Morphological Disambiguation for Tocharian

Interdisciplinary Project Report

by Gabriel Breiner supervised by Hannes Fellner, Gabor Recski, Judit Acs and Bernhard Koller

Overview

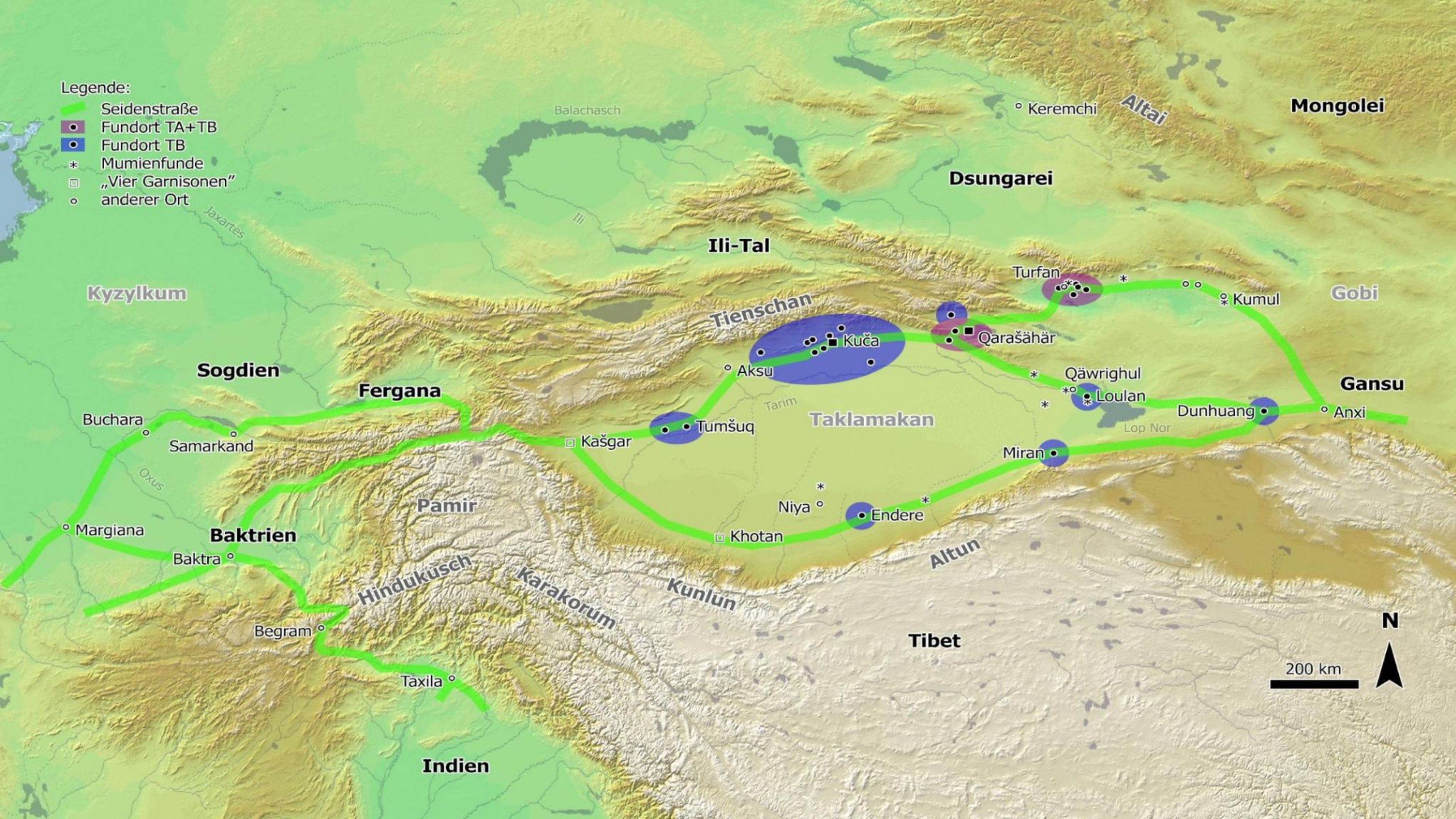
- Introduction
- What is Tocharian?
- The Problem
- Approach
- Evaluation
- Results
- Conclusion

Introduction

- Master’s Student in Data Science
- Interdisciplinary Project
  - “Solve a practical problem in interdisciplinary project work”
  - Problem should not be primarily based in IT or Mathematics.
- Got in contact with Project “Tarim Brahmi”
  - “allow the comprehensive paleographic investigation … ” [of Tocharian]
  - by linking relevant data about manuscripts in one place

The Tocharian Language(s)

- Was discovered around 1900
  - along the routes of the Silk Road
  - in the Tarim Basin
- Actually two very similar languages
  - Tocharian A (TA) - found to the east
  - Tocharian B (TB) - found to the west
- 4th to 10th century CE
- Manuscripts are often translations of Sanskrit
  - Buddhist, scientific, administrative documents.

