Neural Polysynthetic Language Modelling

2019 | INTERNATIONAL YEAR OF
Indigenous Languages

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

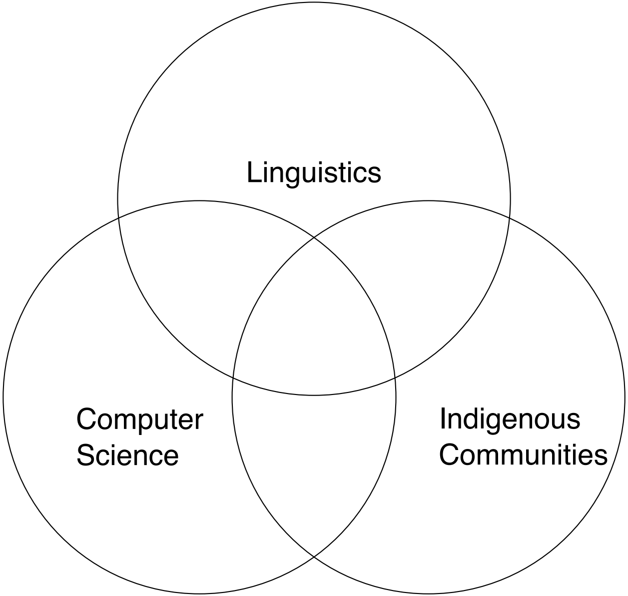
2019 | INTERNATIONAL YEAR OF Indigenous Languages

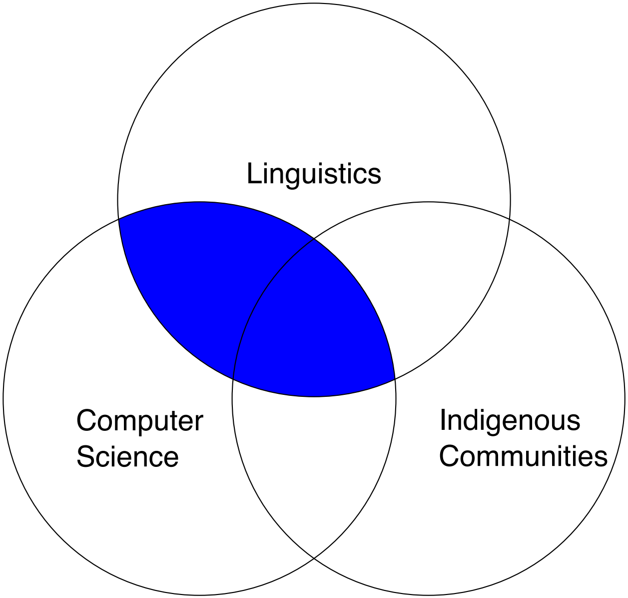
- Increasing understanding, reconciliation and international cooperation.
- Creation of favorable conditions for knowledge-sharing & dissemination of good practices with regards to indigenous languages.
- Integration of indigenous languages into standard setting.
- Empowerment through capacity building.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

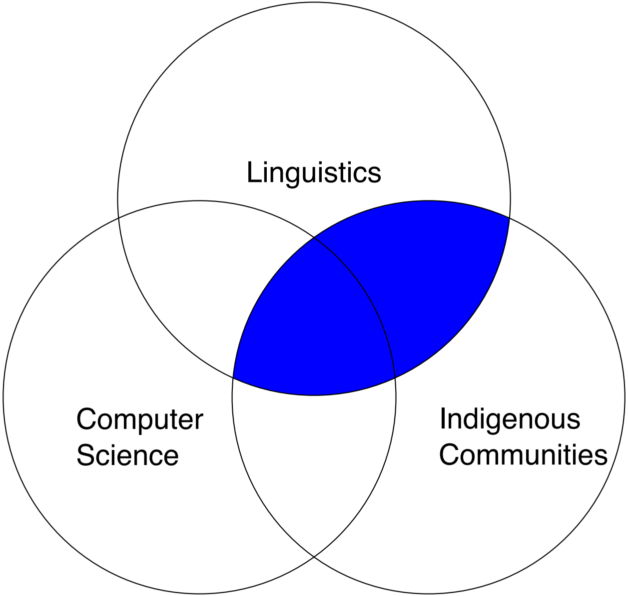
2019 | INTERNATIONAL YEAR OF

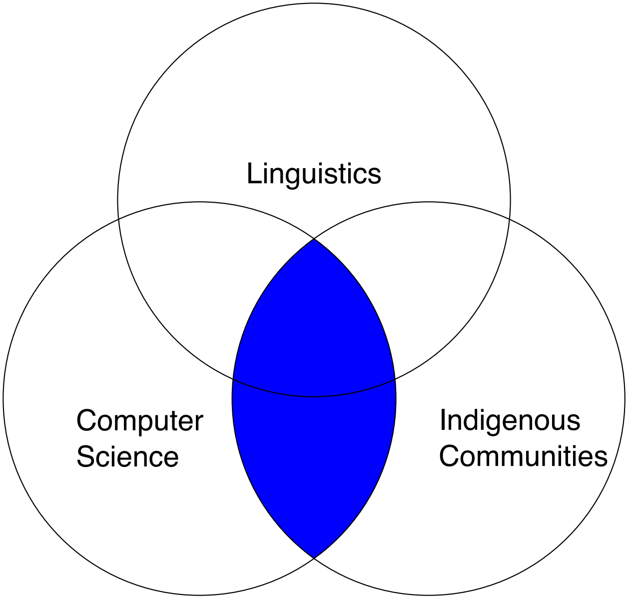
Indigenous Languages

- 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.

