# GURMUKHI PUNJABI (PA) AS A LOW-RESOURCE LANGUAGE THROUGH THE LENS OF THE BLARK MODEL

**KIRANDEEP KAUR**
**Bachelor of Technology, Punjab Technology University, 2019**

A thesis submitted
in partial fulfilment of the requirements for the degree of

**MASTER OF ARTS**

in

**INDIVIDUALIZED MULTIDISCIPLINARY**

Department of English
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

© Kirandeep Kaur, 2023

GURMUKHI PUNJABI (PA) AS A LOW-RESOURCE LANGUAGE THROUGH THE LENS OF THE BLARK MODEL

KIRANDEEP KAUR

Date of Defense: August 23, 2023

| | | |
|---|---|---|
| Dr. D. P. O'Donnell | Professor | Ph.D. |
| Dr. Conor Snoek | Assistant Professor | Ph.D. |
| Thesis Co-Supervisors | | |
| Dr. Yllias Chali | Professor | Ph.D. |
| Thesis Examination Committee Member | | |
| Dr. B. Bordalejo | Assistant Professor | Ph.D. |
| Thesis Examination Committee Member | | |
| Dr. Antti Arppe | Associate Professor | Ph.D. |
| Thesis Examination Committee Member | | |
| Dr. Tabitha Spagnolo | Associate Professor | Ph.D. |
| Chair, Thesis Examination Committee | | |

# DEDICATION

To my maternal grandmother, Surinder Kaur Randhawa, and my mothers, Kawaljeet and HarinderPal Kaur.

# ABSTRACT

We are venturing into the next phase of digital divide (unequal access to digital technology), where the languages which are not ready for Natural Language Processing (NLP) are at the most risk of losing out on the developments in the fields of Speech and Language technologies. This has brought forth a big gap between the readiness of different languages in terms of taking advantage of the recent developments in the field of computational technologies.

Common Language Resources and Technology Infrastructure (CLARIN) - a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily usable, has developed the Basic Language Resource Kit (BLARK) model to assess the readiness for speech and language technology developments in any language.

Punjabi, despite being a major language with millions of native speakers and a significant diaspora population around the world, has received limited attention in the computational technologies. The thesis aims to provide a comprehensive overview of the existing resources, tools, and techniques for Punjabi NLP, as well as to identify the gaps and opportunities for future research using BLARK model as a framework. The thesis, after giving the current (sorry) state of Punjabi in terms of its readiness for computation technologies, concludes with some suggestions for directions and effort which are needed for making Punjabi ready for development of speech and language technologies. The thesis contributes to the field of Punjabi language processing by proposing a generic model for comparing and enhancing Punjabi linguistic resources.

# ACKNOWLEDGEMENTS

I'd like to thank my supervisor, Professor Daniel Paul O'Donnell for his significant support and his insights that helped me write this work. He is a very kind and generous person, and I am grateful for his ongoing mentorship and never-ending support. His humble approach to Digital Humanities is exemplary. This approach is evident in his simple but obvious writing style, which I aspire to emulate throughout my career. He helped me in every possible way to settle down in Canada by providing me with Graduate Assistantships in the Computer Science department and rendered me every possible help that he could to give my references for the software development projects. Dr. O'Donnell used to meet me weekly or sometimes twice in a week to address my issues and to give me valuable feedback for the work done by me. He always boosted my morale by giving me positive assessments and he always motivated and inspired me to carry forward with my work whenever I feel stuck at some point. His relentless support helped me to refine my writing skills.

My heartfelt gratitude to my co-supervisor, Dr. Conor Snoek, for his patience and understanding during the two years it took to complete this work. He assisted and guided me in every possible way. Coming from a Computer Science background, it was sometimes hard for me to play around with linguistics terms and terminologies, but he would very gently and calmly explain everything to me in detail. His feedback and recognition mean a lot to me. I appreciate the time and efforts he undertook to prepare me to write my thesis.

