LDC-IL: The Indian repository of resources for language technology

Narayan Choudhary¹

Accepted: 10 December 2020 / Published online: 3 January 2021
© The Author(s) 2021

Abstract This paper introduces the Government of India Initiative on linguistic data creation in Indian languages. The Linguistic Data Consortium for Indian Languages (LDC-IL) is a fully funded Government of India scheme established in 2007 to cater to the needs of linguistic resources required for the development of language technology in Indian languages. LDC-IL worked silently for more than a decade with a team of around regular fifty people and involving thousands of resource persons, covering twenty major languages of India. Part of the output of LDC-IL was launched in April 2019 by the Hon'ble Vice President of India. This paper, the first introductory paper in an academic journal, aims to give a brief of the works done by LDC-IL and how these works are crucial in the development of language technology for Indian languages. Within a short span of eight months of its release, the language resources released by LDC-IL have been procured and utilized by scores of industry and academic bodies and individual researchers, including major industry leaders like Microsoft, Google, Samsung etc.

Keywords Corpora · Indian languages · Language resources · Data distribution · Linguistic repository

---

✉ Narayan Choudhary
choudharynarayan@gmail.com

¹ Central Institute of Indian Languages, Manasgangotri, Mysore, Karnataka, India

1 Introduction

Language Technology is the need of the hour in a fast-moving society. As per Ethnologue,¹ the world at present has approximately seven thousand languages spoken in the world, and roughly 40% of them are endangered because of various reasons. While there might be various reasons for their endangerment, one of the latest reasons among them is lack of information/communication technology support in a language.

According to Ethnologue, India is the fourth most linguistically diverse country in the world: home to about 453 languages. However, if we consider the 1961 census data (Mitra 1964) of the Government of India, the number of “mother tongues” in India stands at 1652. As per 2011 Census report (Census of India 2011) of the Government of India, India has 121 languages that are being spoken by more than ten thousand people. This does not include the mother tongues/varieties that are often grouped under these 121 languages. For example, under Hindi, 55 other mother tongues (e.g. Bhojpuri, Marwari, Haryanvi etc.) are grouped and many of them are significantly different from what we consider as the standard Hindi, so much that the differences are enough to break any linguistic model trained to serve the standard Hindi. This has a greater bearing in technology development scenario and these varieties need to be covered separately to provide support for them.

Despite being home to 1300 million people (close to 18% of the world population), the information technology support in Indian languages has been lagging by decades compared to other languages of the world (e.g. English, Japanese, Russian etc.). This is so because of several factors including fewer amount of e-content/language resources required for the development of such technology.

The Central Institute of Indian Languages (CIIL), a subordinate office of the Department of Higher Education, Ministry of Human Resource and Development, Government of India, got sensitized about this situation and thus came the Linguistic Data Consortium for Indian Languages (LDC-IL) into existence. LDC-IL is an initiative of the Government of India established in 2007 to address such needs of the language technology development community (but also common linguistic research community) and thereby also promoting the Indian languages and research, as per its mandate.

LDC-IL works under the guidance of a Project Advisory Committee that constitutes heads and representatives of various Indian Institutions working in the area of language technology development (e.g. IITs, IIITs, IISc, Central and State Universities etc.) as well as leading industrial/commercial entities (e.g. Google, Microsoft, Intel etc.) working on language technology development. The consortium takes inspiration from similar organizations elsewhere, primarily the Linguistic Data Consortium (LDC) at the University of Pennsylvania² and European Language Resource Association (ELRA), Luxembourg.³

¹ <https://www.ethnologue.com/guides/how-many-languages>, accessed December 23, 2019.

² <https://www ldc upenn edu/>.

³ <http://www.elra.info/>.

Table 1 Scheduled languages of India (in descending order of speaker strength—2011)

SL	Language	Number of speakers (All India)	% of Total population
1	Hindi	528,347,193	43.63
2	Bengali	97,237,669	8.03
3	Marathi	83,026,680	6.86
4	Telugu	81,127,740	6.7
5	Tamil	69,026,881	5.7
6	Gujarati	55,492,554	4.58
7	Urdu	50,772,631	4.19
8	Kannada	43,706,512	3.61
9	Odia	37,521,324	3.1
10	Malayalam	34,838,819	2.88
11	Punjabi	33,124,726	2.74
12	Assamese	15,311,351	1.26
13	Maithili	13,583,464	1.12
14	Santali	7,368,192	0.61
15	Kashmiri	6,797,587	0.56
16	Nepali	2,926,168	0.24
17	Sindhi	2,772,264	0.23
18	Dogri	2,596,767	0.21
19	Konkani	2,256,502	0.19
20	Manipuri	1,761,079	0.15
21	Bodo	1,482,929	0.12
22	Sanskrit	24,821	Negligible