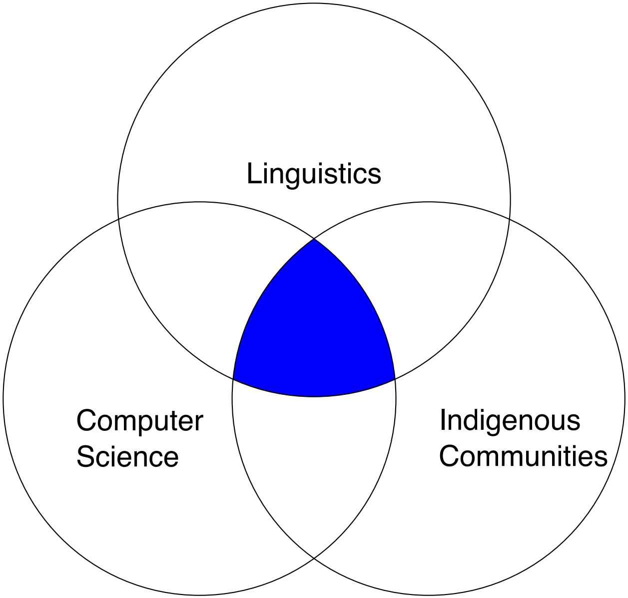

Linguistics
Computer Science
Indigenous Communities

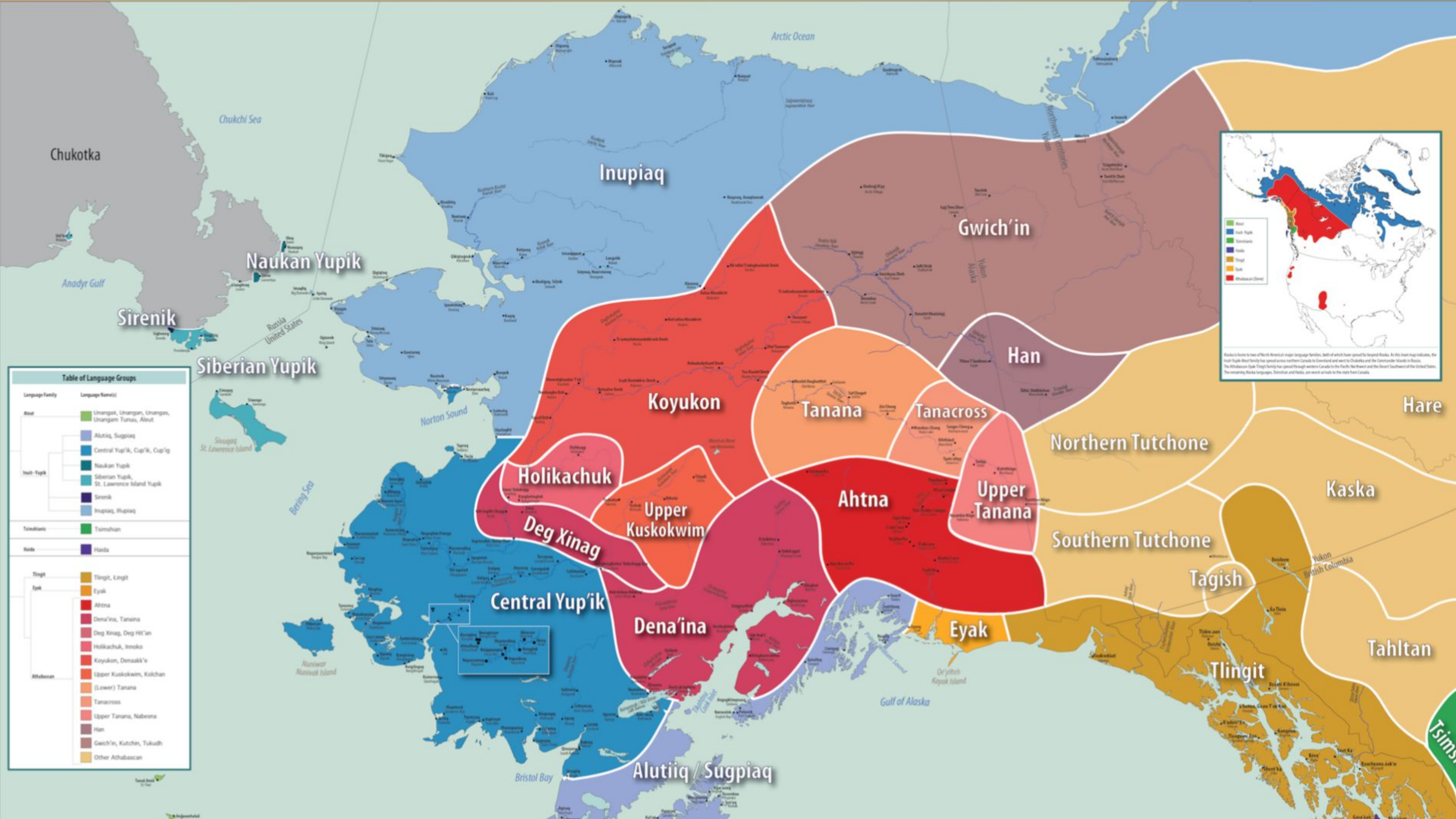

Linguistics
Computer Science
Indigenous Communities


Linguistics
Computer Science
Indigenous Communities

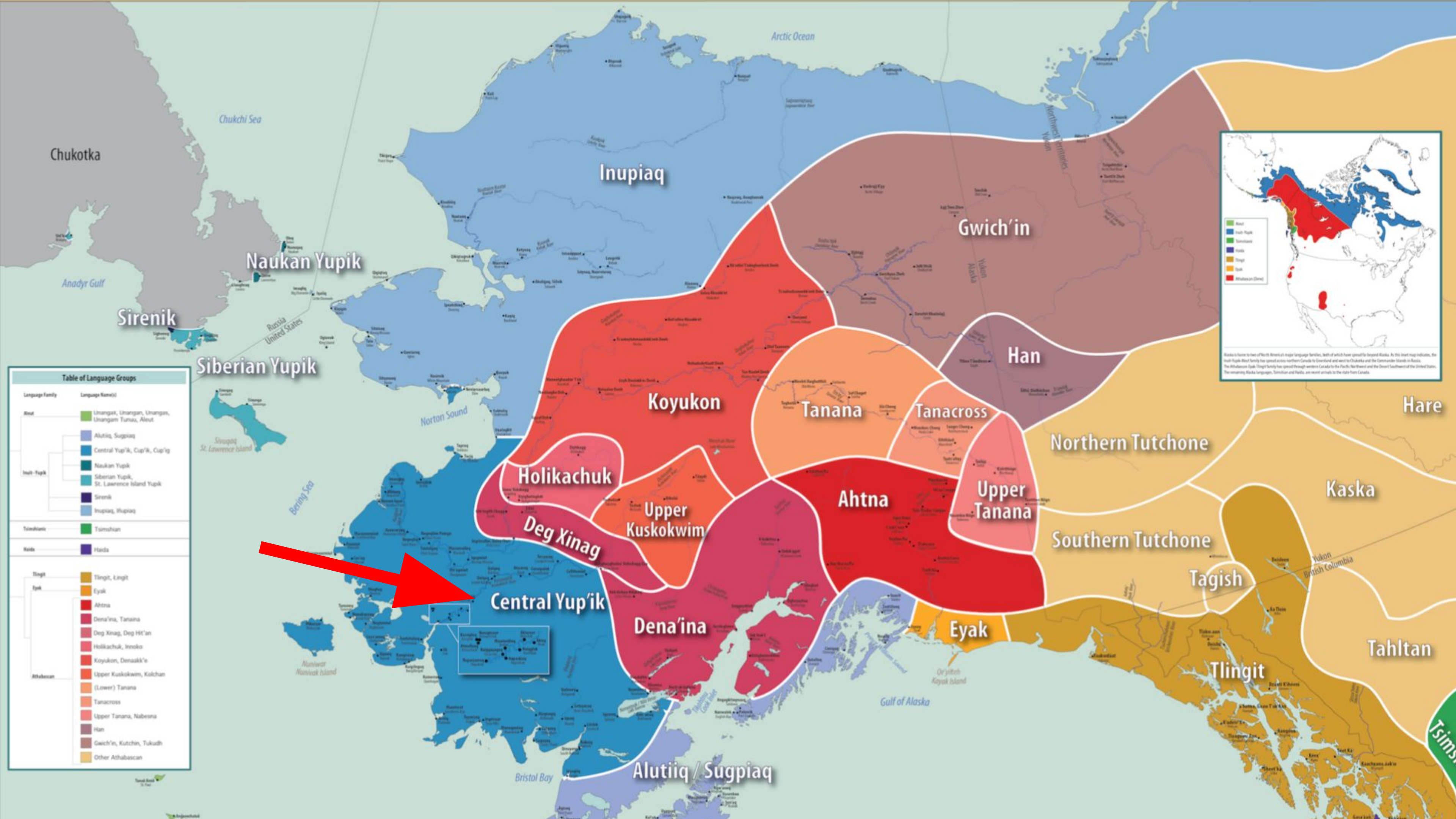


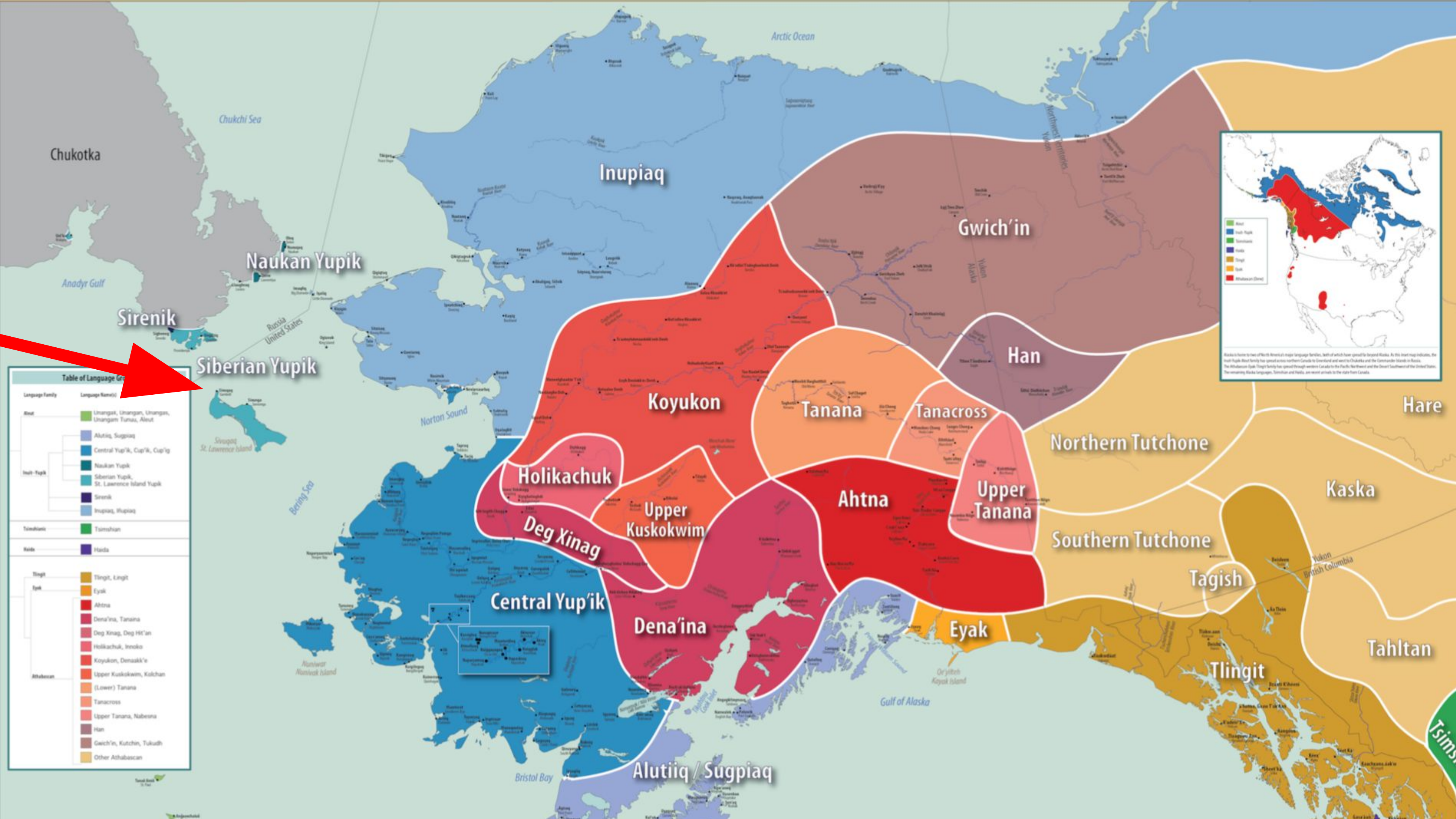
Linguistics
Computer Science
Indigenous Communities