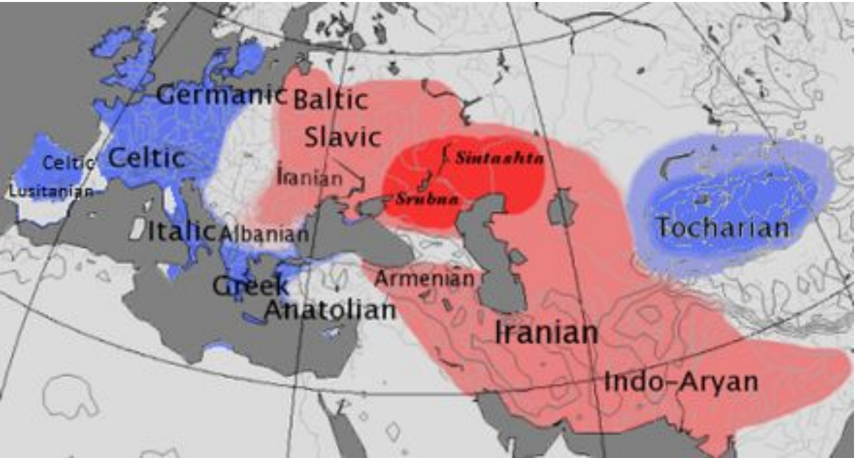


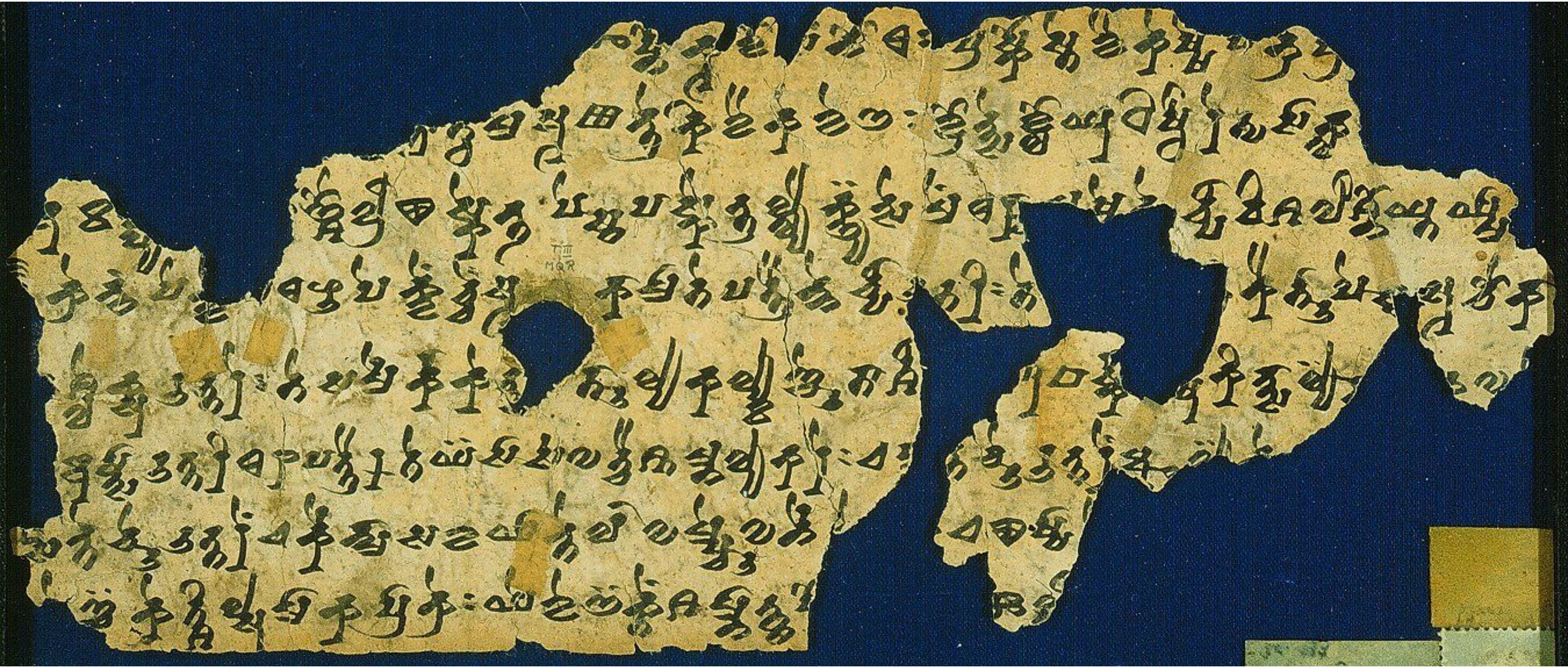
Legende:
Seidenstraße
Fundort TA+TB
Fundort TB
Mumienfunde
„Vier Garnisonen"
anderer Ort
Balachasch
Keremchi
Altai
Mongolei
Dsungarei
Ili
Jaxartes
Ili-Tal
Turfan
Kumul
Gobi
Kyzylkum
Tienschan
Kuča
Qarašähär
Sogdien
Aksu
Qäwrighul
Loulan
Gansu
Fergana
Tarim
Taklamakan
Dunhuang
Anxi
Buchara
Tumšuq
Lop Nor
Samarkand
Kašgar
Miran
Oxus
Pamir
Niya
Endere
Margiana
Baktrien
Khotan
Baktra
Altun
Hindukusch
Karakorum
Kunlun
Begram
Tibet
N
200 km
Taxila
Indien

IT'S HERE
Russia
Finland
Estonia
Latvia
Lithuania
Belarus
Ukraine
Moldova
Romania
Bulgaria
Greece
Black Sea
Georgia
Azerbaijan
Turkey
Kazakhstan
Uzbekistan
Turkmenistan
Kyrgyzstan
Tajikistan
Mongolia
China
North Korea
Sea of Japan
South Korea
Japan
Yellow Sea
East China Sea
Syria
Lebanon
Israel
Jordan
Iraq
Iran
Afghanistan
Pakistan
Nepal
Bhutan
Egypt
Persian Gulf
United Arab Emirates
Saudi Arabia
Oman
Red Sea
Bangladesh
India
Myanmar (Burma)
Laos
Taiwan
Philippine Sea
Sudan
Eritrea
Yemen
Gulf of Aden
Djibouti
Arabian Sea
Bay of Bengal
Thailand
Vietnam
South China Sea
Luzon
Cambodia
Andaman Sea
Gulf of Thailand
Philippines
Panay
Palawan
Negros
Mindanao
Ethiopia
South Sudan
Sri Lanka
Laccadive Sea
Malaysia
Basilan Island
Celebes Sea
Somalia
Uganda
Kenya
Singapore
Rwanda
Burundi
DRC
Tanzania
Java Sea
Indonesia
Banda Sea

Indo-European Languages

TA	TB	english	Latin	Sanskrit	Persian
känt	kante	hundred	centum	śatām	sad
pracar	procer	brother	frāter	bhrātṛ	barâdar

