



Building Resources: Language Comparison and Analysis

Miriam Butt

University of Konstanz

SIGTYP 2020

DFG Deutsche
Forschungsgemeinschaft

DAAD Deutscher Akademischer Austauschdienst
German Academic Exchange Service



This talk looks at the use of **comparative linguistic insights** in building up computational resources for South Asian languages via two case studies: Urdu and Tamil.

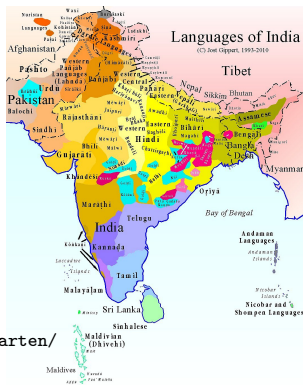
Structure:

- 1 Some Background on South Asian languages and NLP
- 2 Experience with Building Urdu Resources
- 3 The ParGram Effort: Multilingual NLP
- 4 Building Tamil Resources
- 5 Conclusions



Background: South Asian Languages

- My research focus: South Asian languages, primarily Urdu/Hindi.
- The major South Asian languages are spoken by millions all over the world.
- Some major languages: Bangla, Gujarati, Kashmiri, Kannada, Malayalam, Marathi, Nepali, Pashto, Punjabi, Sinhala, Sindhi, Siraiki, Telugu, Urdu/Hindi.
- Hundreds more “minor” ones.
- All mostly studied by only a handful of linguists (if any).

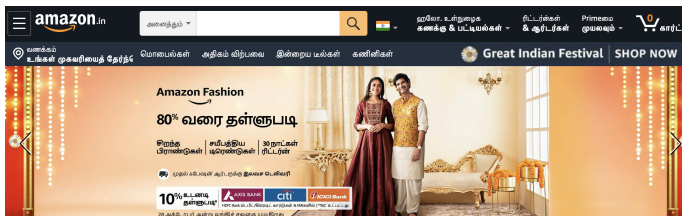


<http://titus.fkidg1.uni-frankfurt.de/didact/karten/indi/indicm.htm>



Background: NLP and South Asian Languages

- Due to British colonialism, English is a major language in South Asia.
- So until recently, little incentive for investment in NLP from the perspective of companies.



- Amazon introduced a Hindi interface in 2018 – before that available in English
- Flipkart (Walmart) started Hindi in 2019, Tamil, Telugu and Kannada in 2020.
- Amazon just added Kannada, Malayalam, Tamil and Telugu in **September 2020** (in time for Divali).



Background: NLP and South Asian Languages

- Hindi is an **Indo-Aryan** language.
- Kannada, Malayalam, Tamil and Telugu are all **Dravidian**.
- Most South Asian languages share broad structural characteristics (word order, verbal predication, case system, mostly agglutinative morphology).
- But Indo-Aryan and Dravidian languages are also quite different.
- Why this focus on Dravidian?
Perhaps because the major e-commerce companies in India are in Dravidian speaking areas?



Amazon, Hyderabad: Telugu



Flipkart, Bangalore: Kannada

Background: NLP and South Asian Languages



South Asian languages are severely under-resourced in terms of NLP.

- Few and usually small annotated corpora.
- Few and usually small lexical resources.
- Few robust NLP tools/software.
- Incipient efforts at standardization across projects/sites/languages (e.g., POS Tags, Dependency labels).

(Will the recent industry involvement result in a significant push forward?)

This talk: lessons learned while building NLP resources for Urdu and Tamil.



Center for Language Engineering (CLE)
at University of Engineering and Technology
(UET), Pakistan: **Urdu**



University of Moratuwa,
Sri Lanka: **Tamil**

Starting Point

- The cooperation with Pakistan began in the early 2000s.
- The cooperation with Sri Lanka in 2018.
- State-of-the-art in NLP has changed vastly in this time, but the process of building NLP resources has been essentially the same:
 - 1 Initial experimentation with ML methods, with poor results.
 - 2 Realization that for many tasks, need annotated and large corpora for learning.
 - 3 Realization that computer scientists lack the necessary linguistic knowledge to build up high quality linguistically annotated resources.
 - 4 Search for linguists to partner with.
 - 5 Realizations in South Asia:
 - a. There are very few linguists with knowledge about the language.
 - b. If there are any, they do not seem to be able to organize their knowledge in a way that is useful for NLP.
 - c. This also tends to be true for any grammars or dictionaries available (in South Asia, many of these were written in the 1800s).



- If one is able to partner up with a linguist, a period of language analysis follows.
- Some typical major issues:
 - What are the main syntactic categories and structures of the language?
 - How does the morphology work?
 - How to deal with phonological/orthographic variation?
 - What lexical resources are needed?
 - How does verbal predication work (complex predicates, case marking, agreement, etc.)?



Resource Building

- Once these questions have been resolved on even a basic level, useful resources can be built.
- CLE has been very good at making their resources available (some of them for a fee).