Linguistics
Computer Science
Indigenous Communities



Arctic Ocean
Chukchi Sea
Chukotka
Anadyr Gulf
Russia
United States
Bering Sea
Norton Sound
Bristol Bay
Gulf of Alaska
Sivuqaq
St. Lawrence Island
Nuniwar
Nunivak Island
Qe'yileh
Kayak Island
Inupiaq
Naukan Yupik
Sirenik
Siberian Yupik
Gwich'in
Han
Koyukon
Tanana
Tanacross
Holikachuk
Upper Kuskokwim
Deg Xinag
Ahtna
Upper Tanana
Central Yup'ik
Dena'ina
Eyak
Alutiiq / Sugpiaq
Northern Tutchone
Southern Tutchone
Hare
Kaska
Tagish
Tahltan
Tlingit
Yukon
British Columbia
Northwest Territories
Yukon
Alaska
Table of Language Groups
Language Family
Language Name(s)
Aleut
Unangax, Unangan, Unangas, Unangam Tunuu, Aleut
Alutiiq, Sugpiaq
Central Yup'ik, Cup'ik, Cup'ig
Inuit - Yupik
Naukan Yupik
Siberian Yupik, St. Lawrence Island Yupik
Sirenik
Inupiaq, Iñupiaq
Tsimshianic
Tsimshian
Haida
Haida
Tlingit
Tlingit, Łingit
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

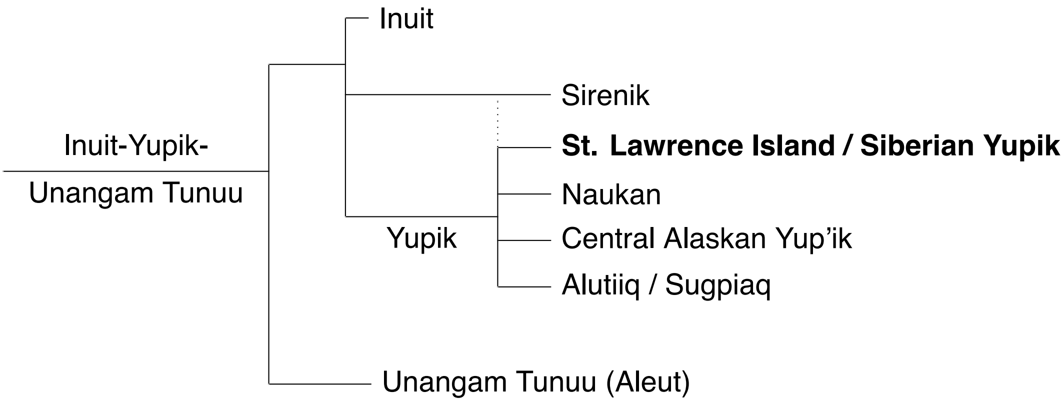

Arctic Ocean
Chukchi Sea
Chukotka
Anadyr Gulf
Inupiaq
Naukan Yupik
Sirenik
Siberian Yupik
Russia
United States
Norton Sound
Bering Sea
Sivuqaq
St. Lawrence Island
Gwich'in
Han
Koyukon
Tanana
Tanacross
Holikachuk
Upper Kuskokwim
Deg Xinag
Ahtna
Upper Tanana
Northern Tutchone
Southern Tutchone
Hare
Kaska
Tagish
Yukon
British Columbia
Central Yup'ik
Dena'ina
Eyak
Tlingit
Tahltan
Alutiiq / Sugpiaq
Nuniwar
Nunivak Island
Bristol Bay
Gulf of Alaska
Kayak Island
Table of Language Groups
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, Lingít
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


Arctic Ocean
Chukchi Sea
Chukotka
Anadyr Gulf
Inupiaq
Gwich'in
Naukan Yupik
Sirenik
Russia
United States
Siberian Yupik
Sivuqaq
St. Lawrence Island
Norton Sound
Bering Sea
Koyukon
Tanana
Tanacross
Han
Holikachuk
Upper Kuskokwim
Deg Xinag
Ahtna
Upper Tanana
Northern Tutchone
Southern Tutchone
Kaska
Hare
Tagish
Tahltan
Central Yup'ik
Dena'ina
Eyak
Tlingit
Alutiiq / Sugpiaq
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

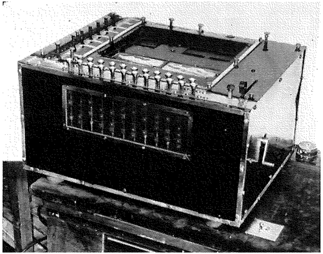
Inuit-Yupik-Unangam Tunuu language family

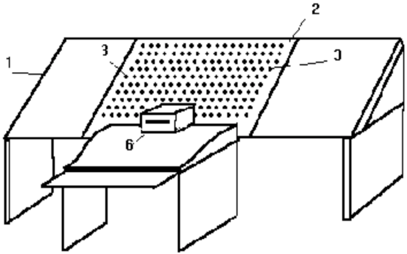
- Greenland (Inuit)
- Northern Canada (Inuit)
- Northern Alaska (Inuit)
- Western Alaska (Yup'ik)
- Southwestern Alaska (Sugpiak, Unangam Tunuu)
- **St. Lawrence Island (Yupik)**
- Big Diomede (Inuit)
- Far eastern Russia (Yupik, Sirenik)

Inuit-Yupik-Unangam Tunuu language family

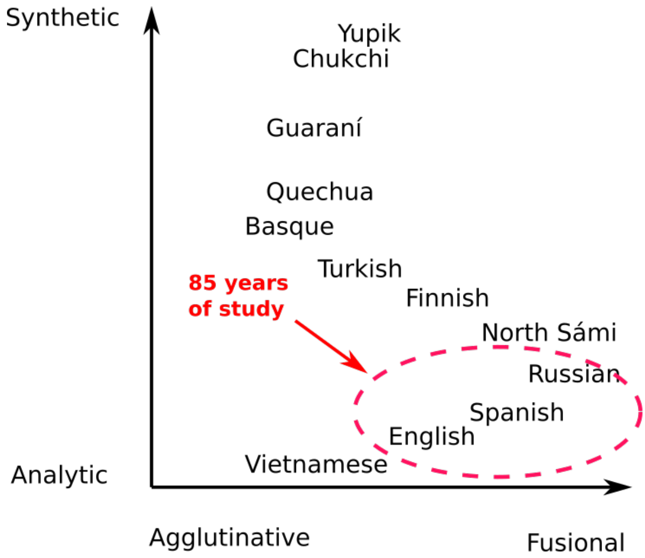


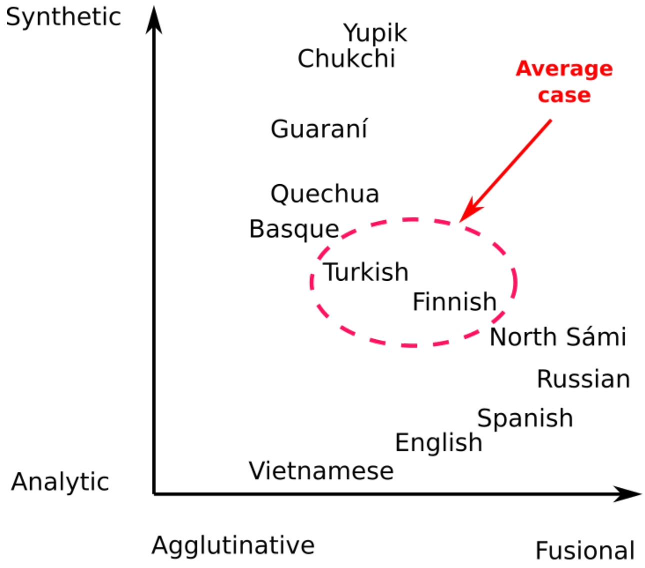
Since 1933, NLP technology has overwhelmingly focused on languages & methodologies in which the word is the primary meaning-bearing unit

