CoRSAL Symposium, University of North Texas, November 17, 2017

Existing (and Potential) Language and Linguistic Resources on South Asian Languages

Elena Bashir, The University of Chicago

Resources or published lists outside of South Asia

Digital Dictionaries of South Asia in Digital South Asia Library (dsal), at the University of Chicago. http://dsal.uchicago.edu/dictionaries/ . Some, mostly older, not under copyright dictionaries. No corpora.

Digital Media Archive at University of Chicago
https://dma.uchicago.edu/about/about-digital-media-archive

Hock & Bashir (eds.) 2016 appendix. Lists 9 electronic corpora, 6 of which are on Sanskrit. The 3 non-Sanskrit entries are: (1) the EMILLE corpus, (2) the Nepali national corpus, and (3) the LDC-IL — Linguistic Data Consortium for Indian Languages

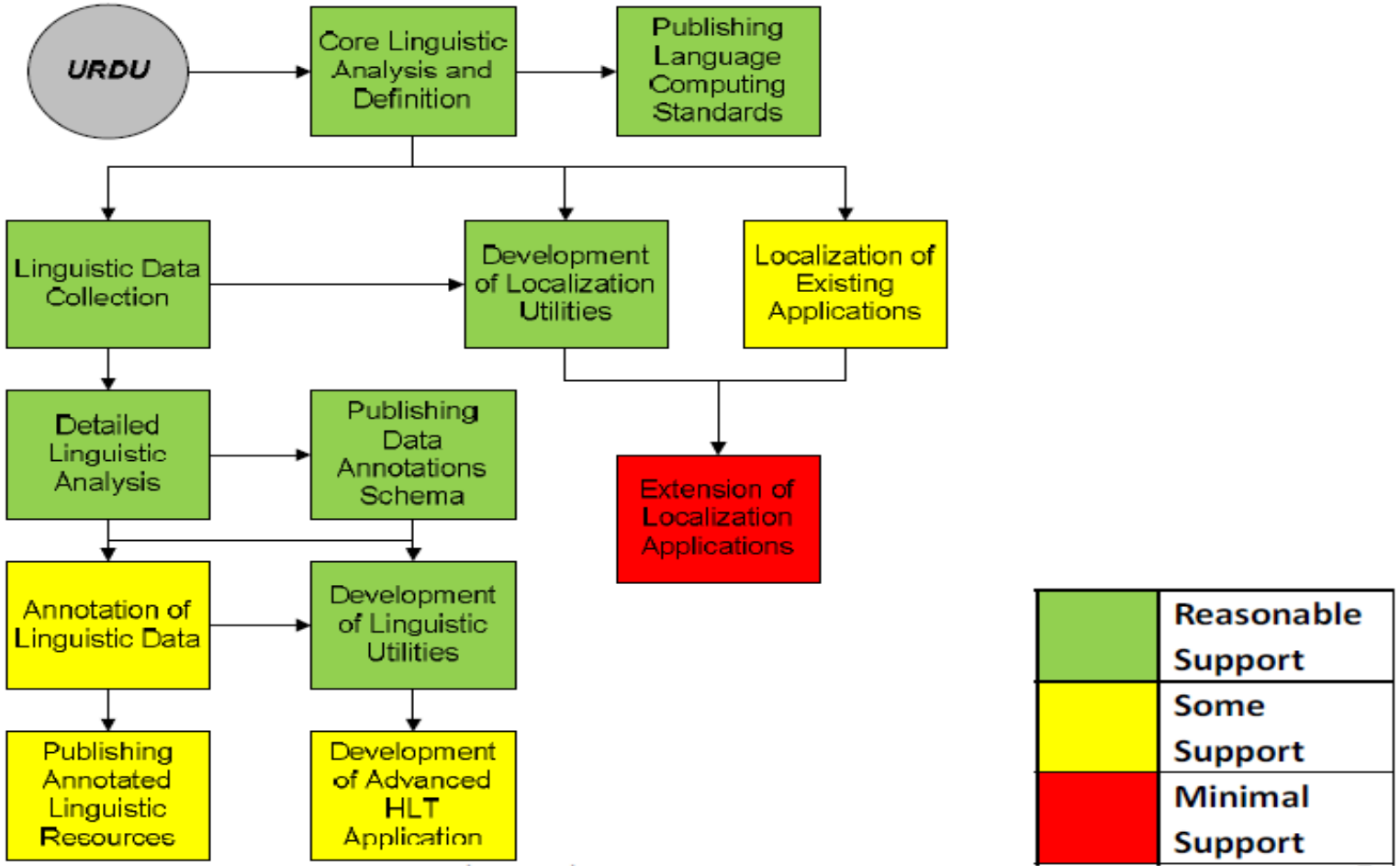
Focus on Pakistan

Developing Standards and Linguistic Resources for Computational Research in Pakistani Languages

www.CLE.org.pk

Sarmad Hussain
Professor
Center for language Engineering (CLE)
Al-Khawarizmi Institute of Computer Sciences (KICS)
University of Engineering and Technology (UET) Lahore
sarmad.hussain@kics.edu.pk

Status of Human Language Technology



Urdu

Most work has been done on Urdu, prioritized at government institutions like the Center for Language Engineering at the University of Engineering and Technology in Lahore (CLE).