LDC-IL started with a focus on 22 scheduled languages of India (languages that are included in the 8th schedule of the Constitution of India). According to the Census of India (2011) report, these scheduled languages are mother tongues of a total of 96.71% of the total population of India. The table below shows these scheduled languages along with the number its speakers as per Census of India report, 2011 (Table 1).

Since almost a decade of its existence, the LDC-IL has created large to medium sized resources for Indian languages that are essential for any kind of research and development works in language technology, statistical analysis and various other kinds of linguistic studies on these languages. Till now, LDC-IL has developed and released datasets in 20 scheduled languages of India (except Sanskrit and Sindhi, where finding the right human resource has been an issue). These resources are the largest resources available in the public domain so far in each of these languages and we hope this will usher in a new era of linguistic and language technology research in Indian languages.

The rest of this paper is arranged as follows: Other Similar Initiatives Across World, Types of Resources Covered, Methodology of Resource Creation, Datasets

Released, Datasets under Preparation, Distribution and Distribution Mechanism, Costing and Memberships, Collaborations, LDC-IL portal as a Data Distribution Platform, Future Goals and Conclusion.

2 Other similar initiatives across the world

The first among a similar initiative is that of the Linguistic Data Consortium run under the aegis of the University of Pennsylvania, based on the inspiration of which, the LDC-IL was established in India. This initiative came into being in 1992, with support from the Government of the United States, to focus mainly on the requirements of the English language. Later, the LDC became more independent and delved into several languages across the world and at present is probably the largest repository of linguistic resources catering to a wide range of academic and commercial research organizations as well as individual researchers.

Soon after the establishment of LDC, the European Language Resources Association came into being in 1995, focusing first on the European languages and, later on, endeavouring into cataloguing other language resources as well (Choukri et al. 2016; Heuvel and Choukri 2017).

LDC and ELRA are two major initiatives in this direction that are most visible. Given the requirement of huge data, there are also a few start-ups that have ventured into developing and selling such datasets in different languages.

In India, the LDC-IL initiative is the first one to have been established with constant regular support from the Government of India. It is worthwhile to note here that the Government of India has been investing in different types of language resource (LRs) creation since 1990s, but none of the resulting language resources has so far been available for commercial research, for want of a centralized data distribution system and a clear pricing policy. Some works undertaken though were being distributed sporadically by the respective developing agencies (such as CDAC). The Technology Development Programme for Indian Languages (TDIL) under the Ministry of Electronics and Information Technology started its own distribution and cataloguing portal at www.tdil-dc.in which mostly catalogued the datasets and distributed only for non-commercial research purposes to Indian organizations/academics.

The LDC-IL data distribution portal hosted at <https://data.ldcil.org>, is the first easy to use data distribution portal in India where all the datasets developed so far by the LDC-IL has been catalogued and is available for order/request for both the non-commercial and commercial categories of users.

3 Architecture of the data distribution portal

The data portal is designed on an open-source distribution of OpenCart that runs on Apache, PHP and MySQL. Necessary modifications have been done to suit the needs of a data portal. The process is not fully automatic in the way that one cannot make payments automatically as the request for a dataset needs to be first approved

offline before which a payment is accepted. While the registration for commercial users is self-approved, the registration requests for a non-commercial user requires a manual verification of the documents submitted before which the non-commercial registration is approved. A registered user can access the sample data, documentation of the datasets and other details of the dataset. The datasets cannot be downloaded automatically as each dataset request requires manual approval of the competent authority in the organization before the same is made available for download through a different distribution system.

4 Types of resources covered

The LDC-IL works on several types of linguistic resources that include text corpora, annotation of the text corpora, speech corpora, annotated speech corpora (at the sentence level and word level), curated lexica (generated out of the text corpora and vetted), dictionaries (procured from other internal and external sources), image corpora for OCR in Indian languages (as a by-product from another project, called Bharatavani⁴).

These are resources from within LDC-IL or other units of CIIL, the parent institution of LDC-IL. LDC-IL has also planned to catalogue and distribute the language resources developed by other agencies (mostly government aided agencies including universities, institutes etc.) that may or may not have been developed with some aid from the government and host them on the LDC-IL data distribution portal.