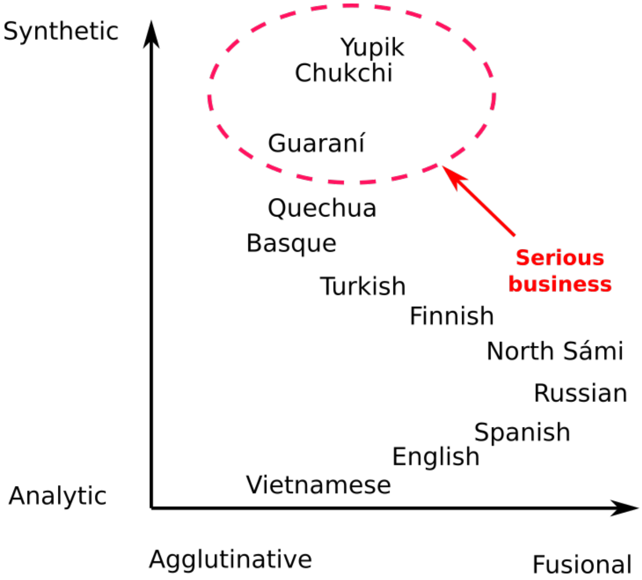




For ***most*** human languages, this assumption is **fundamentally broken**

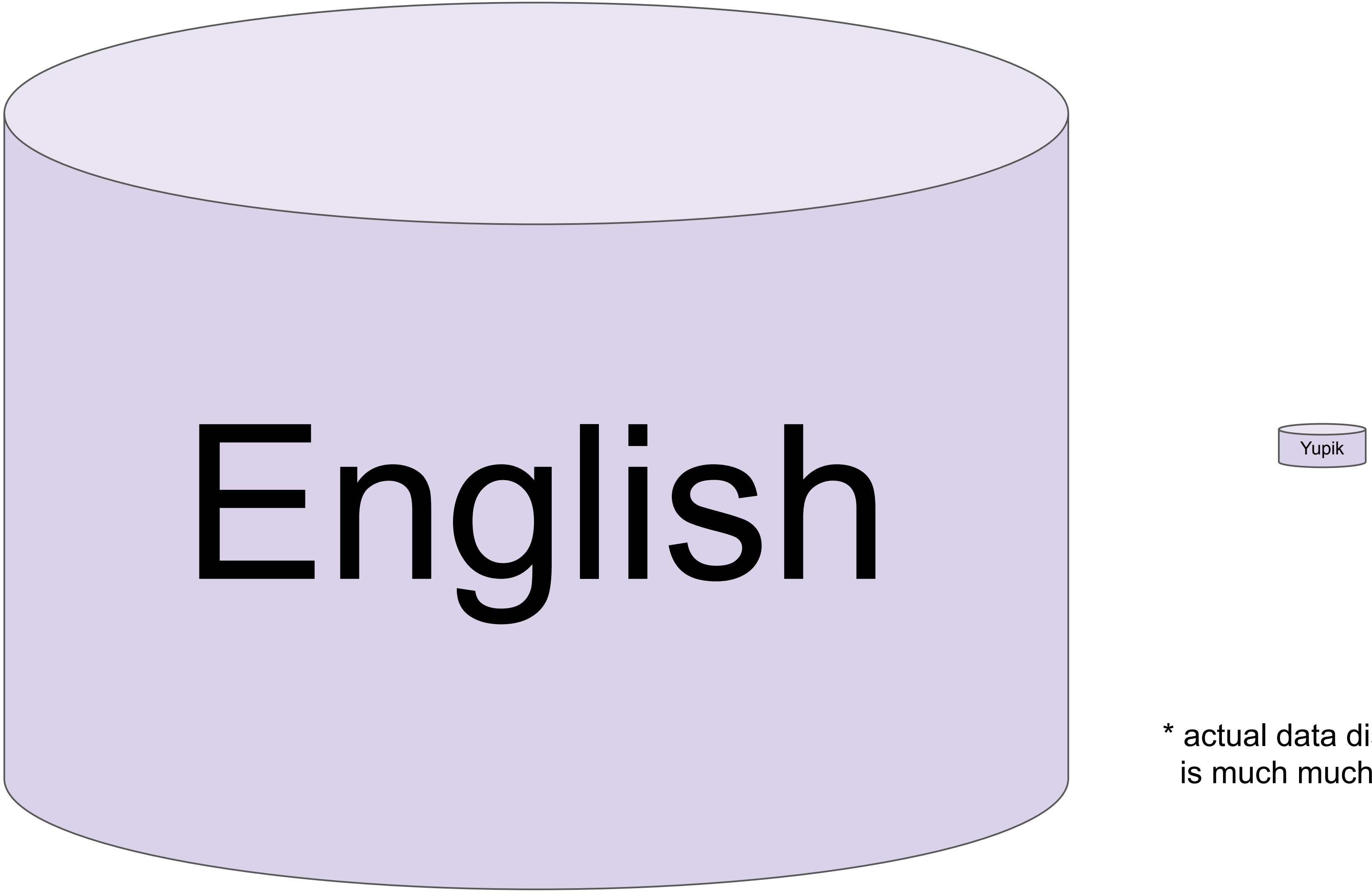

Synthetic
Yupik
Chukchi
Guaraní
Quechua
Basque
Turkish
85 years
of study
Finnish
North Sámi
Russian
Spanish
English
Vietnamese
Analytic
Agglutinative
Fusional


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
Serious business
Turkish
Finnish
North Sámi
Russian
Spanish
English
Vietnamese
Analytic
Agglutinative
Fusional

$$p\left(\tau_t \middle| \tau_1 \cdots \tau_{t-1}\right)$$

$$p(\tau_t|\tau_1 \ldots \tau_{t-1}) = \frac{\text{count}(\tau_1 \ldots \tau_t)}{\text{count}(\tau_1 \ldots \tau_{t-1})}$$



* actual data disparity
is much much larger

dog

dogs

qikmiq

qikmik

qikmit

qikmiq

qikmik

qikmit

qikmigka

qikmigken

qikminka

qikmii

qikmikek

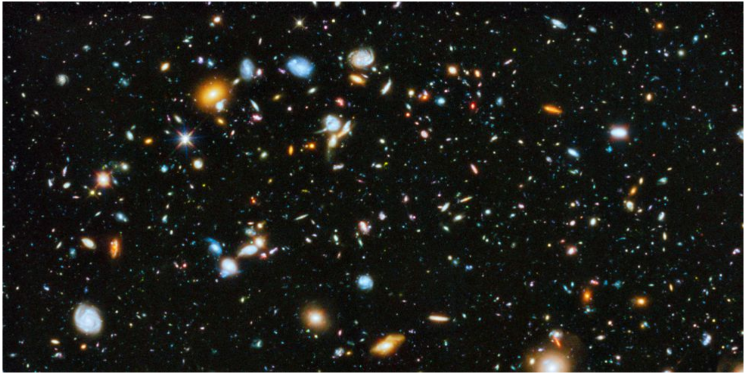
qikmiik

qikmiqa

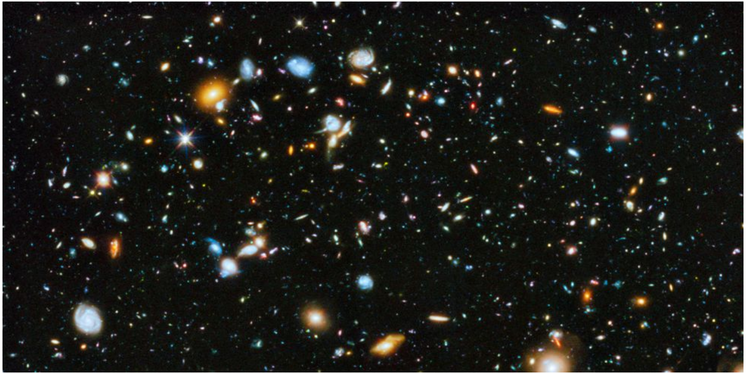
qikmighpung

qikmighput

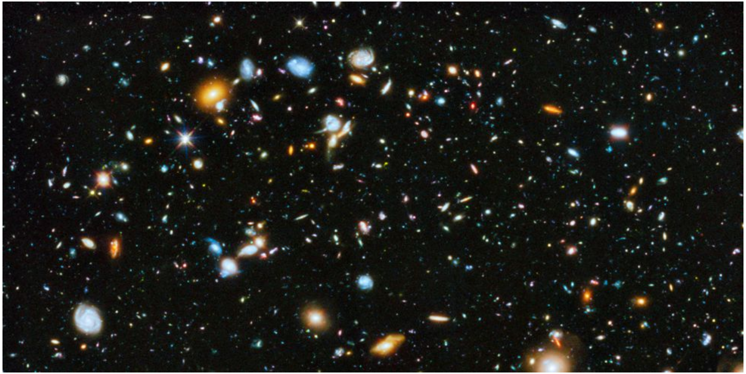
There are 1.2×10^{23} stars in the observable universe.



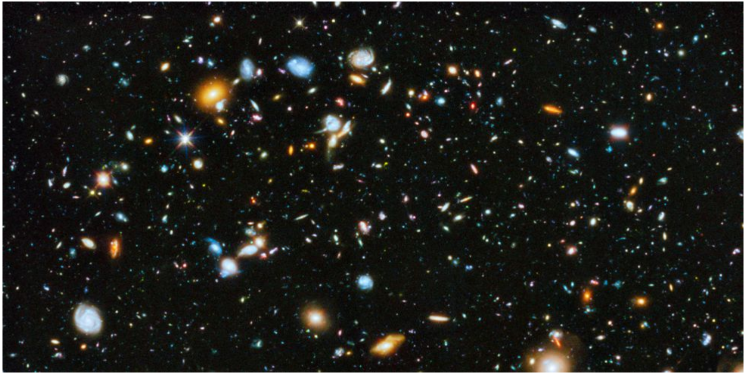
There are 1.2×10^{23} possible Yupik word forms.