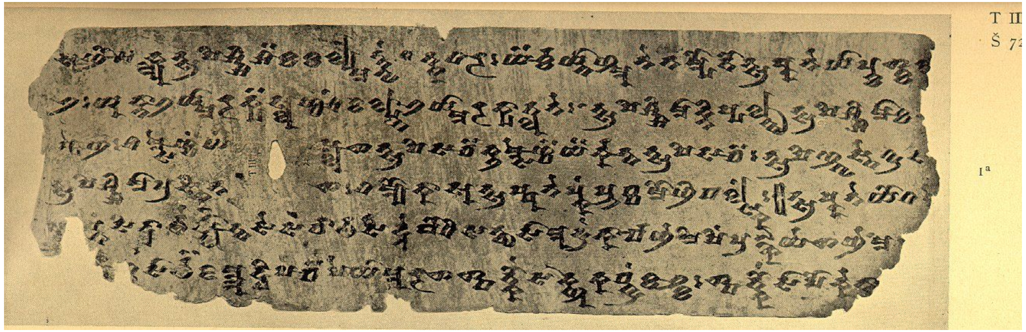




Project “Tarim Brahmi”

- Cooperation between University of Vienna & Austrian Center for Digital Humanities and Cultural Heritage
- “... *link the text witnesses to their digital facsimiles on the character level and to publish this material together with a* **TEI-encoded dictionary** *in an online database*” [7]
- “*... all quantifiable features of all characters, ligatures and words will be extracted and compared using software tools*” [7]
- Currently working on the dictionary

Current Progress



Transcription

... (kā)-
a1 -su ñom klyu tsraṣiśśi śäk kälymentwaṃ sätkatär : yärk ynāñmune nam poto tsraṣṣuneyā p$_{u}$käṣ käl(pnā)-
a2 -l : yuknāl ymāräk yäsluñcäs kälpnāl ymāräk yātlune : 1 tsraṣiśśi māk niṣpalntu tsraṣiśśi māk (śkaṃ) (ṣña)-
a3 -ṣṣeñ : nämseñc yäsluṣ tsraṣisac kumseñ~~c yärk~~ant tsraṣisac : tsraṣiñ waste wrasa(śśi)
a4 tsraṣiśśi mā praski naṣ : tämyo kāsu tsraṣṣune p$_{u}$kaṃ pruccamo ñi pälskaṃ : || tsraṣṣuneyo täm(n)e (ne)-
a5 -ṣ (pra)ṣṭaṃ siddhārthes lānt se sarvārthasiddhe bodhisattu sāmudraṃ kārp ñemiṣiṃ praṅkā yeṣ ñemi – – –
a6 – l·i – sārth jambudvipac pe yāmuräṣ ṣpät koṃsā kñukac wraṃ kälk : ṣpät koṃsā pokenā – – – –
b1 – (kä)lk ṣpät koṃsā lyomaṃ kälk ṣpät koṃsā wälts pältwāyo oplāsyo wraṃ opläṣ oplā kärnm(āṃ) (kälkorä)-
b2 -ṣ päñ kursärwā ār{p/ṣ}lāsyo rarkusāṃ tkanā kälk : tmäṣ rākṣtsāśśi dvipaṃ yeṣ tmäṣ yakṣāśśi – –
b3 baladvipaṃ yeṣ tmäṣ śtwar-wäknā ārṣlās{l/y}o rarkuñcäs iṣanäs kcäk śtwar-wäknā spe(ṣinä)-
b4 -s«†ä» kluṃtsäsyo sopis sāgares lānt lāñci waṣt pāṣäntās śāwes empeles (n)ā(kā)-
b5 -s āsuk kätkoräṣ sāgareṃ läntäṣ cindāmaṇi wmār toriṃ kälpāt poñcäṃ jambudvipis e(kro)-
b6 -rñe wawik ślak śkaṃ || ṣāmnernaṃ || māski kätkāläṃ ktäṅkeñc tsraṣiñ sāmuddrä : traidhātuk saṃs(ār) (tsra)-
... -(ṣṣuneyo)

sätkatär

Lemma: sätkā-
Word class: Finite verb form
Language: TA
Lexeme variants: sätkatär
Lexeme family:
- sätkā- *Root*
  - Kausativum 1 “to spread”
  - Grundverb “to spread out”
    - Present 3
    - Subjunctive 5
      - sätkālune verbal abstract
    - Preterite 1
      - sätko preterite participle

Person: 3
Number: Singular
Tense/Mood: Present
Stem: Present
Stem class: 3
Voice: Middle

Inflectional paradigm

Present

	Singular		Plural		Dual	
	Active	Middle	Active	Middle	Active	Middle
First						
Second						
Third		sätkatär		sätkantär sätkantr-āṃ		

Preterite

	Singular		Plural		Dual	
	Active	Middle	Active	Middle	Active	Middle
First						
Second						
Third	stäk sätkā-m		satkar satkar-āṃ			

Occurrences: 4

sätkatär

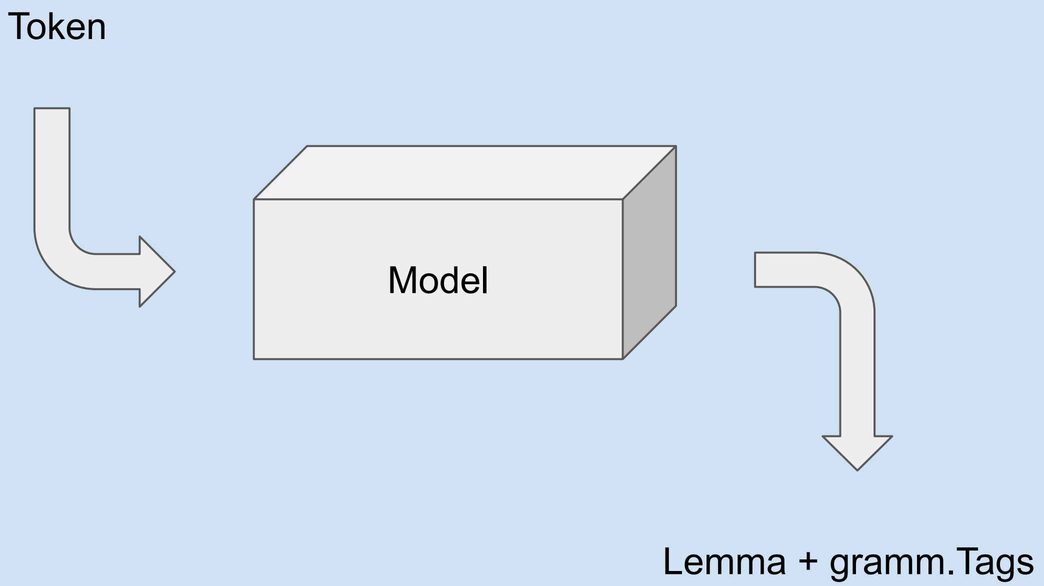
1	A 1 a1	<a1> ñom klyu tsraṣiśśi śäk kälymentwaṃ	**sätkatär**	: yärk ynāñmune nam poto tsraṣṣuneyā
2	A 2 b2	kāsu ñom klyu amoktsāp kälyme kälyme	**sätkatär**	: yärkā yāmäl mäskatär potal (k)r(o)<b3>pal
3	A 37 b5	<b5> -trä : kāsu ñom klyu	**sä(tkatär)**	/// <b6> -nträ kutsmātāñ nämseñc ///
4	A 79 b5	: kāswoneyo yā(mu) nu päl ākntsāśśi	**sätkatär**	tri āpāytwaṃ : 1 (yä)sluntaṃ pe

