely related languages in order to maximize the utility of sparse digitized resources?

NPLM Team Members

- **Lane Schwartz (University of Illinois)**
- **Francis Tyers (Indiana University)**

- Lori Levin (CMU)
- Christo Kirov (Google AI)
- Patrick Littell (NRC Canada)
- Jackie Lo (NRC Canada)
- Emily Prud’hommeaux (Boston College)

- **Hayley Park (Illinois)**
- **Kenneth Steimel (Indiana)**
- **Lonny Strunk (Washington)**
- **Coleman Hayley (JHU)**
- **Katherine Zhang (CMU)**

- Jeffrey Micher (CMU)
- Rebecca Knowles (JHU)
- Robbie Jimmerson (BC)
- Vasya Adriyanets (HSE)
- Han Liu (UC Boulder)

Let’s learn lessons that will apply to the >6700 human languages that are not like English