5 Methodology of resource creation

Creating a linguistic resource in all the scheduled languages of India poses several challenges. At the time when this project formally started in 2007, e-content were rarely produced in Indian languages. The situation was grim even for the third most spoken language of the world i.e., Hindi such that there was not enough content even on the internet or elsewhere which could be readily used for creating even the raw text corpora. Other languages simply did not have any electronic content.

5.1 Collecting the raw text corpus

Therefore, creating electronic content was a big challenge. Different sets of language resources had their own unique set of challenges. The challenges in creating the raw text corpora and that of raw speech corpora and how they were tackled has been well described in Choudhary and Ramamoorthy (2019) and Choudhary et al. (2019), respectively. As there was no electronic text readily available, the language resource persons for each language were sent to respective language-speaking regions and asked to collect samples of text in all possible

---

⁴ www.bharatavani.in.

domains. This exercise resulted in a collection of hard printed books, journals, magazines etc. all published after 1990 (to ensure that the language is contemporary). These books/extracts/excerpts were brought to Mysore, the headquarters of CIIL wherein the text were typed with the help of a large team of data inputters, vetted by the respective language experts and then finalized and archived to be part of the respective language corpus. The metadata about each of the samples have been meticulously maintained and is part of the language corpus. The metadata give information about the source of the text, its author, publication year, domain and sub-domain. The information given in the metadata makes the corpus even more useful for various types of analysis as it is feasible to tweak the corpus to design a domain-specific or more generalized corpus out of the bigger chunk.

Even though the texts were collected in extracts, this had created a hurdle in releasing the data in time as the copyrights were not sought prior to including them as part of the respective corpora. Communication with the legal advisors of the Government of India clearly stated that even the extracts could not be released without permission from the concerned copyright holders. This forced LDC-IL to seek explicit permissions from the respective copyright holders. LDC-IL sent out more than 10 thousand letters to the respective copyright owners to seek their permissions as required by the Indian Copyright Act. Most copyright holders happily gave their permissions with the exceptions of a few extracts that needed to be excluded from the corpora. LDC-IL being a government organization helped a lot in building the trust with the general public as some stakeholders were more enthusiastic and requested us to use the whole of their works for this purpose. There were also a few stakeholders who did not respond (due to various reasons, such as the address not being the same as mentioned in the publication, missing contact details, copyright holders having no legal heirs etc.). In such cases, we have construed the silence as permission granted. In any case, if we ever get a notice of non-consent, the corresponding extracts will be removed from the corpus in subsequent distributions with adjustments elsewhere as necessary.

5.2 Collecting the raw speech corpus

The speech corpus prepared and released also had a similar issue. While at present it is rather easy to collect speech samples from remote corners of the world using the internet and mobile apps, LDC-IL had to undertake fieldworks to collect speech data. The details of the methodology of collecting the speech data has been given in Choudhary et al. (2019). To sum up the methodology, it was a controlled data collection method wherein a defined, concise set of text transcripts were pre-created to cover different types of linguistic aspects of a language, ensuring that it covers all the sounds, grammatical/syntactic structures of the language. Additionally, to ensure a unique set of speech from each of the speakers, a unique piece of the transcript taken from different sources (mostly contemporary news articles) was also given to each of the speakers to read and the same was recorded in a natural, quiet environment. Each language had a target of six hundred speakers, giving equal weight to male and female, across three standard age-groups. It was also ensured

that speech varieties were covered by taking the samples from different zones of the speech community.

Each of the datasets comes with documentation wherein the methodologies of the data collection procedure have been discussed in detail. The same is also published in book format titled *Linguistic Resources for AI/NLP in Indian Languages* (Choudhary (ed.) 2019). This book, available online at no cost on the LDC-IL portal,⁵ documents the 31 datasets so far released by LDC-IL.

5.3 Creating the parts of speech annotated corpora

The parts of speech (PoS) corpora have been created out of a sub-set of the raw text corpora of the respective languages. The PoS tagset applied is drawn from the Bureau of Indian Standards (BIS) draft tagset (as approved and circulated for four languages, including Hindi⁶) as being discussed in the meetings of the BIS committee formed to look into this issue. The same tagset has been applied in various PoS annotated corpora developed for other languages. For example, the same tagset has been applied to develop the PoS annotated corpora for 12 languages undertaken under the ILCI project (Choudhary and Jha 2011; Jha 2010).