CLE
Unlocking Information for
Human Development

CENTER FOR LANGUAGE ENGINEERING

English

Home

About Us

Research

Teaching

Downloads

Resources

CLE Store

Web Services

Careers

Site Map

Contact Us

[Text Corpora] [Image Corpora] [Speech Corpora] [Lexical Resources] [NLP Applications]
[How to Order]

CLE is making these linguistic resources available without cost for supporting academic, non-commercial research. The processing fees being charged will be used to maintain these resources. You are requested to contact CLE directly for any discounts (applicable only for selective public organizations in Pakistan) or for commercial licensing options.

Text Corpora

CLE Urdu Digest Corpus 100K	[Pakistan] [International]
CLE Urdu Digest Corpus 500K	[Pakistan] [International]
CLE Urdu Digest Corpus 1M	[Pakistan] [International]
CLE Urdu Text Corpus 14 Point Size	[Pakistan] [International]
CLE Urdu Text Corpus 16 Point Size	[Pakistan] [International]
CLE Urdu Text Corpus 18 Point Size	[Pakistan] [International]
CLE Urdu Text Corpus 20 Point Size	[Pakistan] [International]
CLE Urdu Text Corpus 22 Point Size	[Pakistan] [International]
CLE Urdu Text Corpus 24 Point Size	[Pakistan] [International]
CLE Urdu Text Corpus 28 Point Size	[Pakistan] [International]
CLE Urdu Text Corpus 32 Point Size	[Pakistan] [International]
CLE Urdu Text Corpus 36 Point Size	[Pakistan] [International]
CLE Urdu Text Corpus 40 Point Size	[Pakistan] [International]
CLE Urdu Digest POS Tagged Corpus 100K	[Pakistan] [International]
CLE Urdu Digest 10B Tagged Corpus	[Pakistan] [International]

CENTER FOR LANGUAGE ENGINEERING

English

CLE NLP Webservices are now available for FREE

- Urdu Text-To-Speech Service (V1.1) - Updated on September 2020
- Urdu Speech-To-Text Service (V1.0) - Updated on July 2020
- Roman to Urdu Script Service
- Camera Captured Address Recognition and Structuring Service
- Urdu Nastalique OCR for Low-Resolution Images Service
- Speech-To-Speech Translation Service

[View All]

News


18 September 2020: Urdu Text to Speech (TTS) V1.1 has been launched. It enables the computer to read out Urdu content in human-sounding voice available in digital forms such as emails, websites and documents. In this version, prosody model has been implemented to improve naturalness of the voice.

7 September 2020: Al-Buruz (Brahui language journal), Journal of Managerial Sciences (journal published in English language) and Noore-Maarfat (multilingual journal) are now available in Tehqeegat: A Research Indexing System. Tehqeegat is hosting research journals in a wide array of Pakistani languages. Indexed Journals List.


28 August 2020: CLE has developed Roman to Urdu script conversion service. This service helps in the conversion of text written in Roman-Urdu into equivalent Urdu script.

Quick Links

- Tehqeegat
- Urdu SAPI Voice
- CLT20
- Urdu OCR - Desktop Version
- Online Punjabi Dictionary
- Online Urdu Dictionary
- Online Torwali Dictionary
- Sindhi English Dictionary



Be the first of your friends to like this



Center for Language Engineering

<http://www.cle.org.pk/index.htm>



Machine Learning

- Once the resources have been built, ML techniques can be applied to level up/extend the existing resources and to build new NLP applications (e.g., Ehsan and Butt 2020).

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5202–5207

Marseille, 11–16 May 2020

© European Language Resources Association (ELRA), licensed under CC-BY-NC

Dependency Parsing for Urdu: Resources, Conversions and Learning

Toqeer Ehsan¹, Miriam Butt²

¹Department of Computer Science, University of Gujrat, Pakistan

²Department of Linguistics, University of Konstanz, Germany

¹toqeer.ehsan@uog.edu.pk, ²miriam.butt@uni-konstanz.de

Abstract

This paper adds to the available resources for the under-resourced language Urdu by converting different types of existing treebanks for Urdu into a common format that is based on Universal Dependencies. We present comparative results for training two dependency parsers, the MaltParser and a transition-based BiLSTM parser on this new resource. The BiLSTM parser incorporates word embeddings which improve the parsing results significantly. The BiLSTM parser outperforms the MaltParser with a UAS of 89.6 and an LAS of 84.2 with respect to our standardized treebank resource.

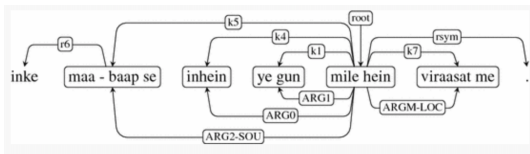
- However, a meaningful evaluation continues to be difficult if existing resources such as high-quality treebanks are still on a (relatively) small scale.