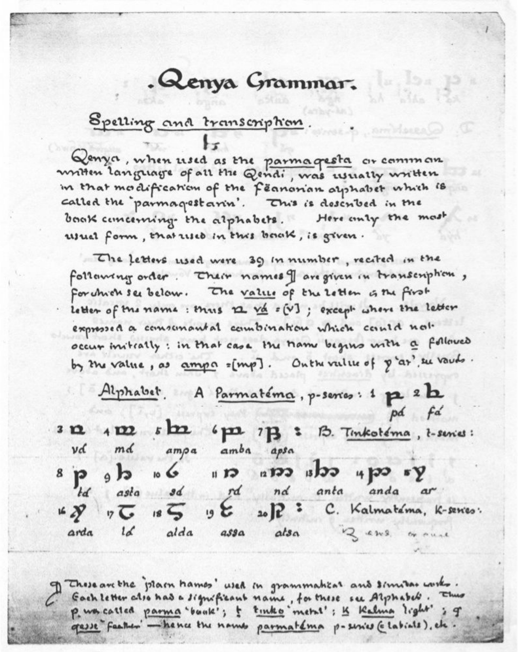


Big data is NOT the solution.



Modelling only at the word-level is like modelling only at a galaxy-level.





.Qenya Grammar.

Spelling and transcription.

Qenya, when used as the parmaqesta or common written language of all the Qendi, was usually written in that modification of the Fëanorian alphabet which is called the 'parmaqestarin'. This is described in the book concerning the alphabets. Here only the most usual form, that used in this book, is given.

The letters used were 39 in number, recited in the following order. Their names ¶ are given in transcription, for which see below. The value of the letter is the first letter of the name: thus vá = [v]; except where the letter expressed a consonantal combination which could not occur initially: in that case the name begins with a followed by the value, as ampa = [mp]. On the value of 'ar', see vowels.

Alphabet. A Parmatéma, p-series: 1 pá 2 fá 3 vá 4 má 5 ampa 6 amba 7 apsa : B. Tinkotéma, t-series: 8 tá 9 asta 10 sá 11 rá 12 ná 13 anta 14 anda 15 ar 16 arda 17 lá 18 alda 19 assa 20 atsa : C. Kalmatéma, k-series:

¶ These are the 'plain names' used in grammatical and similar works. Each letter also had a significant name, for these see Alphabets. Thus p was called parma 'book'; t tinko 'metal'; k kalma 'light'; q qesse 'feather' — hence the names parmatéma p-series (= labials), etc.

PARMA ELDALAMBERON XXII

The Feanorian Alphabet · PART 1

❖

Quenya Verb Structure

by J. R. R. TOLKIEN

+

=

+

=

ash	nazg	durb-	-at-	-ul-	-ûk
one	ring	to.dominate	[Ptcp]	[3Pl]	[Compl]

...

agh	burzum	ishi	krimp-	-at-	-ul
and	darkness	inside	to.bind	[Ptcp]	[3Pl]

Course goals

- Learn about a new language from a reference grammar
  - Demonstrate your understanding through writing and teaching
- Select a topic from computational linguistics applicable to this language
  - Conduct a literature review, resulting in an annotated bibliography & report on state of the art
- Perform research on this topic
  - Identify state-of-the-art baseline, implement & extend it, run experiments, write a paper
- Conduct extended research in a group
  - Collaborate, experiment, and jointly author a paper
- Act as a peer reviewer for your classmates’ work

Demographics

- 1300 Yupiget on St. Lawrence Island
- 800 Yupiget on Russian mainland
- 300-400 Yupiget on Alaskan mainland

Education

- 1930s-1950s Yupik materials developed in Russia
- 1970s-1990s Yupik materials developed in Alaska

- By mid-20th century, shift away from Yupik in Russia
- Current estimate of < 200 L1 Yupik speakers in Russia
- Youngest L1 Russian Yupiget estimated age > 70

Language shift - St. Lawrence Island

- In 1980, nearly all St. Lawrence Island Yupiget children spoke Yupik at home
- By mid-1990s through early 2000s, shift away from Yupik among SLI youth
- All SLI Yupiget born 1980 or earlier assumed to be L1 Yupik
- Current estimate of at least 540 L1 Yupik speakers on SLI
- Youngest L1 SLI Yupiget not known

Phonology & Orthography

Close Vowels	**i**	**u**	Latin
	i	u	IPA
	и	у	Cyrillic
Mid Vowel	**e**		Latin
	ə		IPA
	ы		Cyrillic
Open Vowel	**a**		Latin
	ɑ		IPA
	а		Cyrillic

Syllable structure

- Word-initial V(C)
- Otherwise CV(C)

- V may be short (e, a, i, u) or long (aa, ii, uu)
- Adjacent consonants only at syllable boundaries
- Adjacent consonant generally must agree in voicing

Phonology & Orthography

	Labial	Alveolar		Palatal	Retroflex	Velar	Velar (rounded)	Uvular	Uvular (rounded)	Glottal	
Unvoiced Stops	**p**	**t**				**k**	**kw**	**q**	**qw**		Latin
	p	t				k	k^w	q	q^w		IPA
	п	**т**				**к**	**кӱ**	**ӄ**	**ӄӱ**		Cyrillic
Voiced Continuants	**v**	**l**	**z**	**y**	**r**	**g**	**w**	**gh**	**ghw**		Latin
	v	l	z	j	ɻ	ɣ	$ɣ^w$	ʁ	$ʁ^w$		IPA
	в	**л**	**з**	**й**	**р**	**г**	**(г)ӱ**	**ӷ**	**ӷӱ**		Cyrillic
Unvoiced Continuants	**f**	**ll**		**s**	**rr**	**gg**	**wh**	**ghh**	**ghhw**	**h**	Latin
	f	ɬ		s	ʂ̥	x	x^w	χ	$χ^w$	h	IPA
	ф	**ль**		**с**	**ш**	**х**	**хӱ**	**ӽ**	**ӽӱ**	**г**	Cyrillic
Voiced Nasals	**m**	**n**				**ng**	**ngw**				Latin
	m	n				ŋ	$ŋ^w$				IPA
	м	**н**				**ӈ**	**ӈӱ**				Cyrillic
Unvoiced Nasals	**mm**	**nn**				**ngng**	**ngngw**				Latin
	m̥	n̥				ŋ̊	$ŋ̊^w$				IPA
	мь	**нь**				**ӈь**	**ӈьӱ**				Cyrillic

Legacy Digitization

- 3-volume Lore of St. Lawrence Island
- 3-volume Elementary Yupik readers
- 1-volume of Russian Yupik stories

Intersecting machine learning & linguistic fieldwork

- Yupik is polysynthetic, allowing for morphologically-complex words

(1) **mangteghaghllangllaghyugtukut**

mangteghagh-	-ghllag-	-ngllagh-	-yug-	-tu-	-kut
house-	-big-	-build-	-want.to-	-INTR.IND-	-1PL

'*We want to build a big house*'

- Yupik words typically adhere to the following template:

Root + 0-7 Derivational Morpheme(s) + Inflectional Morphemes + (Enclitic)

- Yupik is polysynthetic, allowing for morphologically-complex words

(1) **mangteghaghllangllaghyugtukut**

mangteghagh-	-ghllag-	-ngllagh-	-yug-	-tu-	-kut
house-	-big-	-build-	-want.to-	-INTR.IND-	-1PL

‘*We want to build a big house*’

- Yupik words typically adhere to the following template:

Root + 0-7 Derivational Morpheme(s) + Inflectional Morphemes + (Enclitic)

- Yupik is polysynthetic, allowing for morphologically-complex words

(1) **mangteghaghllangllaghyugtukut**

mangteghagh-	-ghllag-	-ngllagh-	-yug-	-tu-	-kut
house-	-big-	-build-	-want.to-	-INTR.IND-	-1PL

‘*We want to build a big house*’

- Yupik words typically adhere to the following template:

Root + 0-7 Derivational Morpheme(s) + Inflectional Morphemes + (Enclitic)

- Yupik is polysynthetic, allowing for morphologically-complex words

(1) **mangteghaghllangllaghyugtukut**

mangteghagh-	-ghllag-	-ngllagh-	-yug-	-tu-	-kut
house-	-big-	-build-	-want.to-	-INTR.IND-	-1PL

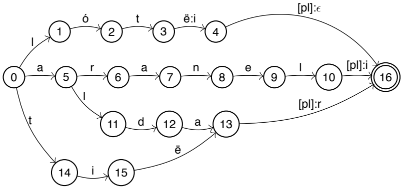
'*We want to build a big house*'

- Yupik words typically adhere to the following template:

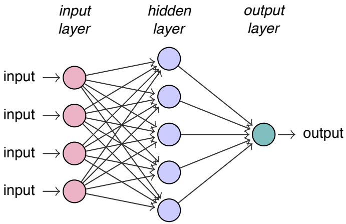
Root + 0-7 Derivational Morpheme(s) + Inflectional Morphemes + (Enclitic)