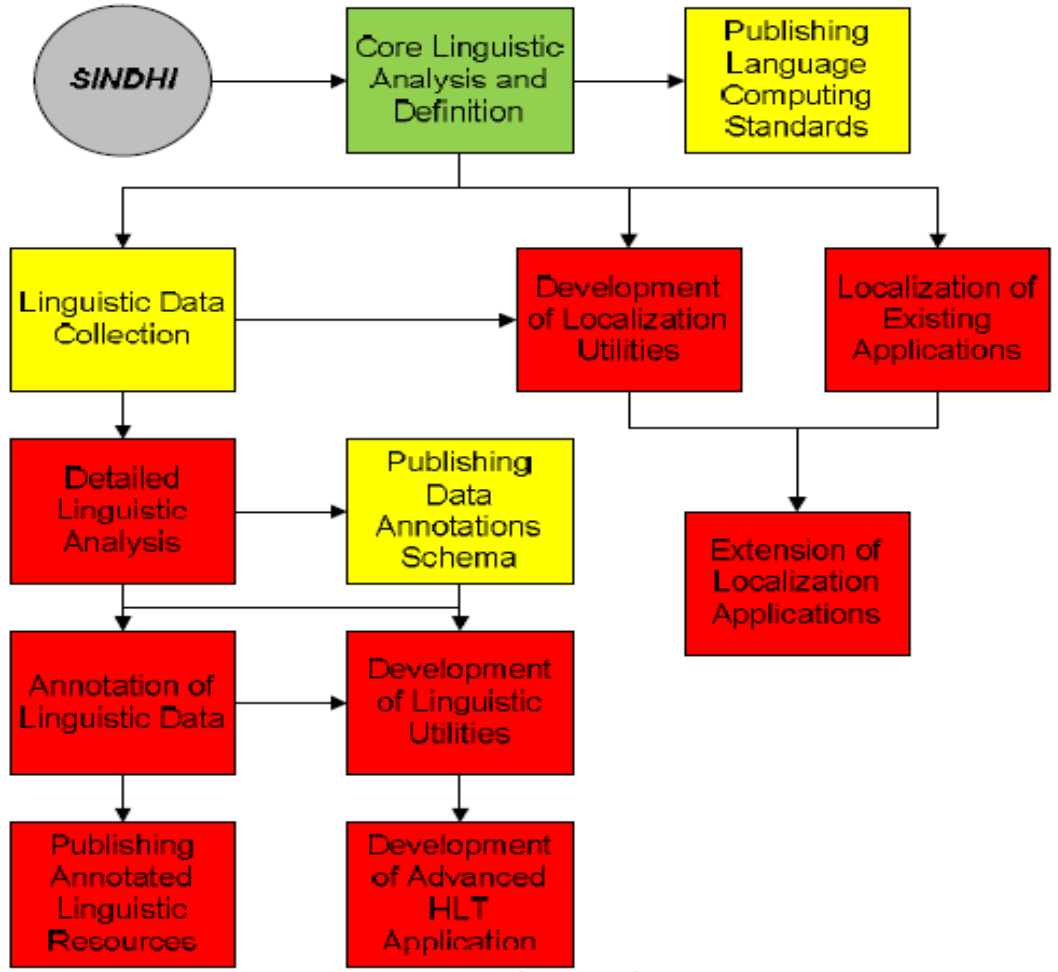
Text corpora: http://cle.org.pk/clestore/index.htm (largest is a 1 million word Urdu corpus from the *Urdu Digest*.

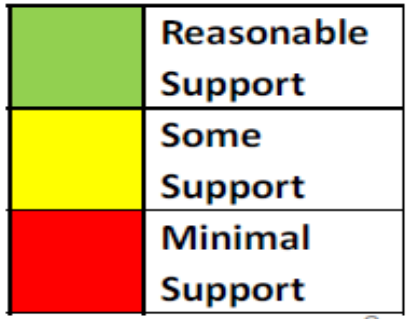
Work on Essential Urdu Linguistic Resources: http://www.cle.org.pk/eulr/

Tagset for Urdu corpus:
http://cle.org.pk/Publication/papers/2014/The%20CLE%20Urdu%20POS%20Tagset.pdf