Though the PoS annotated corpora for several languages, with various sizes, have been developed, they have not been released yet. The reason being that none of these datasets have not undergone the round-robin method of vetting and arbitration. This work is currently on and we expect to release the datasets very soon, one by one as it gets completed. The table below gives a rough summary of the raw text corpora (with WC i.e. word counts), the parts of speech annotated text corpora and the raw speech corpora prepared so far (Table 2).

5.4 Creating the sentence aligned corpora

The sentence aligned corpora are a subset of the raw text corpora wherein the speech files have been transcribed, sentence aligned and irrelevant parts (e.g. noise, coughing, long pauses in between sentences etc.) in the speech files have been removed (or marked as such that can be automatically removed). Other general principles of speech data annotation/transcription for the purposes of Automatic Speech Recognition (ASR) have been followed, for example, numerals are always spelt out, speech disfluency, cut-offs are marked. It is ensured that no utterance (considered an equivalent sentence) is of more than 30 s in length. The transcriptions are given both in the script of the language as well as in Roman transliteration using the LDC-IL transliteration schema.⁷ If it is observed that the speech does not follow the given text, it is ensured that the transcription is provided in consonance with what is spoken by the speaker. This leads sometimes to

⁵ https://data.ldcil.org/upload/pubs/LDCIL_Release_Documentation.pdf.

⁶ <http://tdil-dc.in/tdil-dcMain/articles/134692Draft%20POS%20Tag%20standard.pdf>, accessed December 24, 2019.

⁷ <http://www.ldcil.org/download/TranslationStandards.pdf>.

Table 2 Vital stats of LDC-IL corpora developed so far

SL	Language	RawText (WC)	Parts of speech (WC)	Speech data particulars			
				Speech duration	Speakers (speech)	Speech segments	Speech (WC)
1	Assamese	12,433,551	161,570	–	–	–	–
2	Bengali	11,066,289	367,093	128:46:59	476	73,470	544,439
3	Bodo	2,878,208	290,025	176:53:28	456	77,443	465,737
4	Dogri	1,070,102	–	–	–	–	–
5	Gujarati	5,927,264	661,272	–	–	–	–
6	Hindi	37,440,276	663,823	118:40:03	489	73,695	744,242
7	Kannada	12,005,237	651,741	179:32:52	656	99,109	685,731
8	Kashmiri	1,059,596	92,737	–	–	–	–
9	Konkani	3,995,616	284,430	156:37:51	504	72,938	559,495
10	Maithili	5,542,238	371,989	72:02:12	300	35,109	306,311
11	Malayalam	8,507,084	1,079,637	164:01:02	458	43,670	673,348
12	Manipuri	6,537,256	499,581	156:28:32	620	66,231	709,261
13	Marathi	4,729,821	–	89:17:25	307	58,544	339,261
14	Nepali	7,359,314	213,715	87:14:44	350	48,975	286,618
15	Odia	4,103,781	630,504	138:11:46	474	73,418	547,590
16	Punjabi	12,106,648	1,474,667	101:09:28	467	76,230	545,420
17	Santali	76,546	–	–	–	–	–
18	Tamil	13,947,875	1,548,133	129:20:26	450	60,291	510,871
19	Telugu	6,549,493	30,118	22:43:59	80	10,510	135,151
20	Urdu	5,159,617	1,664,745	99:18:21	499	88,708	577,693
	Total	162,495,812	10,685,780	1820:19:08	6586	958,341	7,631,168

misspellings also. However, to mitigate it, standard pronunciation is also provided that would be useful in developing ASR models.

5.5 Other language resources

Apart from the above four resources, LDC-IL is also planning to release other resources that include standard lexicon in each of the languages which will be useful in tools such as spelling correction and proofing tools. Over the years, CIIL and several other government agencies have published several monolingual, bilingual and multilingual dictionaries. These dictionaries may be useful in various NLP tasks as well. LDC-IL is working on a model to create digital dictionaries out of these works and release them on LDC-IL portal as additional electronic resources. Similarly, the Bharatavani project is engaged on a mass scale to produce electronic text, mostly in the knowledge text domains (i.e. non-fiction text), in 121 languages of India. The effort here is to identify the existing knowledge text in these languages and then digitize them. The method of digitization includes scanning of these texts (often old printed materials published by universities, government bodies, public and private academic bodies across India, including non-governmental organizations) and then getting them typed with the help of DTP operators/vendors across the country. This exercise creates a by-product of a huge amount of image corpus (with a minimum of 300dpi, as required in any OCR training algorithm) and the respective vetted text against it for each page. It is proposed to create an image corpus out of this for 121 languages. These image corpora can be used to train specific OCR models for these 121 languages. The Bharatavani portal at present has hosted knowledge text in close to 100 languages of India that are spoken by more than ten thousand people in India.