- Morphological analyzers may be implemented as a

Finite-State Transducer



Neural Network



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

- **OBJECTIVE**: Analyze inflected Yupik nouns with no derivational morphology
- **TRAINING DATA**: Every nominal surface form and its respective lexical form
  - 3873 Yupik noun roots
  - 273 inflectional suffixes
  - $3873 \times 273 =$ **1,057,329 total nouns**
  - 658,410 after removing duplicate surface forms (case syncretism)

Surface Form	Lexical Form
mangteghaq	mangteghagh[N][ABS][SG]
mangteghaat	mangteghagh[N][ABS][PL]
mangteghaak	mangteghagh[N][ABS][DU]
mangteghaa	mangteghagh[N][ABS][SG][3SGPOSS]
⋮	

- **EVALUATION OBJECTIVES**
  - Evaluate on a neutral dataset
  - Contrast performance with the FST analyzer
- **NEUTRAL DATASET**: *Mrs. Della Waghiyi's St. Lawrence Island Yupik Texts With Grammatical Analysis* (Waghiyi & Nagai, 2001)
  - Identified **344 inflected nouns with no derivational morphology**
- Supplemented the FST analyzer with a guesser module

- Results:

	Coverage (%)	Accuracy (%)
FST (No Guesser)	85.96	79.82
FST (w/Guesser)	100	84.50
Neural	100	**91.81**

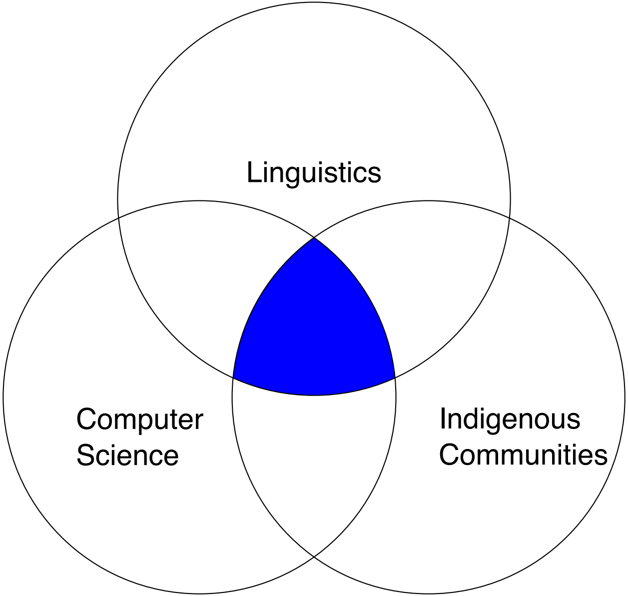
- An **out-of-vocabulary** (OOV) root is an unattested root that appears in the Waghiyi & Nagai (2001) evaluation dataset but does not appear in our data

OOV Root	FST	NN
aghnasinghagh	–	–
aghveghniigh	–	✓
akughvigagh	✓	✓
qikmiraagh	–	–
sakara	✓	–
sanaghte	–	–
tangiqagh	–	✓

- A root with a **spelling variant** is one that differs in the Waghiyi & Nagai (2001) evaluation set from its form in our data

Root Variant	FST	NN
melq**i**ghagh	✓	✓
piites**ii**ghagh	–	✓
uqf**ii**lleghagh	–	✓
*uk**u**sumun	–	✓

Building a virtuous cycle


Linguistics
Computer Science
Indigenous Communities

- Digitization of legacy materials
- Pedagogical materials & tools
- Orthographic experimentation
- Identify under-described phenomena
- Real-time morphological analysis

- Digitization of legacy materials
- Pedagogical materials & tools
- Orthographic experimentation
- Identify under-described phenomena
- Real-time morphological analysis

- Digitization of legacy materials
- Pedagogical materials & tools
- Orthographic experimentation
- Identify under-described phenomena
- Real-time morphological analysis

- Digitization of legacy materials
- Pedagogical materials & tools
- Orthographic experimentation
- Identify under-described phenomena
- Real-time morphological analysis

- Digitization of legacy materials
- Pedagogical materials & tools
- Orthographic experimentation
- Identify under-described phenomena
- Real-time morphological analysis

Moving forward

$$p(\boldsymbol{e}) = p(e_t | e_1 \ldots e_{t-1})$$

$$p(\boldsymbol{e}) = p(e_t | e_1 \ldots e_{t-1})$$
$$\approx p(e_t | e_{t-1})$$

1. Zero-order approximation

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGXYD QPAAMKBZAACIBZLHJQD

2. First-order approximation

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL

3. Second-order approximation

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE

5. First-Order Word Approximation

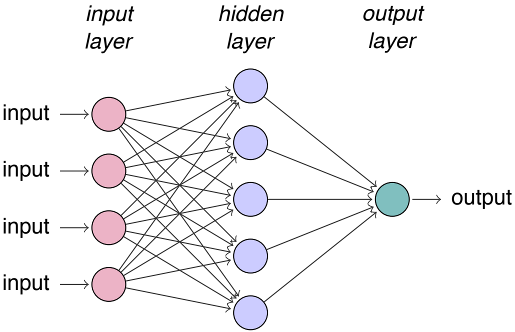
REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. Second-Order Word Approximation

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

$$
\begin{aligned}
p(\boldsymbol{e}) &= p(e_t | e_1 \ldots e_{t-1}) \\
&\approx p(e_t | e_{t-1})
\end{aligned}
$$

$$p(\boldsymbol{e}) = p(e_t | e_1 \ldots e_{t-1})$$


input layer
hidden layer
output layer
input
input
input
input
output

- Legacy text digitization
- Web portal / interactive e-books
- App-based dictionary
- Language learning lessons
- foma-based spell-checker
- Forced aligner / speech recognizer
- Machine translation

Feature-rich Open-vocabulary Interpretable Language Modelling

Interpretable Tensor Morpheme Representation

ɣaławtəma

"with the head"

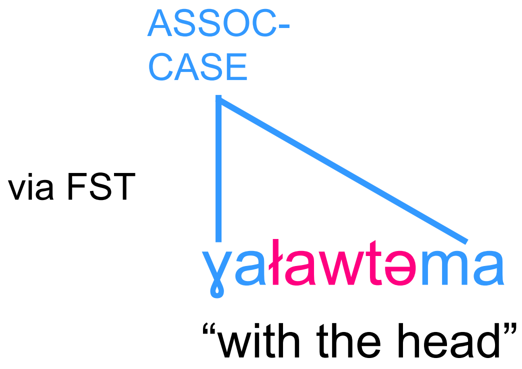
-lewət- "head" (Chukchi)

Interpretable Tensor Morpheme Representation

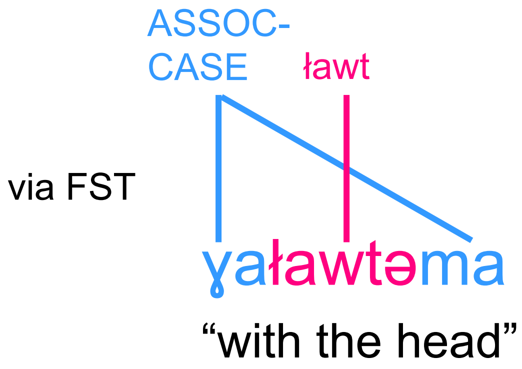
ɣaławtəma

“with the head”

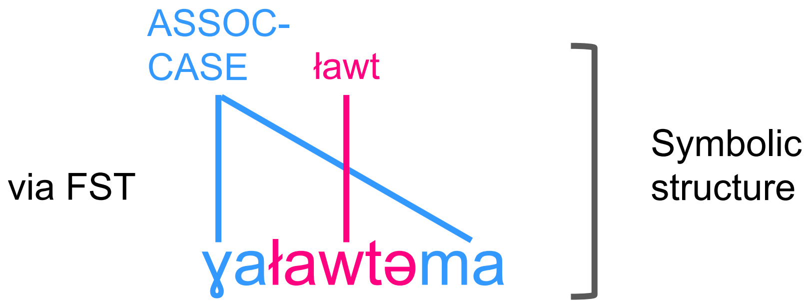
Interpretable Tensor Morpheme Representation



Interpretable Tensor Morpheme Representation

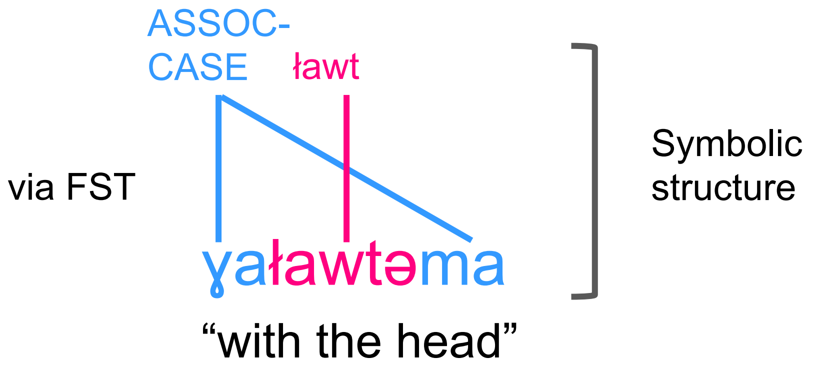


Interpretable Tensor Morpheme Representation



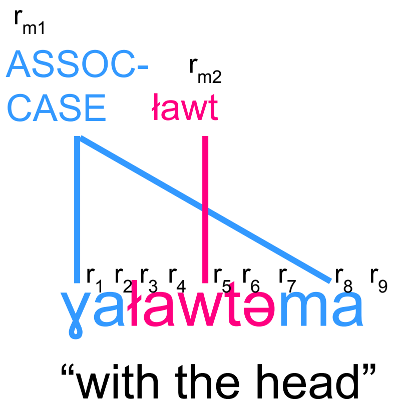
"with the head"