Urdu OCR: http://cle.org.pk/clestore/urduocr.htm

Status of Human Language Technology





Sindhi

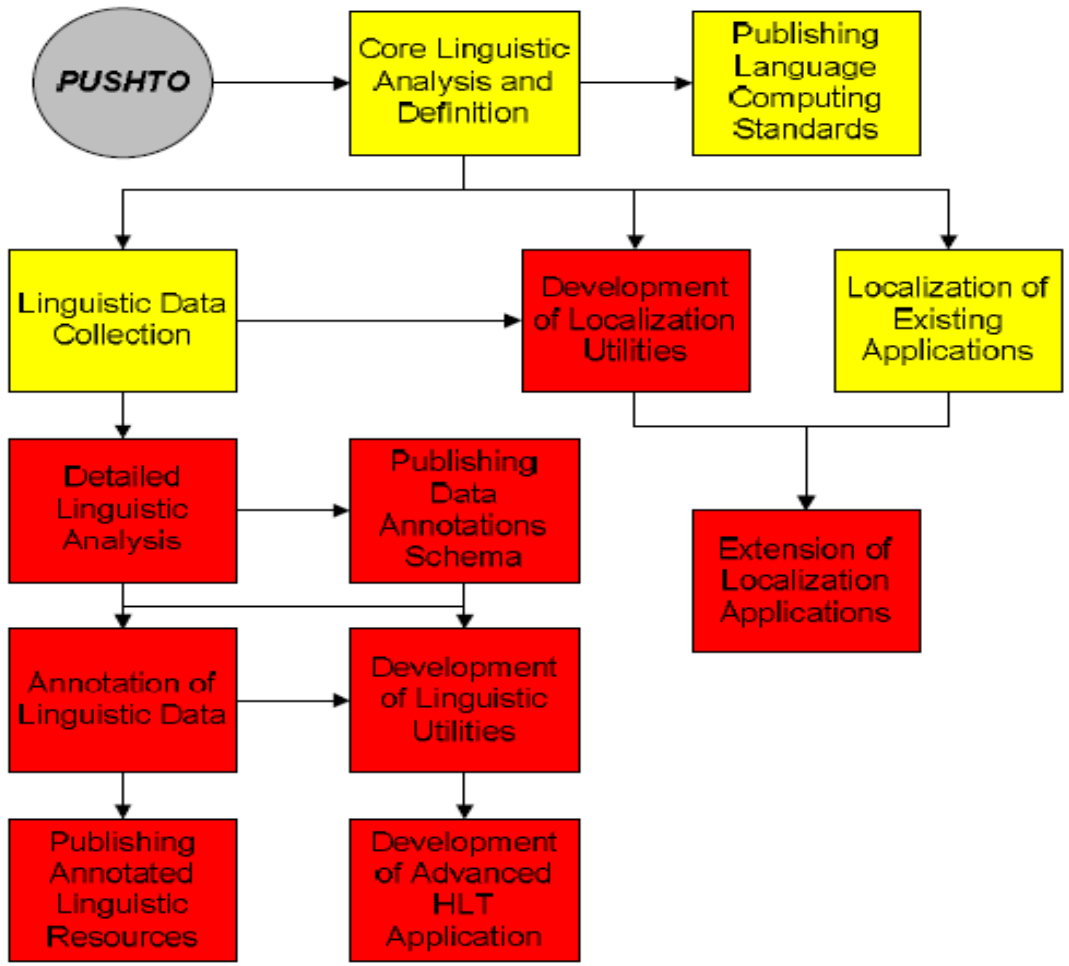
Sindhi is the medium of education in some schools in Sindh
Has more institutional backing and consequent research than other languages, especially Panjabi.

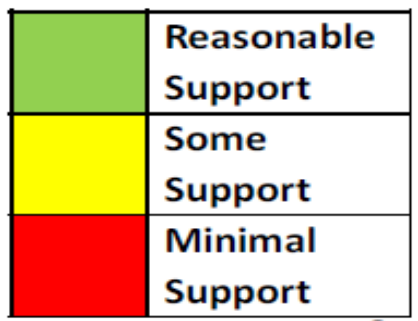
Sindhi-English dictionary developed jointly by Jennifer Cole at the University of Illinois Urbana-Champaign and Sarmad Hussain at CLE (http://182.180.102.251:8081/sed1/homepage.aspx). Has clickable alphabetical word index and is searchable by either Sindhi-script entries or by roman representation.

Work toward Sindhi corpus building (Rahman 2010)

Sindhi Language Authority (http://sindhila.edu.pk/index.php) – quite active
- online Sindhi dictionary (http://dic.sindhila.edu.pk/index.php?txtsrch)
- online Sindhi learning portal (http://learn.sindhila.edu.pk/)
- online specialized dictionaries, Sindhi>English and English>Sindhi