6 Datasets released and soon to be released

After almost a decade of work and surpassing several challenges, this huge resource of LDC-IL was recently released by the Hon'ble Vice President of India in April 2019. A total of 31 datasets have been released on the LDC-IL data portal. This includes gold standard raw text corpora in 18 languages and gold standard raw speech corpora in 13 languages.

Additional, 30 corpora in the four categories mentioned above are at various stages of preparation and will be released soon.

7 Data distribution and distribution mechanism

Taking a cue from the LDC and ELRA, the data distribution has been made easy with the portal allowing users to register in two categories of commercial users and non-commercial users. The non-commercial user category is meant for the academic and non-commercial research organizations within India. Thus, any academic, researcher or student in an Indian university can get the data at no cost for

non-commercial research purposes with the condition that any product coming out of it will not be commercialized for non-public gains.

The commercial user category of registration allows commercial entities, including any individual and non-Indian entity (including non-Indian individual/researcher/student or non-profit organizations) to register and raise the request for the datasets.

The idea of limiting free access only to the Indian academic/non-commercial use may sound rather “not so global” at present. However, this is mandatory as the LDC-IL is a project/scheme of the Government of India, run under the policy of promotion of Indian languages. However, it is possible that, at a later point of time, the resources may be made available for non-commercial uses across the globe on the same terms as applied to the non-commercial entities within India.

LDC-IL has also started talks with the LDC and ELRA to allow cross-listing the datasets on their respective portals. At present, while the cross-listing on other data distribution portals is possible, the data itself cannot be hosted on any portal apart from the ones designated for this purpose within India.

8 Costing and licensing

Putting a price to any dataset developed over the years was one of the biggest challenges faced by LDC-IL. There were several factors involved which affected the price calculation. The Project Advisory Committee of the LDC-IL had set up a sub-committee to look into the pricing of the datasets. However, despite at least three meetings of the sub-committee, no price calculation mechanism could evolve. There were proposals that the total investment made so far in LDC-IL (close to 200 million Indian rupees i.e. around 3 million USD) should be targeted to be recovered from the sale proceeds. However, it was found that it would make the price of the datasets too high to start with.

A proposal was put up by Rajeev Sangal, chairperson of a TDIL standing committee on language technology. He suggested that the guiding principle should be as per the current cost of developing such a language resource. This meant that there might be some tasks or sub-tasks that would be costly, say, 10 years ago, but the same may be cheaper today (e.g. collecting speech samples 10 years ago would be a difficult task, involving a visit to the respective speech areas while the same might be easier today with the use of remote communication technologies).

With the above as the guiding principle, an effort was made to count the number of processes/steps involved in each of the types of the LR creation and calculate the unit cost of it. This resulted in formulae developed for doing a cost analysis for each of the LR types. The present author was assigned to develop a policy document in consultation with other stakeholders in this area to develop formulae for different types of LRs. After incorporating the public comments on the draft of this policy document, the final costing document was published as *Cost Analysis of Linguistic Resources* (Choudhary 3). A sample formula to calculate the cost of a raw text corpus is given in the table below (Table 3):

Table 3 Formula to calculate the cost of a raw text corpus

SL	Particulars	Values
1	Resource type	Text
2	Corpus type	Raw Corpus
3	Language	XX
4	Data source	Typed/Typed + Cleaned/Crawled/Crawled + Cleaned
5	Unit	Character/Key Stroke
6	DTP cost per unit	(INR) 0.007
7	Proofing cost per unit	(INR) 0.004
8	Actual cost of a corpus	
8.1	Typed corpus	= (number of characters*Data Entry cost per character)
8.2	Typed + cleaned corpus	= (number of characters*Data Entry cost per character) + number of characters*proofing cost per character
8.3	Crawled corpus	= (To be updated)
8.4	Crawled + cleaned corpus	= (To be updated)
9	Overall cost of a corpus	= Actual Cost + 10% of actual Actual Cost

The base prices of all the datasets of the LDC-IL currently released (and to be released in future) are based on the formulae given in this cost analysis document. This is a public document, open for feedback and is likely to change with time.