Resource Building

Note: essentially the same development cycle led to the establishment of the **Urdu-Hindi Treebank** (Bhat et al. 2017) and attendant resources.

http://ltrc.iiit.ac.in/hutb_release/

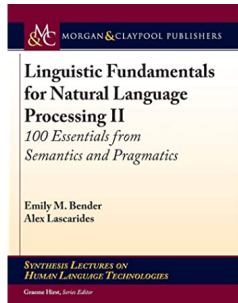
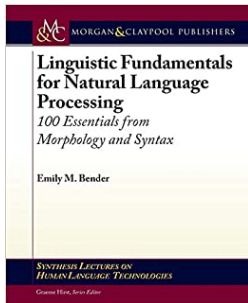


(I was involved as a consultant).



Comparative Linguistic Knowledge

- Linguists acquire a lot of knowledge about language structure in general and how languages differ (universals, comparative typology).
- This helps with tackling underdescribed and underanalyzed languages.
- However, it is difficult to translate linguistic general knowledge into an NLP friendly format.
- One exception: Emily Bender's books





POS Tagging

- Assigning a Part-of-Speech (POS) to an item is one of the very basic tasks in linguistic analysis.

There are ten parts of speech and they are all troublesome.

Mark Twain
The awful German Language

- However, it is also beset with difficulties.
 - How should “in-between” categories be analyzed, e.g., deverbial adjectival participles.
 - How granular should an analysis be?
 - How to deal with foreign/borrowed items?
 - ...

The definitions [of the parts of speech] are very far from having attained the degree of exactitude found in Euclidean Geometry.

Otto Jespersen
The Philosophy of Grammar



POS Tagging

- POS is determined on the basis of **distributional** characteristics (where in a sentence/constituent, what type of morphology, together with which words).
- As such, an ideal task for ML.
- But linguistic analysis has so far had to provide an initial basis for learning.
- Generally cycles in development:
 - 1 Proposal of an initial tagset based on
 - linguistic insight
 - established tagsets (e.g., Penn Treebank Tag-Set)
 - existing generalized guidelines/recommendations (e.g., UD Universal POS tags¹)
 - 2 Manual and/or semi-automatic tagging of a chosen corpus
 - 3 ML based on the tagged corpus
 - 4 Revision of tagset based on experience with manual tagging and ML.

¹<https://universaldependencies.org/u/pos/>



Urdu POS Tagging

- Schmidt's (1999) descriptive grammar of Urdu describes 10 POS tags.
- The first computational tagset for Urdu (following EAGLES guidelines) proposed 350 tags (Hardie 2003; EMILLE corpus).
- Sajjad and Schmid (2009) instead propose 42 tags, a number which is computationally more reasonable to handle.

- The Universal Dependency (UD) guidelines provide 17 high-level tags.²
- An effort to provide unifying guidelines for Indian languages (IL-POS) has 11 major tags and 18 attributes (Sankaran et al. 2008).

²<https://universaldependencies.org/u/pos/>



Urdu POS Tagging

- As part of a German-Pakistan DAAD cooperation we were able to invest time towards building resources for Urdu NLP.
- CLE put together a balanced high quality corpus – the **Urdu Digest** corpus.

دنیا کا ہر فرد کامیابی کا آرزو مند ہے۔ ناکامی سے سب گھبرائے ہیں۔ عزت، دولت، راحت اور عاقبت کی ننگی کے سخی شیدائی ہیں۔ لیکن اصل کامیابی کیا چیز ہے؟ اور حقیقی عزت و راحت کس طرح نصیب ہوتی ہے؟ اس سب سے بہت کم لوگ واقف ہیں۔ اگر آپ حقیقی کامیابی کے گر جانا چاہتے ہیں تو ڈاکٹر زاہد فیضی عمار کی تازہ تصنیف 'ایزند گوارا' پڑھیے۔ ۱۱۳ صفحوں کی اس کتاب کا ایک ایک حرف بھیرت کے درپے کھولنے پر مامور ہے۔

راقم نے اس کتاب کا مطالعہ کیا تو لفظ و معنی کی کھشاکھ دیکھ کر مسحور ہو گیا، جس چیز نے خاص طور پر متاثر کیا وہ ڈاکٹر صاحب کا نظم قرآن ہے۔ بظاہر ڈاکٹر صاحب پنجاب یونیورسٹی کے معلم ادبیات ہیں لیکن درحقیقت وہ ایک داعی، ایک عارف، ایک محقق، ایک مدد، ایک مہر، ایک آئینہ نظر اخلاق اور علم و قلعاس کے فراتر ہیں۔ ٹی وی پر ان کی تقریریں جسے ذوق و شوق سے سنی جاتی ہیں۔ ان کی باتیں عید کی بیویاں ہیں۔ بے اختیار دل میں اتنی جلی جاتی ہیں۔ سب سے زیادہ گرائیڈ غوثی یہ ہے کہ ان کی گفتگو قرآن کریم کی برعمل آیات اور ارشادات رسالت آتب سے یوں بھگکتی ہے جیسے

ع بر تو سے آکتاب کے ذسے میں جان ہے!