Also, I am deeply thankful to Dr. Antti Arppe who helped with the proposition of the thesis and suggested that I should work with the Punjabi BLARK model. His comments on the thesis proposal helped me to narrow down the scope of my thesis. He suggested some useful research articles relevant to the area of my research. Dr. Barbara Bordalejo is always responsive to my thesis presentations, her counter questions on my thesis presentation helped me to make my points more

precise and transparent. Dr. Yllias Chali has always been helpful, he taught me proper guidelines to evaluate assignments in my Teacher Assistantship work. He always motivated me to carry forward my research in the right direction. I would also like to thank my peers AKM Ifthekhar Khalid and Davide Pafumi who have been morally supporting me all the way during my academic years. Sincere thanks to Gurpreet Singh Saini who inspired me to work on my mother tongue, Punjabi.

Last but not the least, a great thanks to my Nanni, my maternal grandmother for all the sacrifices she made to fulfill my dream of pursuing master's education in my dream country, Canada. I am always indebted to all that she has done for me.

I am grateful to everyone who has supported me throughout this process. Without your help and guidance, this thesis would not have been possible.

# TABLE OF CONTENTS

## Table of Contents

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| ANLT | Alvey Natural Language Tools |
| BLARK | Basic LAnguage Resource Kit |
| BNC | British National Corpus |
| CALL | Computer Assisted Language Learning |
| CIIL | Central Institute of Indian Languages |
| CLARIN | Common Language Resources and Technology Infrastructure |
| CLAWS | Constituent Likelihood Automatic Word-tagging System |
| DARPA | Defense Advanced Research Projects Agency |
| ECI | European Corpus Initiative |
| ELRA | European Language Resources Association |
| ELSNET | European Network of Excellence in Language and Speech |
| EMILLE | Enabling Minority Language Engineering |
| HLT | Human Language Technologies |
| IBM | International Business Technologies |
| IGT | Interlinear Glossed Text |
| LDC | Language Data Consortium |
| LR | Language Resource |
| NEMLAR | Network for Euro-Mediterranean LAnguage Resource |
| NLP | Natural Language Processing |
| NSF | National Science Foundation |
| OCR | Optical character Recognition |
| ODIN | Online Database of Interlinear Text |
| OLAC | Open Language Archive Communities |
| POS | Part-Of-Speech |
| TLG | Thesaurus Linguae Graecae |
| WALS | World Atlas of Language Structures |

# CHAPTER 1: INTRODUCTION

This chapter will give a brief introduction to the research work presented in this thesis. It will begin by outlining the background to the research questions addressed and offering a short justification as to why these questions are important. The questions addressed will then be presented and a list of the contributions arising from the investigations will be given. Finally, the structure of the remainder of the thesis will be presented.

## 1.1 Background to the Research

The thesis work will mainly focus on the analysis of the resources that are already available and the tools and algorithms required for Gurmukhi Punjabi. The aim of this study is to provide insights into the various linguistic resources, tools and state-of-the-art techniques applied to the processing of Gurmukhi Punjabi.

Punjabi is written in two different scripts - Gurmukhi and Shahmukhi, which I have explained in detail later in chapter 2. It has millions of native speakers, the language is spoken among a significant overseas population, particularly in Canada, the United States, and the United Kingdom. Gurmukhi Punjabi, being my mother tongue, my thesis will only focus on it. Despite having such a large diaspora, limited work has been done for Gurmukhi Punjabi in the field of Natural Language Processing (NLP) which includes factors like complex linguistic facts and the inclusion of prevalent dialects. Using Punjabi as an example case, this thesis provides the gap analysis of lack of tools and techniques for Punjabi language processing and also describes and motivates the adaptation of the Basic LAnguage Resource Kit (BLARK) approach for less-resourced languages, based on readily available linguistic resources and how the end-user applications can provide benefit to the language communities. A generic model would be prepared

which would highlight the fundamental tools that are required for Punjabi that could assist to build applications like computer-assisted language learning, access control, speech input, speech output, dialogue systems, document production, information access and translation.