Interpretable Tensor Morpheme Representation



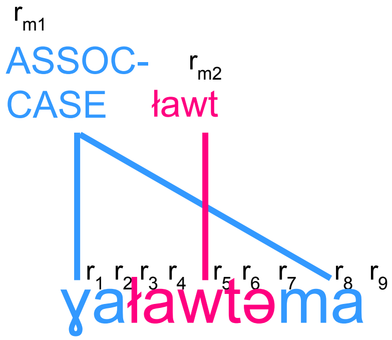
Symbolic structure → Tensor representation → Vector representation

Interpretable Tensor Morpheme Representation



Decompose into *fillers* and *roles.*
(Smolenksy 1990)

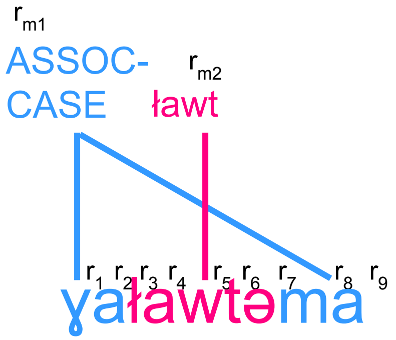
Interpretable Tensor Morpheme Representation



"with the head"

Embed the fillers and roles into *vectors*
(Smolenksy 1990)

Interpretable Tensor Morpheme Representation

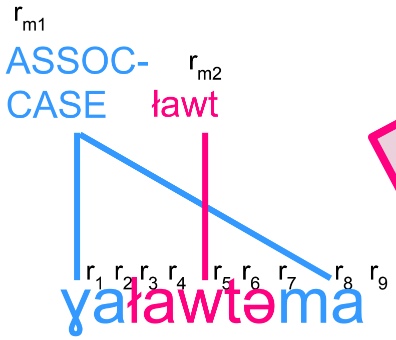


“with the head”

Embed the fillers and roles into *vectors*
(Smolenksy 1990)

$$Repr(\text{ɣa-ma}) = (\hat{\text{ɣ}} \otimes \hat{r}_1 + \hat{\text{a}} \otimes \hat{r}_2 + \hat{\text{m}} \otimes \hat{r}_8 + \hat{\text{a}} \otimes \hat{r}_9) \otimes \hat{r}_{m_1}$$

Tensor Morpheme Representation



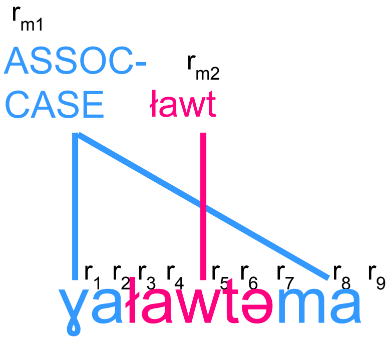
"with the head"

This means we have embeddings that are interpretable! ^_^

Embed the fillers and roles into *vectors*
(Smolenksy 1990)

$$Repr(\text{ɣa-ma}) = (\hat{\text{ɣ}} \otimes \hat{r}_1 + \hat{\text{a}} \otimes \hat{r}_2 + \hat{\text{m}} \otimes \hat{r}_8 + \hat{\text{a}} \otimes \hat{r}_9) \otimes \hat{r}_{m_1}$$

Interpretable Tensor Morpheme Representation



“with the head”

1. Deterministically create these with FST for known sequences

2. Learn them with neural model (e.g. RNN seq2seq) to generalize

Embed the fillers and roles into *vectors*
(Smolenksy 1990)

Deterministically construct morpheme tensors

a. Run morphological analyzer on training data to identify morphemes

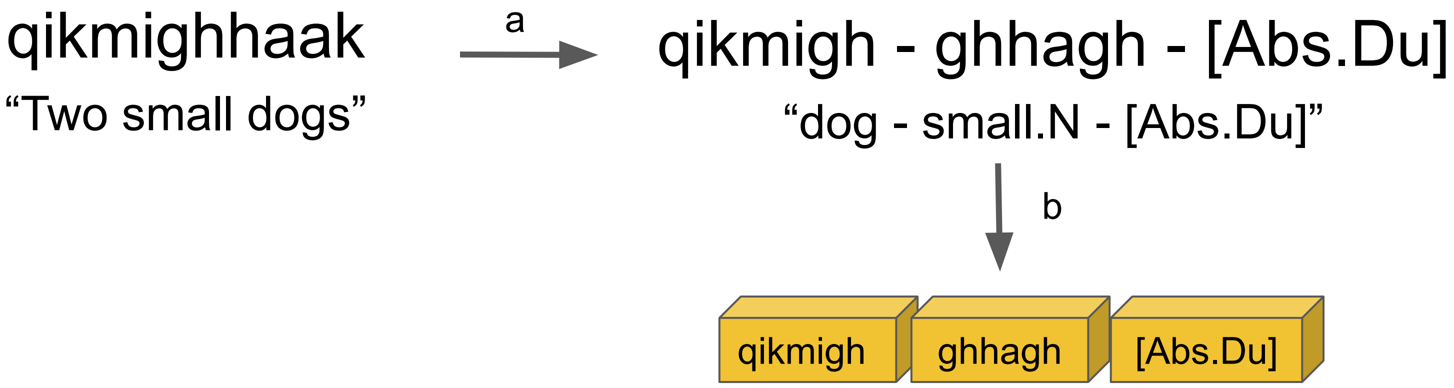
qikmighhaak —a→ qikmigh - ghhagh - [Abs.Du]

"Two small dogs"   "dog - small.N - [Abs.Du]"

(St. Lawrence Island Yupik)

Deterministically construct morpheme tensors

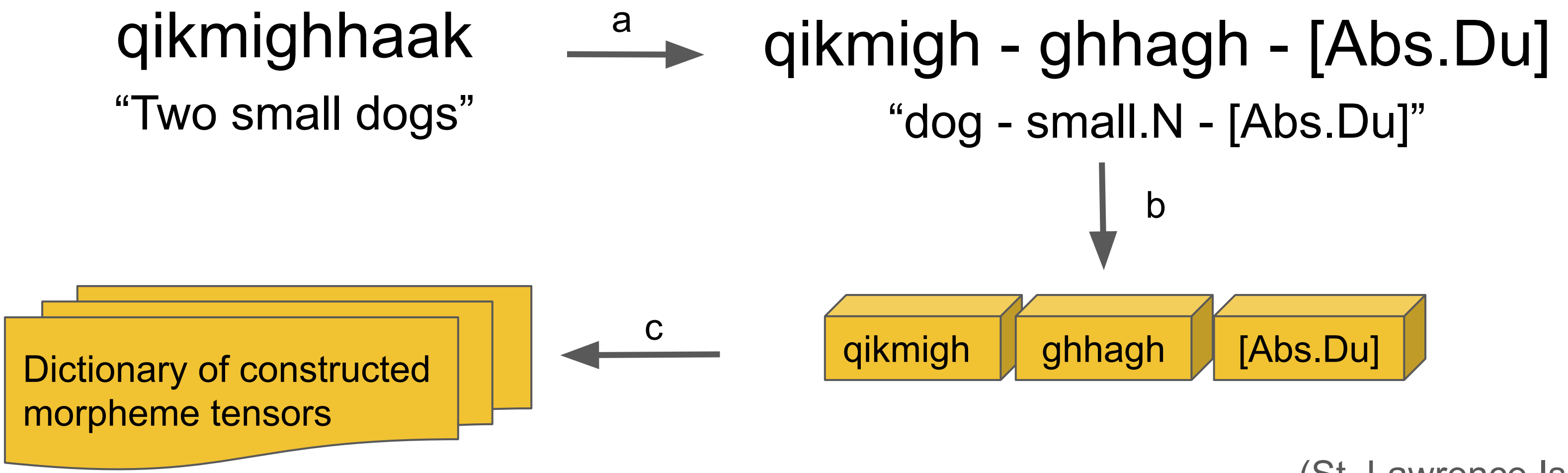
a. Run morphological analyzer on training data to identify morphemes
b. Use **Tensor Product Representation** to deterministically calculate **morpheme tensors**



(St. Lawrence Island Yupik)

Deterministically construct morpheme tensors

a. Run morphological analyzer on training data to identify morphemes
b. Use **Tensor Product Representation** to deterministically calculate **morpheme tensors**
c. Save these morpheme tensors for later use as gold standard labels



(St. Lawrence Island Yupik)