The Problem

The Problem

- Create a dictionary of all tokens + morphological & grammatical information
  - Lemma
  - Grammatical Tags in multiple categories (case, pos, gender, …)
- Manually annotating each token/type is very time consuming
- Solution: Train predictors for lemma and grammatical information
  - Iterate over the tokens of all documents
- This project was **not:**
  - beating a benchmark
  - comparing different architecture
  - exploring the best parameter setting

Goal




```
“Worten” => “Wort” + [plural,dativ,...]
```

Morphological Disambiguation

- Finding the correct grammatical parse for the morphemes of an inflected word
  - commonly includes lemmatization
- Definition & Approach depends on language
  - english: inflective (barely), but very ambiguous
  - turkish: agglutinative, vast vocabulary, modular

```
dogs → dog(N+pl)
are → be(aux+pres+3+pl)
are → be(pres+2+sg)

saglamlastirmak → saglam/(adj) las(verb+become) tir(verb+caus) mak(noun+nom)
    = “the thing that causes something to become strong”
```

Tocharian Declension

- 10 Cases
  - 4 Primary Cases: Nominative, Genitive, Accusative, Vocative (only TB)
  - 6 Secondary Cases: Perlative, Comitative, Allative, Ablative, Locative, Instrumental (only TA)
- Secondary cases attach to [stem] + [acc.]
- Seems inflective

"King"	
nom.sg	walo
acc.sg	lānt
all.sg	lāntäś

Derivational Morphology

- Productive morphemes can stack together
  - Seems agglutinative

snai	preposition	“without”
snai**tstse**	adjective	“poor”
snaitsts**äññe**	noun	“poverty”
snaitstsäññe**ṣṣe**	adjective	“pitiful”

Ultimately “snaitstsäññeṣṣe” is an adjective.

Mode of Disambiguation

```
disambiguate(x) = {lemma(x), p ∈ pos, c ∈ case, …, v ∈ voice, …}
```

pos	`{noun,adj,verb,uninfl,unkn}`
case	`{nom,gen,acc,voc*,per,com,all,abl,loc,ins*}`
gender	`{m,f,n}`
number	`{sg,pl,du}`
person	`{1,2,3}`
tense	`{prs,sbj,impf,opt,pret,imp}`
voice	`{act,mid}`

*language specific

Data Exploration

```
</pos-a>
<pos-n type="entry" id="E_B_prahati" lemma="prahati">
   <meaning>Indian nightshade, Solanum indicum [a medical ingredient]</meaning>
   <pos val="pos-n"/>
   <stem_gender/>
   <nominative val="infl-nom-i"/>
   <accusative val="infl-acc-i"/>
   <plural val="infl-pl-nta"/>
   <etym>Skt. &lt;i&gt;bṛhatī&lt;/i&gt;, Khot. &lt;i&gt;brrahatti&lt;/i&gt;</etym>
   <word id="F_B_prahati">
      <case val="case-nom"/>
      <case val="case-acc"/>
      <number val="num-sg"/>
      <gender/>
      <phon id="B_prahati">
         <form>prahati</form>
         <accent-rule val="v-p-Aə-a"/>
         <accent-rule val="v-p-Ua-a"/>
         <accent-rule val="v-p-Ui-i"/>
      </phon>
      <phon id="B_brähati">
         <form>brähati</form>
         <accent-rule val="v-p-Uə-ä"/>
         <accent-rule val="v-p-Aə-a"/>
         <accent-rule val="v-p-Ui-i"/>
      </phon>
   </word>
   <word id="F_B_prahatinta">
      <case val="case-nom"/>
      <case val="case-acc"/>
      <number val="num-pl"/>
      <gender/>
      <phon id="B_prahatinta">
         <form>prahatinta</form>
         <accent-rule val="v-p-Aə-a"/>
         <accent-rule val="v-p-Ua-a"/>
         <accent-rule val="v-p-Ui-i"/>
      </phon>
   </word>
</pos-n>
<pos-n type="entry" id="E_B_yirmakka" lemma="yirmakka">
   <meaning>~ treasurer (lit: measurer)</meaning>
   <pos val="pos-n"/>
   <stem_gender/>
   <note>
      <bibl xmlns="http://www.tei-c.org/ns/1.0">
         <title corresp="ref:adams1999"/>
      </bibl>: s.v. <hi xmlns="http://www.tei-c.org/ns/1.0">yirmakka*</hi>.</note>
   <word id="F_B_yirmakka">
      <case val="case-nom"/>
      <number val="num-sg"/>
      <gender/>
      <note>
         <bibl xmlns="http://www.tei-c.org/ns/1.0">
            <title corresp="ref:adams1999"/>
         </bibl>: s.v. <hi xmlns="http://www.tei-c.org/ns/1.0">yirmakka*</hi>.</note>
      <phon id="B_yirmakka">
         <form>yirmakka</form>
```