Status of Human Language Technology





Pashto

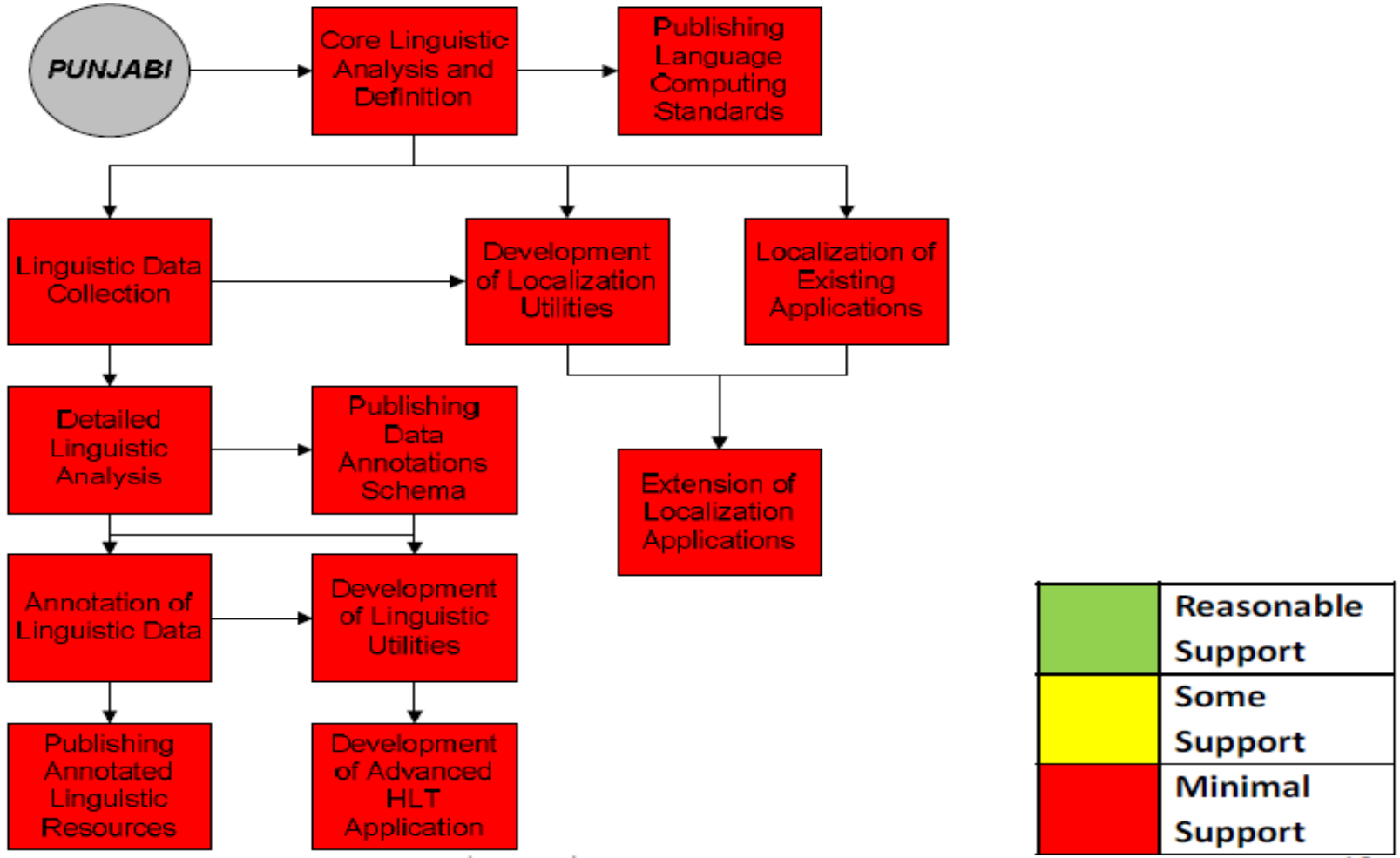
Computational work on Pashto is mainly being carried out at the Department of Computer Science, University of Peshawar, Pakistan.

Khan & Zuhra (2007) describes the development of a 10,000 word, open-ended corpus for Pashto.

A Pashto transcription project was initiated at CLE in 2012, progress to date?

Pashto Academy, Peshawar (located in Peshawar University, some government support), literary studies, no linguistic work

Status of Human Language Technology



Panjabi

Largest number of speakers in Pakistan.

Severely under-resourced in Pakistan; much work (in Gurmukhi script) in India.

Virk (2013), dissertation on computational resources for Indo-Iranian languages: "With around 100 million native speakers, it is the 12th most widely spoken language in the world. When it comes to the computational resources, it is hard to find any grammatical resources for this language." This work is on building a computational grammar for Panjabi, but it too does not focus on lexical or corpus resources.

No Punjabi Language Authority comparable to Sindhi Language Authority

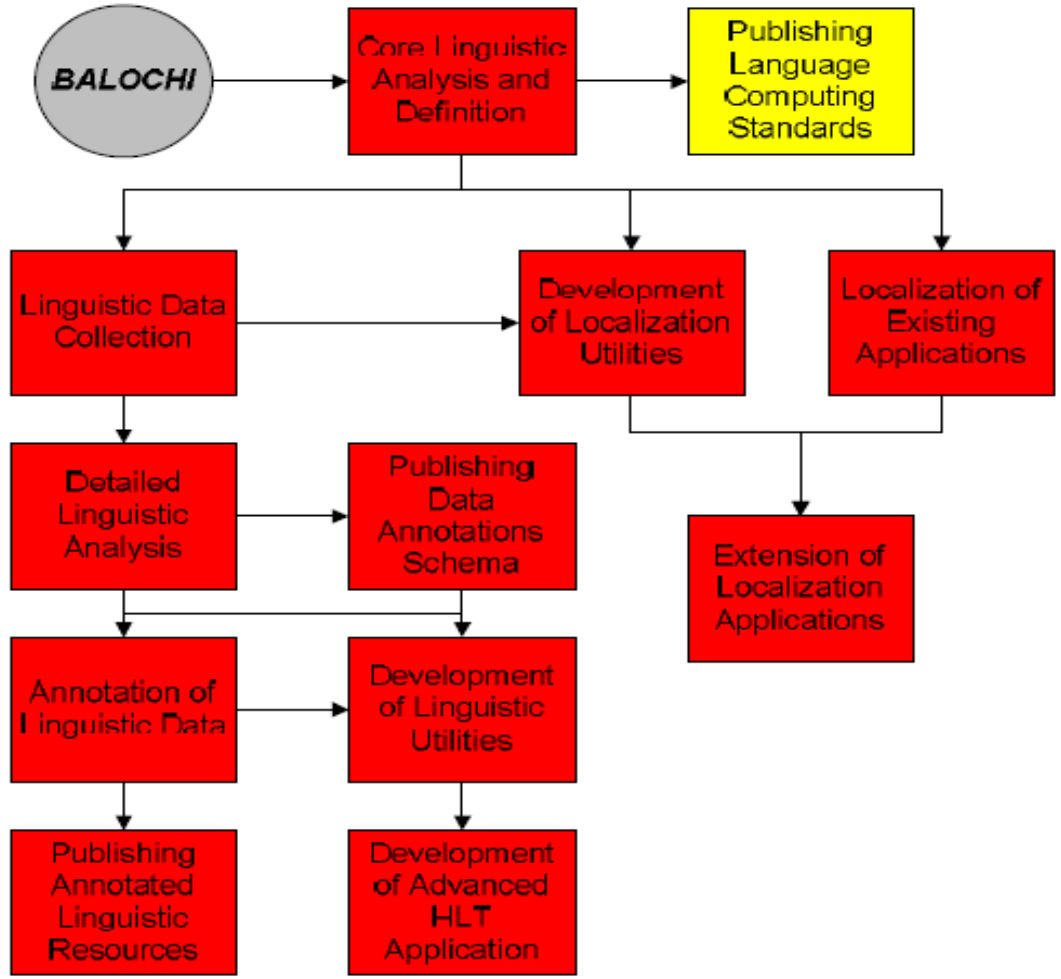
Only one paper on Panjabi (in Perso-Arabic) in Pakistan. http://www.aclweb.org/anthology/Y10-1020