- We wanted to add high quality annotations to it.
- As a first step we took on POS.
 - Informed by previous efforts.
 - And efforts at unifying standards.
 - But mainly through our extensive collaborative LFG-based grammar writing experience in the early parts of our cooperation (cf. ParGram).



POS Tagging — Summary

- ML approaches have yielded high performing POS-taggers.
- POS-tagging is not an end unto itself.
- It is a first analysis step for down stream applications.
- The tagset must therefore be well designed.
- This includes being:
 - computationally tractable
 - linguistically well motivated
- One can also think about building **hybrid systems**.
 - Write rules for those parts which parts are easily identifiable and few in number (closed class, e.g. negation, focus clitics, modals).
 - Combine these with a language model.
 - Thus saving time and computational resources
→ by using linguistic knowledge about language structure.



Text-to-Speech and Prosody

- Having built basic NLP resources, the DAAD German-Pakistan cooperation was able to move on to more challenging tasks.
- CLE has developed a Text-to-Speech System (TTS) for Urdu.
- TTS is particularly critical for areas like South Asia, which have wide-spread illiteracy.
- The production of natural sounding speech requires an integration of **prosody**.
- Prosody is challenging:
 - Still “unsolved” even in well-studied languages like English.
 - Need to understand how phonetic cues translate into (which) linguistic categories.
 - Identify relevant information in the speech signal.
 - Develop viable annotation schemata.

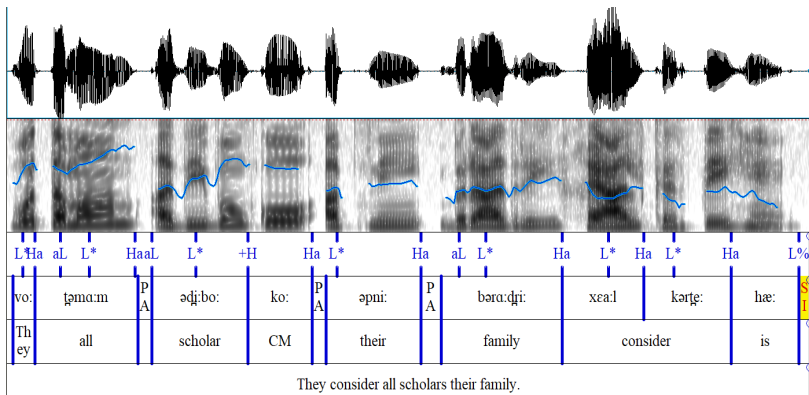


Description of Urdu Speech Corpus

Size of corpus	1303 sentences (3 hours of speech)
Text source	Books, magazines and newspapers (balanced)
Sentence selection	Automatic sentence selection from text corpus using greedy algorithm
Types of sentences	951 declaratives, 183 interrogatives, 151 imperatives, and 18 exclamatory sentences
Length of sentences	5 -15 words
Type of corpus	Read speech
Number of speakers	3 professional speakers (2 males and 1 female)
Levels of Annotation	Word, Stress, Break index (BI) and Intonation



Urdu Prosodic Phrasing (Urooj et al. 2019)

[Play Sound](#)


- Urdu intonation is mainly a pattern of Low-High (L-H) on prosodic phrases.
- The last phrase in declaratives is always Low.



Our Approach

- Manual prosodic annotation is difficult and time consuming.
- CLE uses semi-automatic methodology:
 - Implementation and application of Hussain's (2005) stress algorithm (helps determine prosodic phrasing).
 - Semi-automatic annotation of intonation and prosodic phrases.
 - Extension via ML once enough initial data has been annotated.
- Learned language model for prosodic patterns feeds into TTS model.
- The inclusion of prosodic knowledge was found to indeed improve the TTS.



Prosodic Annotation

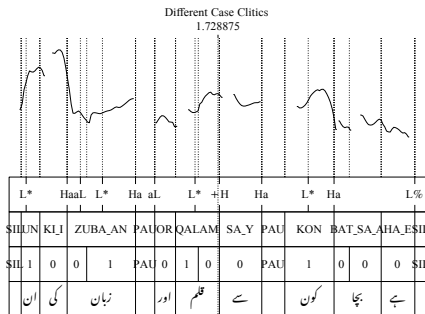
- The annotation again proceeds cyclically.
 - 1 Initial annotation scheme developed.
 - 2 Revisions due to problems/inconsistencies found during manual annotation.
 - 3 Revision due to problems/inconsistencies detected during automatic annotation.
 - 4 Reannotation.

- Some of the non-conforming patterns found by CLE made no sense to them.
- As such they were not sure how to proceed/revise their analyses.

- I have very little expertise in prosody.
- But given my general **comparative linguistic** knowledge I could identify generalizable reasons for seemingly puzzling patterns.
- These insights in turn can feed into the annotations and thereby the language model.