### 1.2 Questions Addressed

Publishing software tools equipped with features related to language processing is an active area of investigation. The thesis investigates the different levels of resource requirement for Punjabi as a low-resource language. Thus, the research questions are:

1. What is Gurmukhi Punjabi?
2. What are Natural Language Processing and Computational Linguistics?
3. What are low resource languages and how can we categorize a language as a low resource?
4. What has been done on Natural Language Processing (NLP) in Punjabi/what resources are available? What remains to be done?
5. How can (4) be done? Why is it important?

Since, I am expanding Arabic and Dutch Basic LAnguage Resource Kit (BLARK) on different grounds, taking Gurmukhi Punjabi as an example, so I am adapting the above questions from the research done for Dutch and Arabic. The first two questions focus on giving background to research, describing what Punjabi is and explaining Natural Language Processing (NLP) in general. The third question is particularly about what low resource languages are and then the fourth question will evaluate the existing and required Gurmukhi Punjabi language resources, tools and applications. The fifth question will define the possible roadmap to how the required things can be achieved.

### 1.3 Importance of this case study

It might seem reasonable to have only a few languages like English or French to do fine in the digital world, so why bother with other languages? However, building Natural Language Processing (NLP) applications for such languages can at the same time reinforce the ties between the world and ensure its diversity (Sciforce 2019). I have mentioned some of the reasons in the following sections.

#### 1.3.1 Preservation

People throughout the world have been using computers and the internet in their own languages. Despite the fact that India produces one-fourth of world engineers (M. Ghosh 2018), still Indian users and Punjabi users, in particular, have been left with the only option which is to use the internet in English. To ameliorate the socioeconomic environment of a country, it inevitably becomes important to support the native languages in technological environments. The need for language-based content and technology has to be addressed. Society at large can benefit from Information Technology effectively if people can communicate with computers in their own languages. Barely 70 % of the Indian population is literate, of which only an elite minority (~10%) can read, write, and speak the English language. This leaves out most of the Indian population to access worldwide and thus, it is essential to have an interface that supports the local language in written and spoken technologies so that it can cater for the needs of illiterate and semi-literate sections of the population. “You lose a language, you lose a culture” (Oregonian/OregonLive 2022), the status of culture and language of Gurmukhi Punjabi is unique and special status is accorded to Punjabi community in India, a status that recognizes distinct cultural, economic and political rights, including the right to continue separate identities and Gurmukhi Punjabi form the basic medium for

the transmission, and thus survival of Gurmukhi Punjabi culture, literature, history, religion, political institutions, and values is important.

### 1.3.2 Software applications

Natural Language Processing (NLP) tools are good to have, and they hold the potential to be useful in language revitalization. Sometimes even languages that were deemed to be extinct come back to thrive. Revival of Hebrew is one of the examples, and with the help of Natural Language Processing (NLP) techniques such revival can be sped up drastically (Oregonian/OregonLive 2022). Revitalization of language can be established through education and accessibility. Availability of language technologies paves the way to practice the language and ensures the continuation of language usage. Language tools are more accessible to young people. Keyboards in mobile phones can be taken as a rudimentary example, if they are available in local languages, it can help the people especially the youth to learn and practice in their own native languages and also, they can exchange text messages or post content on social media in their native language, thus, escalating the usage of the language.

### 1.3.3 Knowledge expansion

Corpus linguistics, which is used as the base for any Human Language Technologies (HLT) application or module, does not tell if the information is correct, incorrect or possible. It usually contains the most common words used in the language or common phrases or what tenses people use most frequently. New developments in Natural Language Processing (NLP) might be helpful in flashing insights by comparing more diverse languages.

### 1.3.4 Monitoring demographic processes

Punjabi has approximately 109 million speakers ("Pakistan Bureau of Statistics" 2017; "Government of India" 2011) and is spoken among a significant overseas diaspora, particularly in the United States, United Kingdom and Canada. Given the large population and geographical territory, Punjabi has a large magnitude of dialectical variations, with speakers often able to determine the city in which a person resides based on their pronunciation and word choice. The worldwide proliferation of digital communications has created the need for language and speech processing systems for Punjabi. With availability of small data sets and large morphological inflexion, Punjabi is facing challenges in developing such systems.