Mention of one article, entitled "Lexical database of Pakistani Regional Languages" found from internet search, but without the author's name and without any way to access it.

Punjab Institute of language, Art & Culture exists, (some support from Punjab Government, but no linguistic content, no events reported since 2012) (http://www.pilac.punjab.gov.pk/)

My own work in preparation on a descriptive grammar of Panjabi, Hindko, and Saraiki.

Status of Human Language Technology



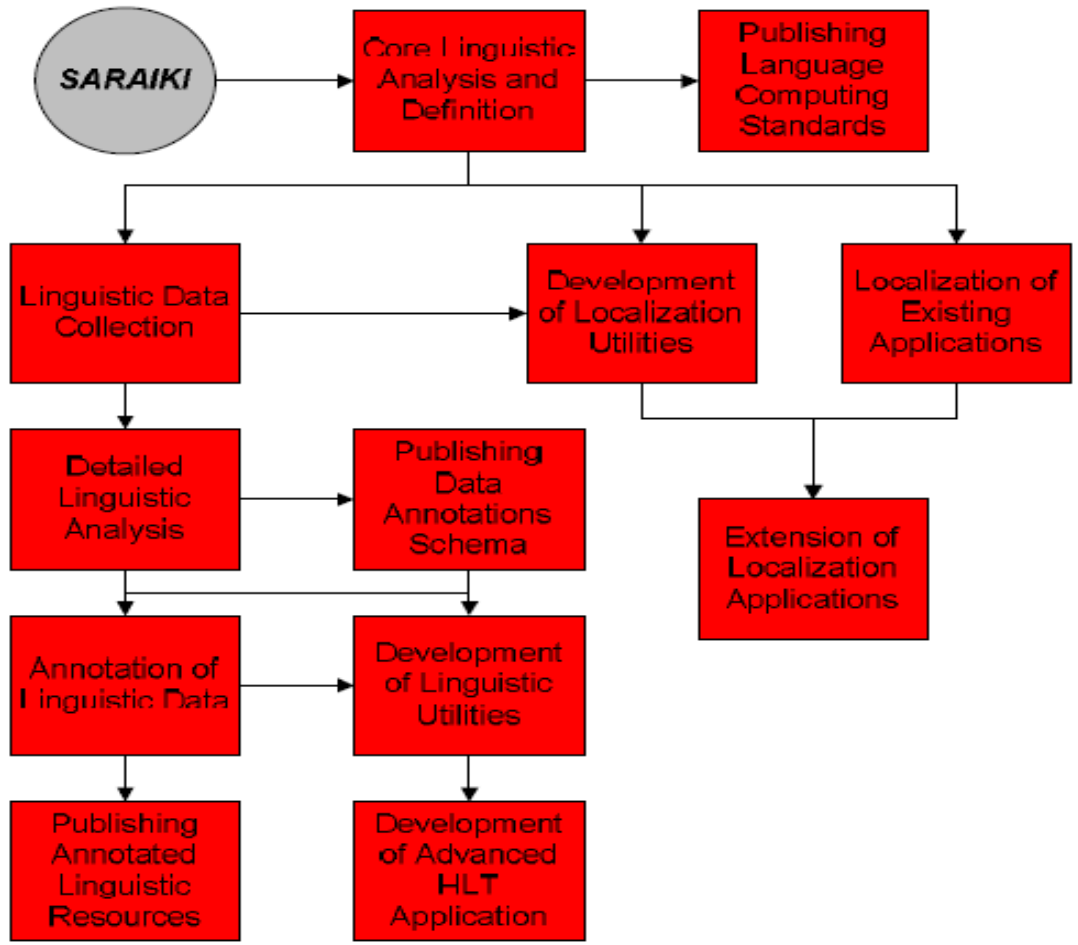
	Reasonable Support
	Some Support
	Minimal Support

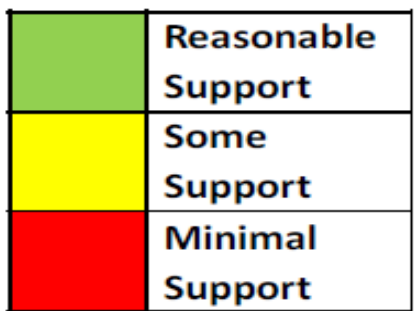
Balochi

Balochi Academy, Quetta (partial government support).
Many print publications, including 20 dictionaries and publications on grammar.
One journal, with only one issue since 2012.