Autoencoder

Dictionary of constructed morpheme tensors

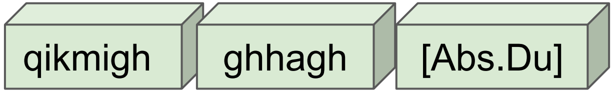
High dimensionality:
10^3 - 10^9 floats per vector

qikmigh ghhagh [Abs.Du]

(St. Lawrence Island Yupik)

Dictionary of constructed morpheme tensors

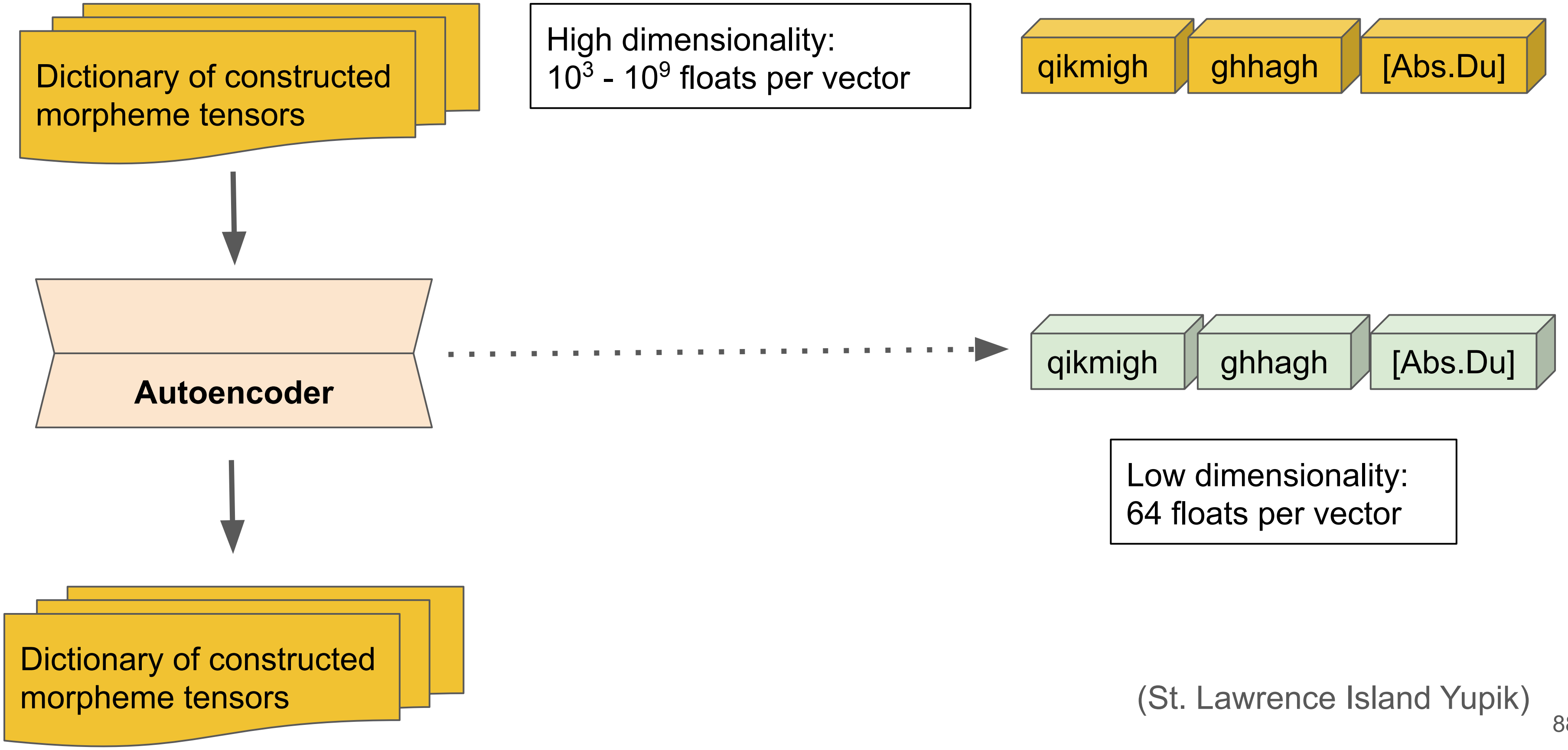
High dimensionality:
10^3 - 10^9 floats per vector



Low dimensionality:
64 floats per vector

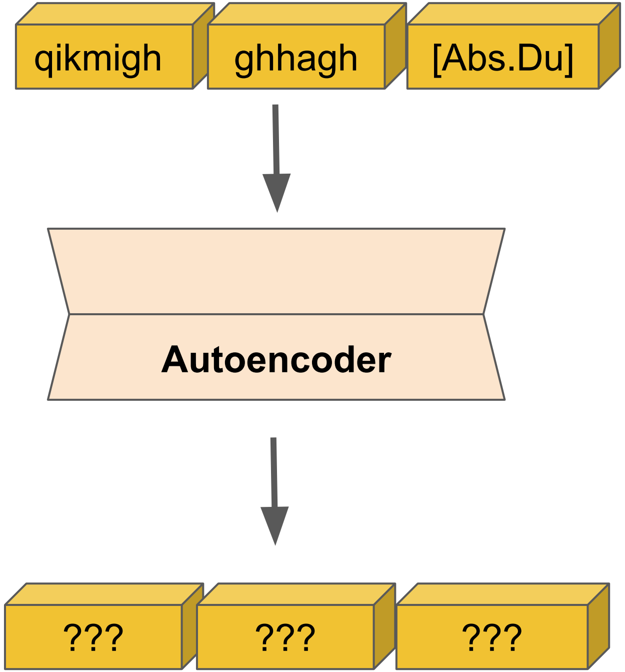
(St. Lawrence Island Yupik)

Use autoencoder to learn morpheme vectors



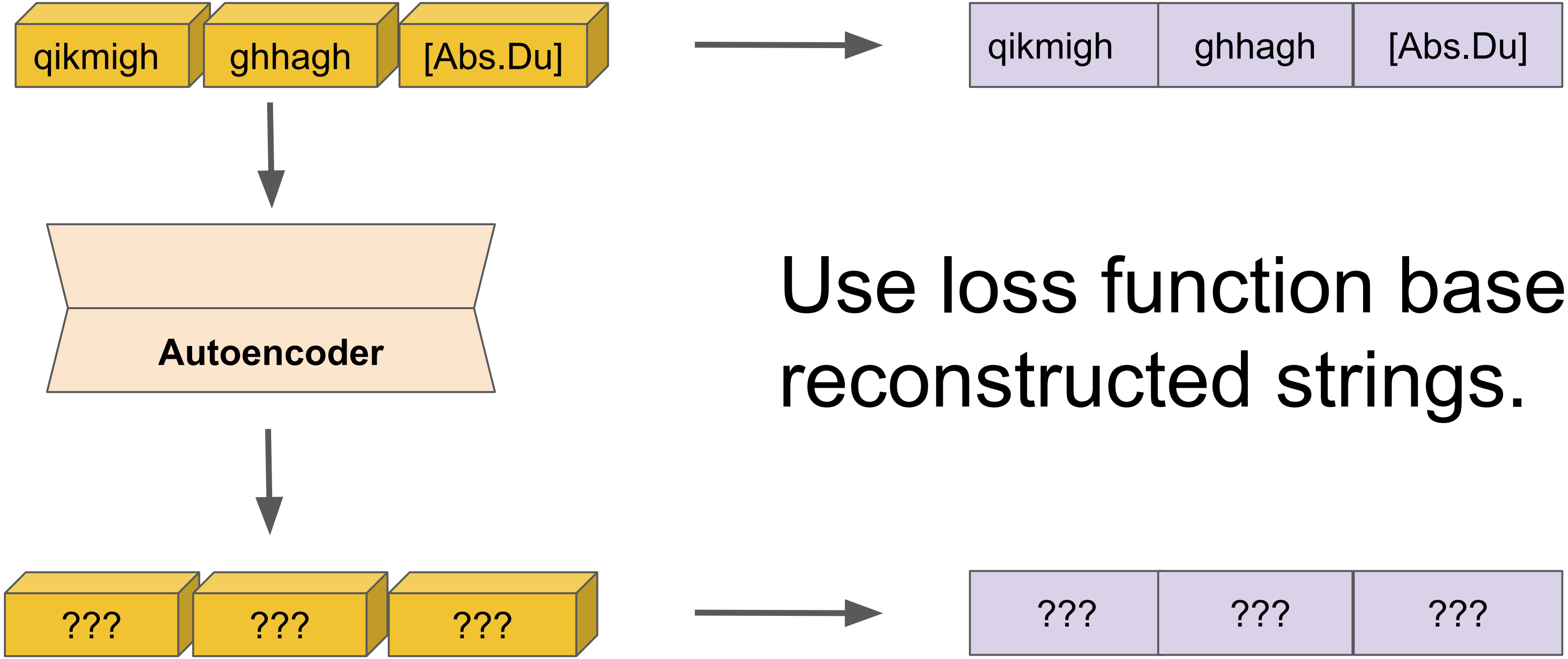
(St. Lawrence Island Yupik)

Problem: Morpheme tensors are sparse



As a result, learning signal is very weak.

Solution: Unbinding Loss



Use loss function based on reconstructed strings.