THT 3206| -ñ- tkālñe - nti dharma ka - ṣa ka ste • - ñe naṃ mä -ce - saṃ khyā - ā -k-ā -
THT 1450.d| - k- r-ñ- śkau sa i - kUce tne sā rñi kwā lype l-e - ntse a nte ke ñā ssa -su wa śāk maiyya -
pelaikne -s- ṣṣ- e -n- rt-i tāñ ersna we śai - kauwwa pyappyaintsa - piś yäknesā ploryaimeṃ -aṃ aṣanaike -e -
tre n-a -
THT 3216| - ñmantse -rä - pāskeṃ : na - ñ- sk- ntā -y- kau rccā -e -
PK AS 16.2|ke ktseñ bram-ñākte mant weña ‖ paṇḍurāṅkāñene ‖ wi-ppewänne kṣattaryi śpālmeṃ : ñākteṃ śāmnaṃts
yärkents= āṣāṃ cai : pelaiknesse wäntre ceṃ saimtsa : centsak saimtsa kantär se śaiṣṣe : 1 ‖ tusāk warñai kṣatri
posa te nauṣ yärkentse aṣāṃ takāre ‖ tumeṃ mant cai ksa alyaik alanmeṃ śāmna eṅkalñene nāki kärsormeṃ ālyauceś
weñare ‖ katarosine ‖ śūkentane treṅkāltsa perne peñyo muskītär : eṅkalñentse ṣarmtsa ywārc yärtoṣ lkāntär
wertsyaṃne : calle ṣ wesäṃ miṣenta lauke tarkam eṅkalñe : warto wṣeṃñai saimtsa wes śayem omte pintwātsa : 1 ‖
tumeṃ cai eṅkalñentse nāki kärsormeṃ koṭanmasa warttone lateṃ • tū no kUce yäknesa ‖ śawaṃñe-kwamane ‖ pw
eṅkalñenta rerīnoṣ śaul śāwṣante airpācce : bram-ñākti ra yayātaṣ warto wṣeṃñai saim yāmoṣ : ompalskoññeś
spelkkessoñc kärsoṣ nāki kleśanmaṃts : śle-maiyyā ywārc ersante abhijñenta piś ṣäp no : 1 ‖ se tane teri ste
ente pañākti śaiṣṣene mā tsämoṣ tākaṃ • twak māka krātayuk preścīyaṃne kUse cai orotstse cämpamñecci bodhisatvi
tākaṃ cai ot tämpak-yäknesa rṣāki mäskentär cenaṃts omte airpācce śaul śaitsi skeyessontāṃts cämpamñe ṣai
tarāltse śaiṣṣeṃ tāñtsi ālyine amalākämpa tasemane po wāntarwa lkātsi raddhisa yatsi ñākcīyai klautsaisa
klyaussi kātkor ekamätte karsatsi • eṃṣke nemce ylai-ñākte bram-ñāktesa warñai ñākteṃts yarkesa yamaṣṣālyi tākaṃ
tu yäknesa aurtsana aiśamñenta pärkān-me • eṃṣke tot naivasaṃjñānāsaṃjñāyataṃ tāñtsi
THT 1389.n| - ca ke lk- cakene pa -te ye kle - tuwak -
Dd 6.1| - krent ṣpane lka -
PK Bois A27| - kunacaṃttre ākṣa • -
THT 147.5| - repacyeṃ wīkastsi -t-ṣc rämttär no räse -
THT 1439.f| - cchati gacch- mā - seṃ -
PK Bois A21| - parra tārka se -
THT 2192| - l-a - ṣpā yi -
THT 587.b| - -ntsa snai kärstau snai ṣotri - wätkaltsa sältaññe ramtä - aikne cpī aksäṣṣäṃ - ksa cpi nervvanne
ṣāñ o -
THT 1857| -r- tw- ntse - ly- s-
THT 3891.a| - pā - sna - : dvā -
THT 2386.k| - nt-r-e -e -i -ī ke -e l-ai -
THT 274|kUce toṃ wñāwa āyorntā maittreyeṃśca śāmtsisā papāṣṣorñe tonts pontants ścmoñña nesāṃ śaiṣṣentse kentsa
pärnna mā stämoṃ ṣāli stāna onolmī papaṣṣorññesa pärna kUc= āyornta yāmorntasa 30-5 kUse toṃn yāmāṃ wāntārwa
eynāke rā ksa śamnā ñāktents ñākteṃ maitreyeṃ lkāṃ su āksau ñāś centsā tonmeṃ kā ks= ālāṃ pälko tākaṃ
aknātsaññesā ñākti lāñco wāntreśśī mā cai lkān-ne klyomñesā 30-6 papāṣṣorñe tākaṃn kā ṣeme ra ksa cok tāṅksā
twāṣṣāṃ ñi śariputra - maitreyeṃ kUse ysāṣṣāna pyappyain rā kātaṃ ñā -ā - kauc krUi -ī - śāmnantso cet= sū
lkātsi 30-7 mā yātaṃ sū yänmātsī meyyāsa epretñesā ś-t- māktewñe lantuññesā mā r= āmokānt= āklorsa ma etreUññai
meyyāsā krent yāmorsa āyorsa yātaṃ śāmtsi maittreyeṣc 30-8 papāṣṣorññeṣṣeṃ krentāṃ yayātāṣāṃ koklentsā ai - sā
āyorṣṣe aiśāmñeṣṣe pantaintsā śāmnā śānmeṃ maittreyeṃśc po śaiṣṣentse - ce saṃtk= ewkān-m= onwāññe lāṃ saṃsārṣṣe
pelemeṃ 30-9 ślek te yatka pūdñākte śāriputri prāśānne śkas yäknesa maiwa keṃ tary yāltse po śaiṣṣenne kodyānmā
sumerntā naittāre po waskāte -
THT 3506| - ma kā śy- n- m-a mai -
THT 1334.m| - meṃ postāṃ - ṣṣ- ā ya -
THT 2819.v| -iś -
PK Bois B104|ñi yaitkorsa maṃt pyām - yaṃ cewāmpa klaina - oksaiṃ trai • kerccappaṃ -
THT 3830.c| - ma -
THT 2795.y| - wa wā -
THT 3291.g| - śtā - k- ne rā -
THT 2379.u| -p- l-e twe - stña k-a - ne kekmoṣā śaiṣṣe - wpśa ntai : a l-e - lyelkorme tāñwā - l- ṣm- ly- ka
-ts- śc tāñe ṣārmntsa - ceṃnts no tā kUsa ñ- me śpā ke ṣṣe kā o -
THT 3916.c| - tu vi ru - y- hā -
THT 2762.c| -s- wi - ś- ści -
THT 3204| - tu lmi - l-r pu -
IOL Toch 257| - r- yamaṣṣa - y-ntär n- w- ln- -meṃ ṣai kUse ṣpā pone ñorameṃ ṣai toy - -na stānampa tasemane •
se ṣuktante pañākte -e kärsanalle 10-7 ‖ tvamoghe - • dvīpa • preṅke • trāṇaṃ • waste • kṣat- -ṣūwa ra
yamaṣṣeñcana pelaiknenta aksaṣṣeñca - -ne leleku ṣai cwī no kUśalamūlānta a - laklesa ñeñusku ṣai ñakti aiśai
yamaṣa - o -
THT 2377.t| - pä ṣṣā -l- ś- ne - ñāśā lkā - lna ra ki t kättwoy yontwe māñce -n-e - ñ-e ne me -t- ñā kte wa -y-