No database or corpus work known to me.

Status of Human Language Technology





Saraiki

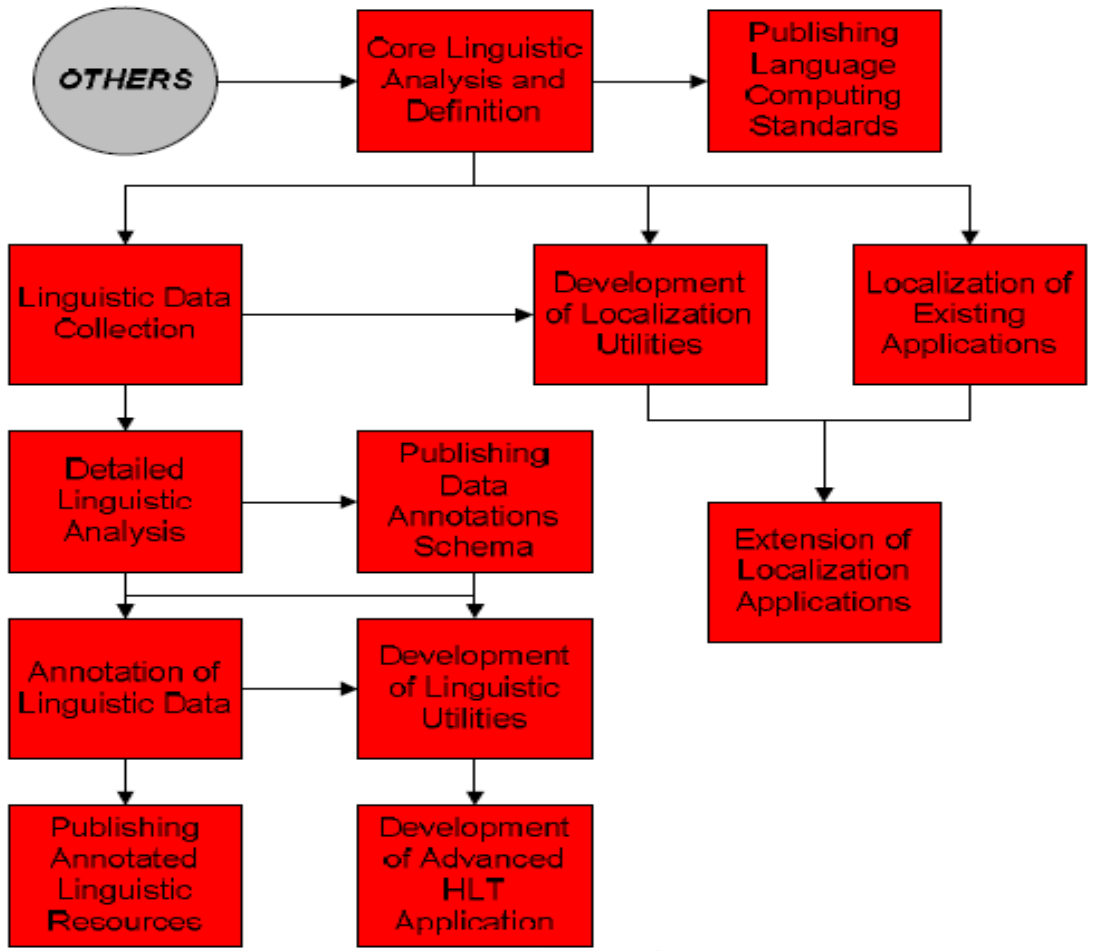
Much interest in Saraiki (Siraiki) language among its speakers.

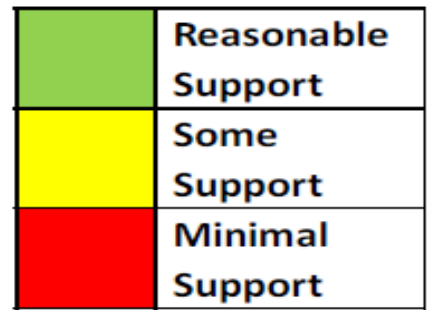
Department of Saraiki at Bahauddin Zakariya University, Multan

Association for Saraiki Advanced Area Studies (ASAAS) - planning to develop a Saraiki online dictionary and corpus

No corpus resources yet underway.

Status of Human Language Technology





Others -1

Brahui

https://brahuiacademy.org/ (some government support)
Print publications.
Held International Conference in 2015.

Burushaski

Burushaski Research Academy – 3-volume Burushaski-Urdu dictionary

Web-accessible Burushaski archive by Sadaf Munshi at UNT:
http://ltc.unt.edu/~sadafmunshi/Burushaski/archive.html

Torwali

Torwali-Urdu dictionary compiled by Inam Ullah, a native speaker of Torwali (http://www.cle.net.pk/otd/)

Example best known to me of community-oriented work initiated and carried out by a member of the local speech community, Inam Ullah. Has made presentations on language documentation and very active in one of the many local, community-based organizations. Has served as Chairperson, Mother Tongue and Heritage for Education and Research (MOTHER). Some excerpts from documents prepared by him follow.