Linguistic Insight

- Generalizable cases involved: case clitics, focus clitics, negation, question words, compounding and (identification of) derivational morphology.

[Play Sound](#)


- (1) [ʊn=ki [zuban or kalam]]=se kon bac-a
 they.Obl.Pl=Gen.F.Sg tongue.F.Sg and pen.F.Sg=Inst who.Nom save-Perf.M.Sg
 hɛ?
 be.Pres.3.Sg
 'Who was able to be escape his/her tongue and pen?'



Prosody Summary

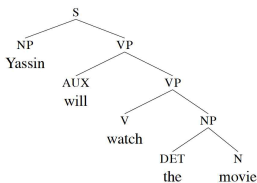
- TTS is an interesting area in which both rule-based algorithms and ML play a role.
- A large part of producing natural sounding speech rests on integrating prosody (intonation).
- Prosody is under-researched, but generalizations or explanations for patterns are available via high-level comparative linguistic knowledge.
- Application of this knowledge allows for the building of better systems.

Linguistic Generalizations — LFG

- Linguistic insights do not occur in a vacuum.
- They are formed and tested as part of expectations generated from a particular **theory of language**.

(1) a. Yassin will watch the movie.

b. e-structure



c. f-structure

PRED	'watch<SUBJ,OBJ>'
SUBJ	[PRED 'Yassin'
OBJ	[PRED 'movie'
TENSE	future

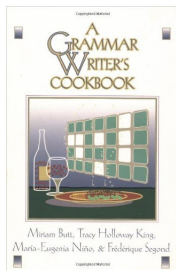
- I work within a theory of grammar that:
 - is strong in terms of integrating **typological** information
 - relies heavily on the **inductive method** — looks at data and formulates hypotheses/expectations on the basis of the data
 - is geared towards enabling **computational** (and psychological) modelling
 - includes both constituent and dependency information (cleanly separated).
- The theory is **Lexical-Functional Grammar (LFG)**.⁴

⁴ <https://ling.sprachwiss.uni-konstanz.de/pages/home/lfg/index.html>



ParGram

- LFG is the basis for the computational **ParGram** (Parallel Grammar) effort.
- Began in 1996 with 3 sites:
 - PARC (English)
 - Xerox Grenoble (French)
 - IMS, Stuttgart (German)
- Still on-going, with many more sites added over the years.
- ParGram⁵ was led by PARC for many years (resources, etc. now maintained in Konstanz and Bergen).
- (PARC dropped out after the successful start-up *Powerset*, which was then bought by Microsoft)
- Goal: Computational grammar development for diverse sets of languages
 - via a joint development platform (XLE)⁶
 - with a common linguistic understanding (LFG)



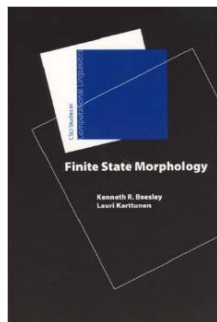
⁵<https://pargram.w.uib.no>

⁶<https://ling.sprachwiss.uni-konstanz.de/pages/xle/index.html>



- Languages over the years:
 - English, French, German, Hungarian, Norwegian, Polish, Welsh
 - Georgian, Malagasy, Indonesian, Urdu, Turkish
 - Chinese, Japanese
 - Tigrinya, Wolof

- Resources created:
 - Language particular stand-alone morphological analyzers (typically implemented with xfst).
 - Language particular lexicons with subcategorization information.
 - Treebanks, a subset of which are parallel and aligned (Sulger et al. 2013).
 - Write-Ups on design decisions.
 - Starter Grammars — to bootstrap new languages on the basis of multilingual experience.





ParGram – Starter Grammars

- Specifications for Features and Feature Spaces based on comparative linguistic knowledge.
- For example: values for gender, number, case, verbal type, quantifiers, determiners.
- Design decisions informed by a typologically diverse set of languages.
- Sample feature space for specifiers:

SPEC: $\rightarrow\langle\langle$ [ADJUNCT AQUANT DET NUMBER POSS QUANT].

DET: $\rightarrow\langle\langle$ [DEIXIS DET-TYPE PRED].

DET-TYPE: $\rightarrow\$\$ {def demon indef int rel}.

def: definite (“the box”)

demon: demonstrative (“this box”)

indef: indefinite (“a box”)

int: interrogative (“which box”)

rel: relative (“the girl whose box broke”)



- Also capture generalizations expected to be found across languages via pre-defined **templates**.
- For example: passivization, coordination.
- Furthermore: interface to allow corpus-based statistical/frequency information to inform grammar analyses (“most likely parse”).