Labeled Data

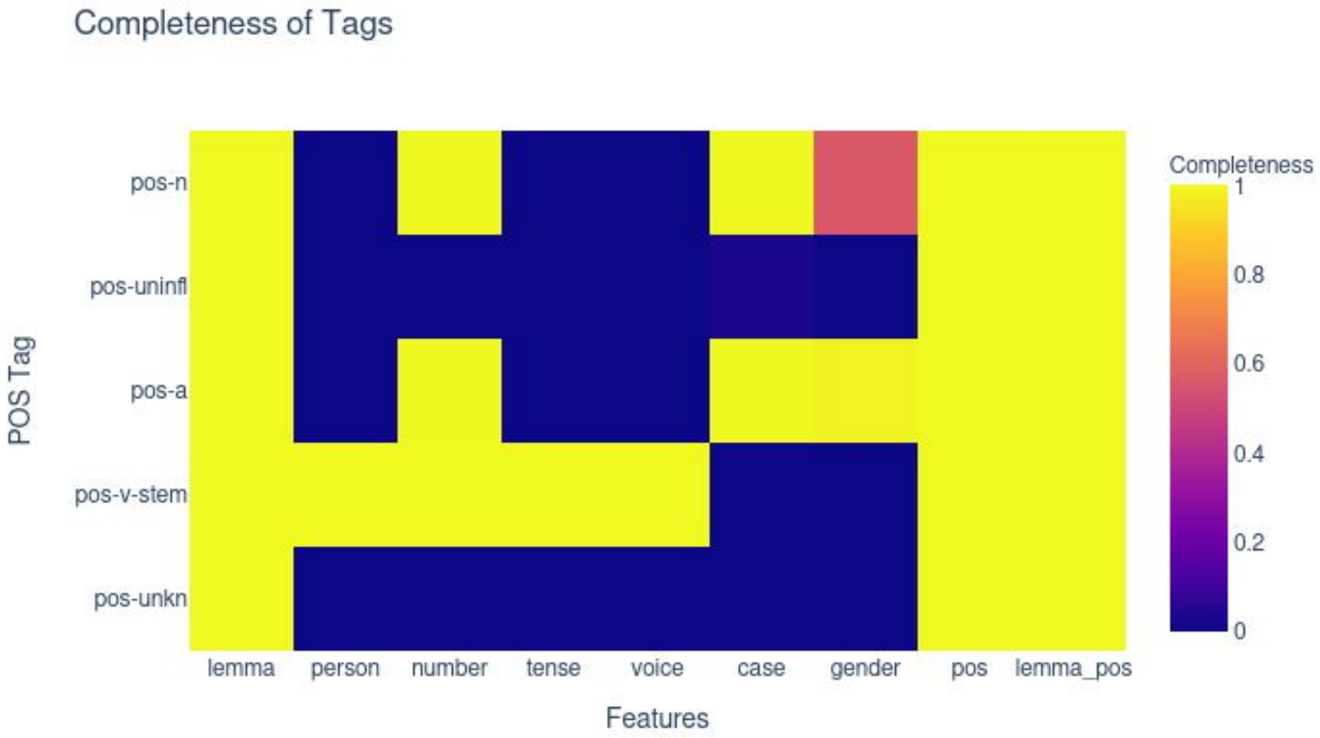
	TB	TA
# types	11268	3955
# tokens	29298	13220
# documents	7057	1635
mean doc length	20.4 tokens	33.24 tokens

41.9%

of samples are annotated (TB)

44.4% for TA

How complete are the annotations (rows)?



Ambiguity: How many words have multiple parses?

- Influences choice of model
  - low ambiguity: character-level
  - high ambiguity: word-level & context
- answer: **30.04%** (3386)
- What kind of ambiguities?

ṣamāññeṣṣe → ṣamāne + **nom** + m + adj
ṣamāññeṣṣe → ṣamāne + **acc** + m + adj
ṣamāññeṣṣe → ṣamāne + **[nom-acc]** + m + adj

Approach

CoNLL-SIGMORPHON [4]

- Challenge in 2018
- Task 1: Inflection
  - given: lemma + tags
  - generate: inflected form
  - with low (10^2), medium (10^3) and high (10^4) numbers of training samples
- On a large number of different languages
  - inflective and agglutinative Ls represented

- Adequate reference for this project
  - *this is the inverse problem*
  - *many languages in the challenge (inflective and agglutinative)*
  - *according to this we have a medium to high number of training samples*