"The goal of this project is to promote and enhance education among the school students of a low-resourced speech community through utilizing native lexical resources created locally during the past couple of decades so as to bring about a behavioral change in them towards the preservation of their ancestral language and culture." (Inam Ullah 2015-16)

"To meet these goals a user-friendly set of books comprising one trilingual dictionary and two reverse indices (or glossaries) will be disseminated among all the community students reading in classes 9th and 10th of the government high schools (total number1000) during academic year 2017." (Inam Ullah 2015-16)

Also, Idara Baraye Taleem Taraqqi (https://www.facebook.com/IdaraBarayeTaleemTaraqi/)

Regional/national level institutions

Forum for Language Initiatives, formerly Frontier Language Institute-Peshawar, (FLI) is an NGO (mother-tongue literacy training, capacity development, provision of technical resources, publication) (http://fli-online.org/site/)

Department of Pakistani Languages at Allama Iqbal Open University in Islamabad.

However, so far no government sponsored effort focused specifically on documentation of languages under threat. Localization efforts by the computational linguistics community in Pakistan are still focused on the larger languages, but there are indications that the techniques developed by computational linguists will also be deployed in efforts to document smaller languages (see Rahman 2004, Inam Ullah 2012).

Linguistics-related civil-society organizations of interested scholars in Pakistan

1. Linguistic Association of Pakistan (LAP)

2. Society for Natural Language Processing, Pakistan (SNLP) http://www.snlp.org.pk/

2. Pakistan Association for Lexicology and Lexicography (PakLex)

3. Pakistan Association for Developing Formal Grammars (PakGram)

4. Pakistan Association for Interpreting and Translation Studies (PAITS)

5. Pakistan Association for Inducing Corpus- Oriented Research (PAICOR)

6. Pakistan Association for Maintaining and Preserving Indigenous Languages (PAMPIL) – very new

Community-based organizations

Khowar. Anjuman Taraqqi-e-Khowar – numerous print publications, no website

http://www.mahraka.com/ (website on language, literature and culture of Chitral) – section on e-Khowar (writing Khowar on various operating sysems

Gawri. Gawri Cultural Society (Gawri)

Hindko. Gandhara Hindko Board and Gandhara Hindko Academy http://www.gandharahindko.com/. 137 published books. Hold frequent awareness building conferences.

Wakhi. Wakhi-Tajik Cultural Association (https://www.facebook.com/Wakhi-Tajik-Cultural-Association-WTCA-302214366610133/)

Shina. Shina Language and Culture Promotion Society (http://www.saskenshina.com/en/books)

Potential institutional and leadership resources

Center for Language Engineering (CLE) (Sarmad Hussain) http://cle.org.pk/

Linguistic Association of Pakistan (LAP) http://www.lap.org.pk/ Presidential address at ICLAP 2017 by Ghulam Raza, himself a computational linguist, was entitled "Building Linguistic Resources and Developing Language Technologies - Synergetic Efforts Needed in Pakistan"

Society for Natural Language Processing, Pakistan (SNLP) http://www.snlp.org.pk/

PAKLEX – Aim is to "promote the research activities relevant to lexicology and lexicography in the country of Pakistan". Not much done so far, but the interest is there. (Tafseer Ahmed Khan and Ghulam Raza) http://www.paklex.org/

Computer Science Department, Peshawar University (Muhammad Abid) http://www.uop.edu.pk/faculties/?q=Faculty-of-Management-and-Information-Sciences

Department of Computer Science, DHA Suffa University, Karachi. Has Center for Research in Computational Linguistic (CRCL) (Tafseer Ahmed Khan) http://www.dsu.edu.pk/index.php/en/academics/computer-science

Documentation studies on smaller languages of Northern Pakistan

Language	Endangerment status	Documentation [D] and Vitality [V] Studies
Bateri (Baterawal Kohistani, *bhaT'esa z'ib*)	Definitely endangered	Zoller (2005) [D]; Strand (2011b) [D]
Brahui	Vulnerable	Bashir (2010) [D]0
Burushaski	Vulnerable	Burushaski Research Academy (2007, 2009) [D]; Jammu and Kashmir Burushaski Munshi (2006, 2010) [V, D]; Hunza Burushaski Munshi (2009, 2010–present) [D]
Chilisso	Severely endangered	Hallberg (1992b) [D, V]
Dameli (Damia)	Severely endangered	Perder (2008) [V, D], Perder (2013) [D]
Domaki	Severely endangered	Weinreich (1999) [D]; Weinreich (2009) [D]; Weinreich (2010) [V]; Tikkanen (2011) [D]

Documentation studies on smaller languages of Pakistan - 2

Language	Endangerment status	Documentation [D] and Vitality [V] Studies
Gawar-Bati	Definitely endangered	Decker, K. D. (1992b) [V]
Gowro (Gabaro)	Severely endangered	Hallberg, D. G. (1992b) [V]
Indus Kohistani (“Maiyã”)	Vulnerable	Hallberg (1992) [D, V]; Hallberg, D. G. & Hallberg, C. E. (1999) [D]; Zoller (2005) [D]
Kalam Kohistani (Bashkarik, Gawri, Kalami)	Definitely endangered	Baart (1997, 1999, 2004) [D];14 Baart & Sagar (2004) [D]
Kalasha (Kalashamon[dr])	Severely endangered	Bashir (1988) [D]; Peterson (2006) [D]; Trail & Cooper (1999) [D]; Mørch (2000) [V]; Decker, K. D. (1992c) [V]
Kati	Definitely endangered	Strand (2011a) [D]