INESS and XLE

- XLE still needs to be obtained via a license from PARC.
- Nice web interface available at Bergen (INESS)⁷
 - Interact with existing diverse set of grammars
 - Upload own grammars
 - Access to treebanks
- **ParGramBank** collects parallel and aligned treebanks across a diverse set of languages (Sulger et al. 2013):
 - English, French, Georgian, German, Hungarian, Norwegian, Polish, Turkish, Urdu, Wolof.
 - Currently adding **Tamil**

The screenshot shows the INESS XLE-Web interface. The browser address bar is `clarino.uib.no/iness/xle-web`. The page title is "INESS XLE-Web". The user is logged in as "Local | Clarin SPFF | eduGAIN".

The interface includes a navigation menu on the left with links for Home Page, Knowledge center, The project, Documentation, FAQ, Publications, Links, and Resources. Below this are sections for Treebanks, Tools, and Grammars.

The main content area shows a form for entering a sentence. The grammar is set to "English". The sentence entered is "Covid is causing shut-downs.". Below the form are several checkboxes for parsing options: "Packed representation" (checked), "Suppress CHECK", "Show XLE messages", "Suppress complex categories", "Show discriminant weights", "GIT update grammar", "Show unoptimal", "PREs only", "Include non-top F-structures", "Show discriminants", "c-structure", and "f-structure".

The results section shows "2+4 solutions, 0.017 CPU seconds, 3.233MB max mem, 111 subtrees unified". Two structures are displayed: "C-structure" and "F-structure".

C-structure: A hierarchical tree diagram for the sentence "Covid is causing shut-downs.". The root is "ROOT", which branches into "Sadj[fin]" and "PERIOD". "Sadj[fin]" branches into "S[fin]" and "VPall[fin]". "S[fin]" branches into "NP" and "VP[prog,fin]". "NP" branches into "NPpobj" and "NPzero". "NPpobj" branches into "N" (labeled "Covid") and "NPzero". "VP[prog,fin]" branches into "AUX[prog,fin]" (labeled "is") and "VP[prog]". "VP[prog]" branches into "V[prog]" (labeled "causing") and "NP". "NP" branches into "NPpobj" and "NPzero". "NPpobj" branches into "N" (labeled "shut-downs") and "NPzero".

F-structure: A table representing the semantic structure of the sentence. The table has columns for "PRED", "TNS-ASP", "OBJ", and "SUBJ". The rows represent different parts of the sentence: "cause", "shut-down", and "Covid".

PRED	TNS-ASP	OBJ	SUBJ
'cause<[<:covid], [2:shut-down]>'	TENSE pres, PROG +, PERF -> MOOD indicate	PRED 'shut-down' NTYPE NSEM COMMON count PERS 3, NUM pl, CASE obl	PRED 'covid' NTYPE NSEM common PERS 3, NUM sg, CASE nom
		VTYPE main, PASSIVE -, CLAUSE-TYPE dcd	

⁷ <https://clarino.uib.no/iness/xle-web>

Developing NLP Resources for Tamil

- DAAD Sri Lanka-German cooperation with the University of Moratuwa (UoM)
- UoM is looking at building NLP applications for Tamil and Sinhalese.
- UoM cooperation partners: K. Sarveswaran, Gihan Dias.
- Usual trajectory:
 - Initial experiments with machine learning yielded low results (not enough useful resources)
 - Though some success with closed domain machine translation.
 - Decided to take a step back and begin with building resources via rule-based systems.
 - These systems can produce data to enable downstream ML.
- Decided to invest time building a ParGram style grammar for Tamil.



- Targeted Resources (work by K. Sarveswaran)
 - (Stand-alone) Finite-State Morphology
 - Lexicons
 - Automatically generated parses that can be stored in a treebank (cf. King et al. 2003, INESS).
 - The corpus in form of an annotated treebank can then be fed into ML systems.
 - Link to Universal Dependencies (UD)
 - LFG already contains a dependency representation
 - UD was in fact inspired by LFG's system of organization
 - Conversion/Relation to UD type treebanks should be trivial



Tamil ParGram Grammar

- Tamil is a challenging case for NLP.
 - Complex orthography (see below).
 - Complex morphophonology (mostly agglutinative) that is poorly described.
 - Complex predication (e.g. 'mistake do' in (2)), not well understood.
 - Interesting syntax, also poorly described.
 - No idea about semantics or pragmatics.

- Example with an embedded clause – note inflections on the complementizer.

(2) [avan pizhai sey-tt-aan **enp-a-athu-ai**] ram
 [pron.3sm.nom mistake do-past-3sm **comp-rel-pron.3sn-acc**] Ram.nom
 nirupi-tt-aan
 prove-past-3sm
 'Ram proved (the fact) that he made a mistake.'



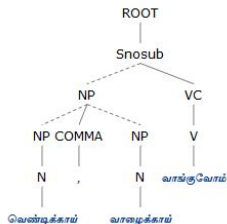
Tamil ParGram Grammar

- Tamil is a linguistically under-researched language.
 - Much of the descriptive material is over 100 years old.
 - Very few linguists with Tamil expertise.
 - Of these, almost none with an understanding of how to organize information to make it of computational use.