A string generation task

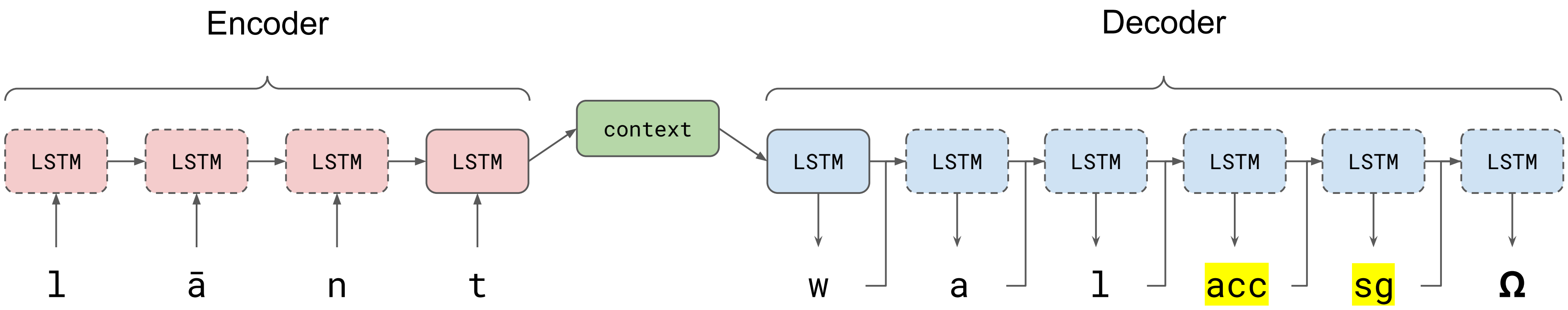
- Common approach to SIGMORPHON
  - and well performing (eg. BME-HAS Acs [5])
- Include the grammatical tags in the target string
  - treat tags as characters
- Train (neural) predictor to generate lemma + tags
  - character by character
  - solves both lemmatization and grammatical tagging

`l+ā+n+t+ä+ś → w+a+l+all+sg`

“King”	
lemma	wal
nom.sg	walo
acc.sg	lānt
all.sg	lāntäś

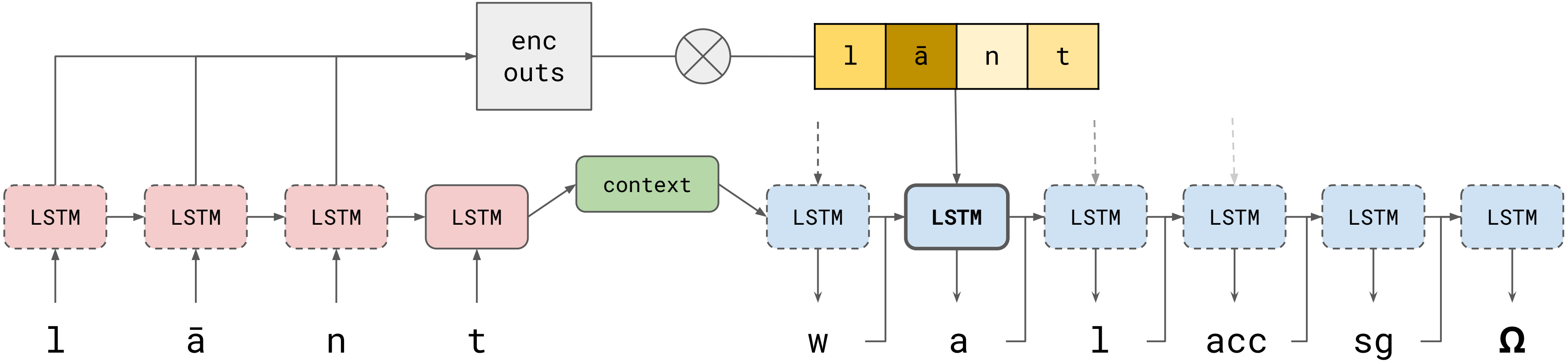
Sequence-to-Sequence Model

- predicts output sequence given input sequence
  - uses some form of RNN
  - enables different input & output lengths
- in this case: encoder-decoder seq2seq model [2]
- common for translation tasks
  - language generation tasks in general
  - often on word-level
  - in our case: character level



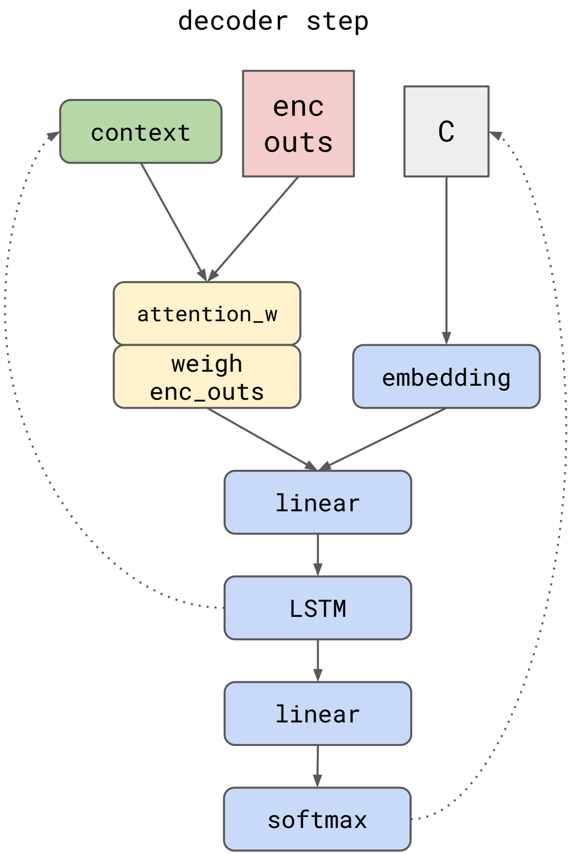
Attention

- Extension of encoder-decoder architecture
  - decoder has access to all encoder outputs
  - weighs encoder outputs (=drawing attention to certain elements)
- Advantages
  - counters long sequence bottleneck
  - provides feedback of the model
  - increase performance



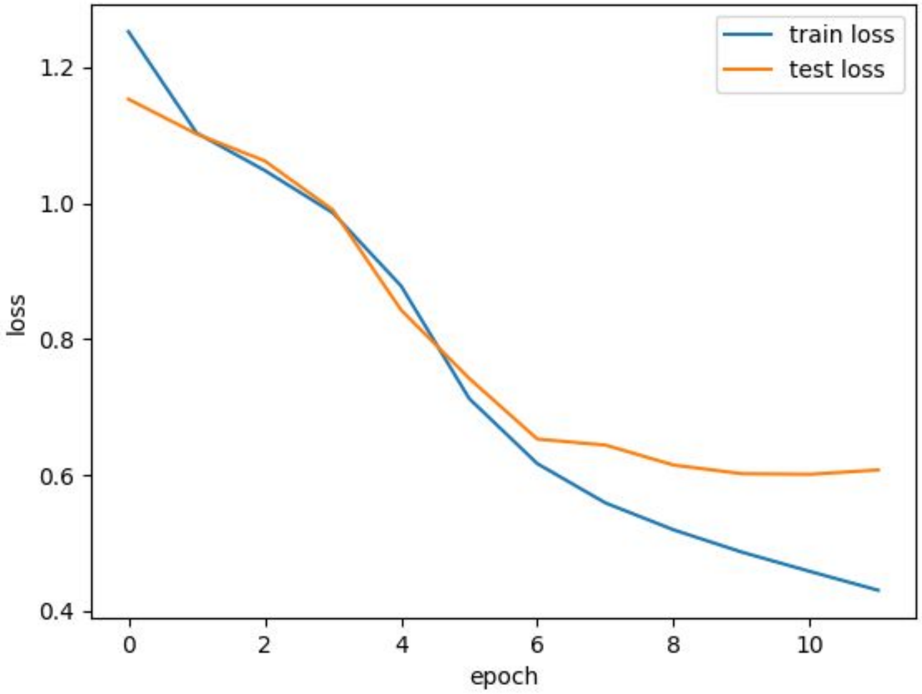
Bahdanau Attention [1]