Documentation studies on smaller languages of Pakistan - 3

Language	Endangerment status	Documentation [D] and Vitality [V] Studies
Khowar	Vulnerable (somewhat less so now)	Bashir (in progress) [D]; Decker, K. D. (1992a) [V]
Kundal Shahi	Definitely endangered	Rehman & Baart (2005) [V] [D]
Ormuri	Definitely endangered	Burki (2001) [V]; Hallberg, D. G. (1992a) [V]; Efimov 2011[1986]), English translation by Baart in 2011 [D]
Palula (Phalura, Dangarikwar)	Definitely endangered	Decker, K. D. (1992a) [V]; Strand (2000) [D]; Liljegren (2008) [D]
Torwali	Definitely endangered	Inam Ullah (2010a, 2010b) [D]
Ushojo	Definitely endangered	Decker, S. J. (1992) [V]

Documentation studies on smaller languages of Pakistan - 4

Language	Endangerment status	Documentation [D] and Vitality [V] Studies
Wakhi	Definitely endangered	Mock (1998) [V, D]; Müller et al. (2008) [V]
Yidgah	Definitely endangered	Janjua (2011) [V]

Nepal

Nepali National Corpus https://www.sketchengine.co.uk/nepali-national-corpus/
Uses tagset at: https://www.sketchengine.co.uk/nepali-tagset/

Speech and written corpora:
http://web.archive.org/web/20150826233511/http://www.bhashasanchar.org:80/aboutnnc.php

Bangladesh

Tagset(s) for Bangla. Being thought about; apparently numerous proposals.

https://github.com/abhishekgupta92/bangla_pos_tagger
http://computing.dcu.ie/~sdandapat/publication/thesis/MS_thesis.pdf
http://www.panl10n.net/english/outputs/Working%20Papers/Bangladesh/Microsoft%20Word%20-%206_N_37.pdf

Extent of consensus so far ?

Selected References
Some classic early works and recent works relevant mostly to corpus construction

Baker, Paul, Andrew Hardie, Tony McEnery, Richard Xiao, Kalina Bontcheva, Hamish Cunningham, Robert Gaizauskas, Oana Hamza, Diana Maynard, Valentin Tablan, Cristian Ursu, B. D. Jayaram, and Mark Leisher. 2004. Corpus linguistics and South Asian languages: Corpus creation and tool development. *Literary and Linguistic Computing* 19(4): 509–524.

Becker, Dara, and Kashif Riaz. 2002. A study in Urdu corpus construction. In: *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization at the 19th International Conference on Computational Linguistics. August 2002*. http://dl.acm.org/citation.cfm?doid=1118759.1118760 (Accessed 8 Dec. 2014).

Dash, Niladri Sekhar. 2003. Corpus linguistics in India: Present scenario and future direction. *Indian Linguistics* 64(1–2): 85–113.

Selected References – 2

Hardie, Andrew, Paul Baker, Tony McEnery, and B. D. Jayaram. 2006. Corpus-building for South Asian languages. In: Saxena & Borin (eds.) 2006: 211–241.

Hock, Hans Henrich, and Elena Bashir. 2016. *Languages and linguistics of South Asia*. DeGruyter Mouton: Berlin.

Humayoun, Muhammad, and Aarne Ranta. 2010. Developing Punjabi morphology, corpus and lexicon. In: *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, 163–172. http://www.aclweb.org/anthology-new/Y/Y10/Y10-1020.pdf (accessed 8 Dec. 2014).

Hussain, Sarmad. 2003. Computational linguistics (CL) in Pakistan: Issues and proposals. In: *Proceedings of EACL 2003 (Workshop in Computational Linguistics for Languages of South Asia), Hungary, 2003*. http://www.cle.org.pk/research/papers.htm (accessed 8 Dec. 2014).
Hussain, Sarmad. 2013. Developing standards and linguistic resources for computational research in Pakistani languages. http://cle.org.pk/linguistic%20Resources-RIU.pdf (accessed 8 Dec. 2014).

Selected References – 3

Ijaz, Madiha, and Sarmad Hussain. 2007. Corpus based Urdu lexicon development. In: *The proceedings of Conference on Language Technology (CLT07)*. Peshawar: University of Peshawar, Department of Computer Science. http://www.cle.org.pk/research/papers.htm (accessed 8 Dec. 2014).

Khan, Mohammad Abid, and Fatima Tuz Zuhra. 2007. A general-purpose monitor corpus of written Pashto. In: *Conference on Corpus Linguistics, Birmingham, 2007*. http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2007/249Paper.pdf (accessed 8 Dec. 2014).

Rahman, Mutee ur. 2010. Towards Sindhi corpus construction. *Linguistics and Literature Review* 1(1): 74–85. Lahore: University of Management and Technology. http://www.cle.org.pk/clt10/papers/Towards%20Sindhi%20Corpus%20Construction.pdf (accessed18 Dec. 2014).

Selected References – 4

Virk, Shafqat Mumtaz. 2013. *Computational linguistics resources for Indo-Iranian languages.* Centre for Language Technology, Gothenburg, PhD dissertation. http://www.cle.org.pk/Publication/theses/2013/shafqat-phd-thesis.pdf (accessed 18 Dec. 2014).

Yadava, Yogendra P., Andrew Hardie, Ram Raj Lohani, Bhim N. Regmi, Srishtee Gurung, Amar Gurung, Tony McEnery, Jens Allwood, and Pat Hall. 2008. Construction and annotation of a corpus of contemporary Nepali. *Corpora* 3(2): 213–225.