- I have been able to
 - 1 leverage **comparative linguistic** knowledge of the structure of South Asian languages to help in building Tamil computational resources
 - 2 clear up misunderstandings or false claims in the literature (e.g., 38 auxiliaries...!)
 - 3 leverage the **ParGram multilingual experience** and knowledge towards an efficient implementation of a Tamil grammar
 - 4 including typologically unusual phenomena like inflected complementizers

Tamil ParGram Grammar on INESS

C-structure



F-structure

PRED	'வாங்கு<[12:pro], [4]>'																						
TNS-ASP	14	TENSE fut, MOOD indicative																					
OBJ	{	<table border="1"> <tr> <td>PRED</td> <td colspan="2">'வெண்டிக்காய்'</td> </tr> <tr> <td>NTYPE</td> <td>11</td> <td>COMMON count</td> </tr> <tr> <td>NSYN</td> <td>10</td> <td>common</td> </tr> <tr> <td>RATIONAL</td> <td>-</td> <td>PERS 3, NUM sg,</td> </tr> <tr> <td>HUMAN</td> <td>-</td> <td>GEND neut, CASE nom,</td> </tr> <tr> <td>ANIM</td> <td>-</td> <td>9</td> </tr> <tr> <td>PERS</td> <td>3</td> <td>GEND neut, COORD-LEVEL NP, CASE nom</td> </tr> </table>	PRED	'வெண்டிக்காய்'		NTYPE	11	COMMON count	NSYN	10	common	RATIONAL	-	PERS 3, NUM sg,	HUMAN	-	GEND neut, CASE nom,	ANIM	-	9	PERS	3	GEND neut, COORD-LEVEL NP, CASE nom
PRED	'வெண்டிக்காய்'																						
NTYPE	11	COMMON count																					
NSYN	10	common																					
RATIONAL	-	PERS 3, NUM sg,																					
HUMAN	-	GEND neut, CASE nom,																					
ANIM	-	9																					
PERS	3	GEND neut, COORD-LEVEL NP, CASE nom																					
	,	<table border="1"> <tr> <td>PRED</td> <td colspan="2">'வாழைக்காய்'</td> </tr> <tr> <td>NTYPE</td> <td>8</td> <td>COMMON count</td> </tr> <tr> <td>NSYN</td> <td>7</td> <td>common</td> </tr> <tr> <td>RATIONAL</td> <td>-</td> <td>PERS 3, NUM sg,</td> </tr> <tr> <td>HUMAN</td> <td>-</td> <td>GEND neut, CASE nom,</td> </tr> <tr> <td>ANIM</td> <td>-</td> <td>5</td> </tr> </table>	PRED	'வாழைக்காய்'		NTYPE	8	COMMON count	NSYN	7	common	RATIONAL	-	PERS 3, NUM sg,	HUMAN	-	GEND neut, CASE nom,	ANIM	-	5			
PRED	'வாழைக்காய்'																						
NTYPE	8	COMMON count																					
NSYN	7	common																					
RATIONAL	-	PERS 3, NUM sg,																					
HUMAN	-	GEND neut, CASE nom,																					
ANIM	-	5																					
SUBJ	12	PRED 'pro' RATIONAL +, PERS 1, NUM pl, GEND epi																					
VTYP	3	main, PASSIVE -, CLAUSE-TYPE decl																					

- (3) vendikkaai, vaalaikkaai vankkuvooom
 okra ash plantain buy.fut.1pl
 '(We) will buy okra and ash plantain.'



- UoM goal: build openly accessible resources.
- Several previous stabs at morphological analyzers for Tamil have been made, none are available.
- Current effort programmed in OpenFST.
- After one year of concerted effort, contains all inflectional forms (Sarveswaran et al. 2019).
- Grammar and morphological analyzer now able to parse elementary school textbooks.



- ML is currently dominant within NLP.
- The knee-jerk reaction to any sort of NLP task is to throw ML at it.
- Often with quite low results (50%-60%), yet these are reported as valuable research.
- Observation:
 - the morphology of a language has only finite-state complexity;
 - the morphological inflections are finite in nature
- ParGram experience has shown that a concerted effort of 1-2 years tends to yield a robust, workable finite-state morphological analyzer for a language
 - the technology is not difficult to use/program
 - the algorithms and complexity issues are well understood

Discussion: Resource Creation



- Not clear why a finite, computationally eminently already solvable problem like the construction of a morphological analyzer should be tried via ML.
- Except — one needs **linguistic knowledge**.
- This is difficult for computer scientists to acquire.
- It is also difficult for anybody to acquire for understudied languages.
- What should the solution be?



- Not clear why a finite, computationally eminently already solvable problem like the construction of a morphological analyzer should be tried via ML.
- Except — one needs **linguistic knowledge**.
- This is difficult for computer scientists to acquire.
- It is also difficult for anybody to acquire for understudied languages.
- What should the solution be?

Invest in Linguistics!