While the datasets are free for non-commercial purposes within India, the prices for the commercial use is kept affordable to attract all the types of commercial entities, including start-ups, medium scale enterprises and multi-national firms. The base price itself is 10% of the total cost incurred in developing such an LR at present. The startups, MSMEs and entities from the SAARC nations (the neighbouring countries of India which share the same languages) get additional discounts on the base price.

9 Future goals

At present, LDC-IL is a wholly owned Government of India (GoI) programme, with the total expense being borne by the GoI. It was established with the vision that after 6 years of its operation, the programme will start generating its own funds and can stand on its own. Due to delay in its data being released, the same was not possible. However, with the release of its datasets in April 2019, it has started generating revenues. These funds at present go to the Government of India. However, the Consortium can now stand on its own. The proposal to make LDC-IL an autonomous body, with minimum support from the Government of India, is underway.

In near future, LDC-IL may also host LRs developed by other organizations within and outside India. To begin with, LDC-IL is already in talks with various organizations and individuals who have agreed in principle to host their datasets on the LDC-IL data portal.

10 Conclusion

Linguistic Data Consortium for Indian Languages (LDC-IL) aims to serve the language technology development community by providing quality data to work with. The datasets are authenticated by following the standard protocols. While researchers within India can take the benefit of these datasets, the datasets are available to the Industry and others at a reduced price. LDC-IL aims to become a repository of linguistic resources with a focus primarily on Indian languages. LDC-IL also aims to collaborate with other organizations of the world serving these types of datasets by way of proliferating these datasets in their catalogues.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain

permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243–257.
- Census of India. (2011). Paper 1 of 2018: Language (Table C-16). Office of Registrar General, Government of India. Retrieved from https://censusindia.gov.in/2011Census/C-16_25062018_NEW.pdf.
- Choudhary, N. (2018). *Cost analysis of linguistic resources*. Mysore: Central Institute of Indian Languages.
- Choudhary, N. (Ed.). (2019). *Linguistic resources for AI/NLP in Indian languages*. Mysore: Central Institute of Indian Languages.
- Choudhary, N. & Girish, N. J. (2011). Creating multilingual parallel corpora in Indian languages. In: *Proceedings of the 5th Language & Technology Conference*. Poznan, Poland. https://doi.org/10.1007/978-3-319-08958-4_43.
- Choudhary, N., Rajesha, N., Manasa, G., & Ramamoorthy, L. (2019). LDC-IL raw speech corpora: An overview. *Linguistic resources for AI/NLP in Indian languages* (pp. 160–174). Mysore: Central Institute of Indian Languages.
- Choudhary, N., & Ramamoorthy, L. (2019). LDC-IL raw text corpora: An overview. *Linguistic resources for AI/NLP in Indian languages* (pp. 1–10). Mysore: Central Institute of Indian Languages.
- Choukri, K., Mapelli, V., Mazo, H., & Popescu, V. (2016). ELRA activities and services. In *Proceedings of the 10th International Conference on Resources and Evaluation*, May 23–28, 2016, Portorož, Slovenia. Portorož: ELRA, pp. 463–468. Retrieved from <https://www.aclweb.org/anthology/L16-1074>.
- DiPersio, D. & Christopher, C. (2019). The linguistic data consortium: Developing and distributing language resources4All. In *Proceedings of LT4All: International Conference on Language Technologies for All: Enabling Linguistic Diversity and Multilingualism Worldwide*. UNESCO Headquarters, Paris. Retrieved from <https://www ldc.upenn.edu/sites/www ldc.upenn.edu/files/lt4all-2019-ldc-language-resources-4all.pdf>.
- Jha, G. N. (2010). The TDIL program and the Indian Language Corpora Initiative (ILCI). In: *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*. May 17–23, Valletta, Malta. Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/874_Paper.pdf.
- Mitra, A. (1964). Census of India 1961, vol. I, part II - C(ii): Language tables. Delhi: Manager of Publications, Government of India Press.
- Henk van den, H., & Khalid, C. (2017). ELRA's central role in managing language resources in Europe. In: *Proceedings of the 15th EFNIL Conference*, October 5–6, 2017, Mannheim, Germany. pp. 71–81. Retrieved from <http://www.efnil.org/documents/conference-publications/mannheim-2017/EFNIL-Mannheim-13-vandenHeuvel-Choukri.pdf>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.