- "Additive attention"
  - as opposed to multiplicative attention (Luong [3])
- computed in decoder step
- Steps
  - calculate alignment from enc_outs and context
  - softmax alignment (= attention_weights)
  - enc_outs * attention_weights
  - concat attended & embedded



Training Process

loss	cross entropy
optimizer	adam
learning rate	1e-4
batch_size	1
embedding_size	50
context_size	100



Evaluation & Results

Evaluation

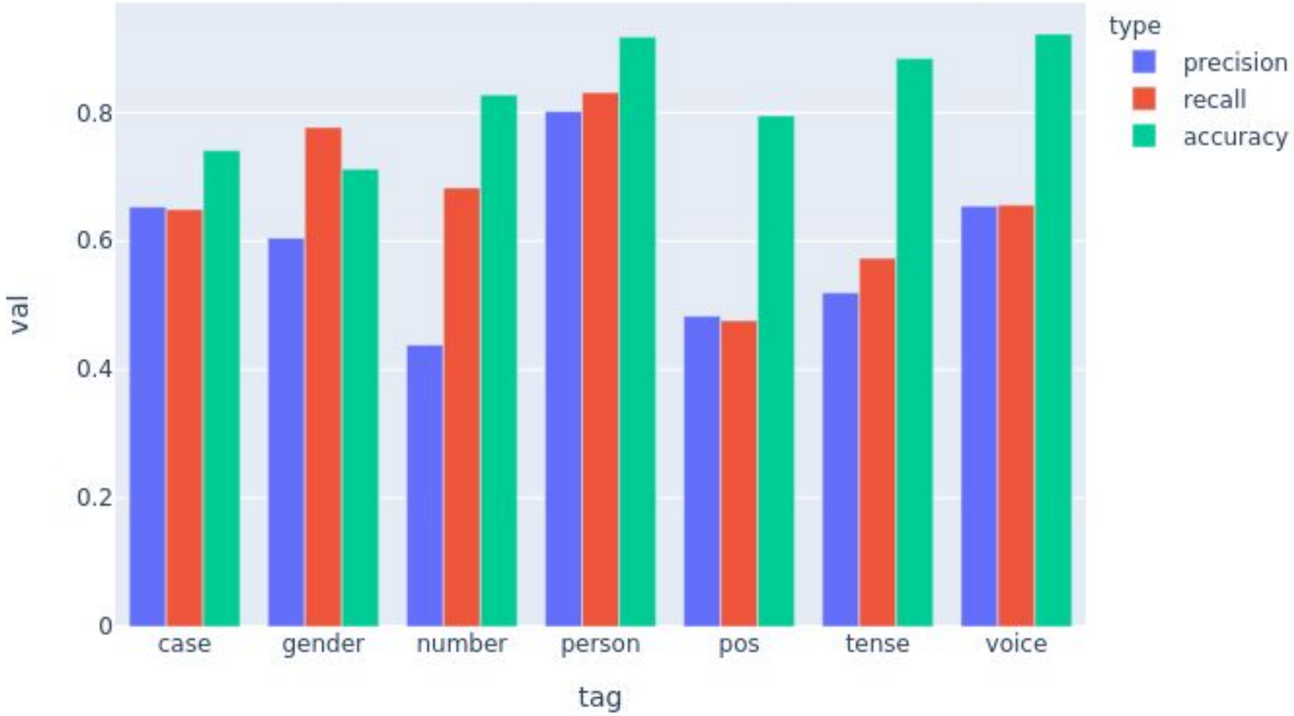
- Split the per-sample evaluation in to two parts
  - Lemma: Levenshtein Distance
  - Gramm.Classes: Precision, Recall, Accuracy
- Levenshtein Distance
  - count of operations to transform given string into target string
  - operations: add, delete, substitute (characters)
- Precision, Recall, Accuracy
  - calculated per gram.category
  - tuning for recall would be ideal

1.21 / 1.87

Levenshtein Distance

test / val

Classification Metrics (TB val set)



average	
precision	57.2%
recall	60.8%
accuracy	82.2%

Conclusion

- Scores are nowhere near perfect
    - they do not need to be, since they will be checked
    - team was actually positively surprised with the results
- Benefits of a black box model
    - Team “Tarim Brahmi” did not need to know much about ML
    - I did not have to learn Tocharian
    - = less interdisciplinary work?
- Neural networks are hard to debug
    - have a proper experimental setup
    - do not get carried away
- Interdisciplinary Work
    - Communication is very important (vocabulary!)
    - Ideally: Linguists in control up until preprocessing step
    - How much do they need to know? How much do I need to know?
    - I was possibly very lucky considering the data situation

References

[1] Neural Machine Translation by Jointly Learning to Align and Translate Bahdanau et al. 2015

[2] Sequence to sequence learning with neural networks Sutskever et al. 2014

[3] Effective Approaches to Attention-based Neural Machine Translation Luong et al. 2015

[4] The CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection Cotterell et al. 2018

[5] BME-HAS System for CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection Ács 2018

[6] CeTOM Website https://www.univie.ac.at/tocharian/?home (25.05.2021)

[7] Project Tarim Brahmi https://www.oeaw.ac.at/acdh/projects/tarim-brahmi/ (25.05.2021)