Concluding Remarks



- Most of the efforts that have produced usable, high-quality resources for further NLP processing/applications:
 - Partnered with (computationally interested) linguists.
 - Invested heavily in (intelligent) manual annotation.
 - Annotation guidelines were developed and refined over time in several cycles.
 - Annotation was guided by theory and linguistic insights, involving deep and sometimes difficult discussions (not invented on the fly within the space of a limited project).
- Successful NLP approaches have been able to build on such resources.

Concluding Remarks



- But there is a draw-back: many annotations and applications are very English oriented.
- English is a typologically odd language.
- Probably due to its long and varied history of language contact.
- NLP methodology that is good for English is not necessarily working out for languages with different structures.
- (Word Embedding is more promising)
- So **comparative linguistic** work and knowledge is becoming ever more crucial as NLP seeks to expand successfully into a wider variety of languages.

→ **SIGTYP a very timely enterprise!**



Thanks!

The work presented here was done in collaboration with many people over many years.

Konstanz: Tina Bögel, Annette Hautli-Janisz, Benazir Mumtaz.

Lahore: Farah Adeeba, Toqeer Ehsan, Benazir Mumtaz, Sarmad Hussain, Sana Shams, Saba Urooj.

Karachi: Tafseer Ahmed.

Colombo: Gihan Dias, Kengathariyer Sarveswaran.

ParGram (close collaborators): Mary Dalrymple, Stefanie Dipper, Helge Dyvik, Anette Frank, Martin Forst, Ronald Kaplan, Tracy Holloway King, John Maxwell III, Paul Meurer, Christian Rohrer, Victoria Rosén, Koenraad de Smedt, Annie Zaenen.

References I

- Ahmed Khan, Tafseer, Saba Urooj, Sarmad Hussain, Asad Mustafa, Rahila Parveen, Farah Adeeba, Annette Hautli, and Miriam Butt. 2014. The CLE Urdu POS tagset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2920–2925. Reykjavik, Iceland: European Language Resources Association (ELRA).
- Bender, Emily M. and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198. Online: Association for Computational Linguistics.
- Bhat, Riyaz Ahmad, Rajesh Bhatt, Annahita Farudi, Prescottt Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, and Fei Xia. 2017. The Hindi/Urdu Treebank Project. In N. Ide and J. Pustejovsky, eds., *Handbook of Linguistic Annotation*. Dordrecht: Springer.
- Ehsan, Toqeer and Miriam Butt. 2020. Dependency parsing for Urdu: Resources, conversions and learning. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5202–5207. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.
- Hardie, Andrew. 2003. Developing a tag-set for automated part-of-speech tagging in Urdu. In D. Archer, P. Rayson, A. Wilson, and T. McEnery, eds., *Proceedings of the Corpus Linguistics 2003 Conference*.
- Hussain, Sarmad. 2005. Phonological Processing for Urdu Text to Speech System. In Y. Yadava, G. Bhattarai, R. R. Lohani, B. Prasain, and K. Parajuli, eds., *Contemporary Issues in Nepalese Linguistics*. Linguistics Society of Nepal.
- Kalouli, Aikaterini-Lida, Richard Crouch, and Valeria de Paiva. 2020. Hy-NLI: a Hybrid system for Natural Language Inference. In *COLING 2020*.

References II

- King, Tracy Holloway, Richard Crouch, Stefan Riezler, Mary Dalrymple, and Ronald Kaplan. 2003. The PARC700 Dependency Bank. In *Proceedings of the EACL03: 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*.
- Sajjad, Hassan and Helmut Schmid. 2009. Tagging Urdu Text with Parts of Speech: A Tagger Comparison. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*, pages 692–700.
- Sankaran, Baskaran, Kalika Bali, Monojit Choudhury, Tanmoy Bhattacharya, Pushpak Bhattacharya, Girish Nath Jha, S. Rajendran, K. Saravanan, L. Sobha, and K. V. Subbarao. 2008. A common parts-of-speech tagset framework for indian languages. In *Proceedings of LREC 2008*, pages 1331–1337. European Language Resources Association.
- Sarveswaran, Kengatharaiyer, Gihan Dias, and Miriam Butt. 2019. Using meta-morph rules to develop morphological analysers: A case study concerning Tamil. In *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing*, pages 76–86. Dresden, Germany: Association for Computational Linguistics.
- Schmidt, Ruth Laila. 1999. *Urdu: An Essential Grammar*. London: Routledge.
- Sulger, Sebastian, Miriam Butt, Tracy Holloway King, Paul Meurer, Tibor Laczkó, György Rákosi, Cheikh Bamba Dione, Helge Dyvik, Victoria Rosén, and Koenraad De Smedt. 2013. Pargrambank: The pargram parallel treebank. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pages 550–560.
- Urooj, Saba, Benazir Mumtaz, and Sarmad Hussain. 2019. Urdu intonation. *Journal of South Asian Linguistics* 10.