## 1.4 Contribution to the Research

The thesis aims to provide a comprehensive overview of the existing resources, tools, and techniques for Punjabi NLP, as well as to identify the gaps and opportunities for future research. It highlights the Basic Language Resource Kit (BLARK) concept and defines the BLARK approach to the needs of Punjabi language. The BLARK is an approach proposed by Krauwer (2003) and Binnenpoorte et al. (2002) for establishing a roadmap for Human Language Technologies (HLT) for a given language, based on this scheme, gaps will be identified and reported. Strategies for filling the gaps will be explained to the major stakeholders of Gurmukhi Punjabi. The outcome of this task will allow us to make suggestions for missing essential resources for Punjabi language. Also, criteria that allow us to assess the quality of language resources and tools will be established, thus enabling the scholars and researchers to set priorities which also consider the concrete needs of projects for low-resource languages like Punjabi. The BLARK schema and its prototypical application to Punjabi will be presented in a report. The major contributions are as follows:

1. To effectively incorporate different information (such as parallel data or language relatedness) to the BLARK model, along with the annotated data.
2. To analyze many language technology solutions for low-resource scenarios such as annotation mapping and limited resources such as monolingual data, bilingual dictionaries, and speech recordings for Punjabi.
3. To identify the modules and data that are needed to build Natural Language Processing (NLP) applications.
4. To survey work done for Punjabi language processing and to show that scattered work has been done but there is little unification. Also, there is scarcity of linguistic resources and open-source software for Punjabi.

### 1.5 Report Structure

The thesis is structured as follows. Chapter 2 and Chapter 3 lists the background needed to understand the thesis, consisting of a brief introduction to Punjabi language and how Natural Language Processing (NLP) works with the languages.

Chapter 4 focuses on the definition of low resource language and BLARK. It introduces the concept of a unique Basic Language Resource Kit (BLARK) for the Natural Language Processing (NLP) and machine learning and explains in more detail what is meant by the term, and why creation is considered as a necessary element for developing and building Natural Language Processing (NLP) tools. It provides an overview of the creation of the original BLARK and previous approaches to the task of defining such a language resource kit.

Chapter 5 presents the prototypical application of the BLARK schema for Punjabi and gives an overview of the existing Natural Language Processing (NLP) tools for Punjabi. It describes how the

lack of development of a corpus and building of a lexicon for Punjabi language is the one of reasons why Punjabi is an under-resourced language.

Chapter 6 compares language resources and tools available for the individual languages. Indian languages like Hindi, Tamil, Bengali or Marathi are compared with Dutch and Arabic. The chapter sketches out that if building BLARK for Arabic can help to enrich the language, the same can be done with Punjabi too.

# CHAPTER 2: INTRODUCTION TO PUNJABI

**Punjabi** (**Panjabi)** is an Indo-Aryan language belonging to the Indo-European language family's Indo-Iranian subgroup. It has been classified to the CENTRAL group under the INNER sub-branch of Indo-Aryan languages under NIA sub classification scheme (Masica 1991). It is natively spoken by the Punjabi people in Pakistan and India. Punjabi has approximately 109 million native speakers and with Pakistan having 76,629,082 million native speakers and India having 22,124,726 million native speakers ("Pakistan Bureau of Statistics" 2017; "Government of India" 2011). As already mentioned in Chapter 1, there are two main varieties of Punjabi: Eastern Punjabi which is written in Gurmukhi script and Western Punjabi, which is written in Shahmukhi script, each of which has a number of dialects. Eastern Punjabi is spoken mainly in India in the states of Punjab, Chandigarh, Haryana, Himachal Pradesh, Kammu and Kashmir, and Rajasthan. Western Punjabi is spoken mainly in Punjab province in Pakistan, and in Punjab state and Jammu and Kashmir state in India. It is also known as Lahanda, Lahnda or Lahndi (Hussain et al. 2020, 283). The language is spoken among a significant overseas population, particularly in Canada, the United States, and the United Kingdom.

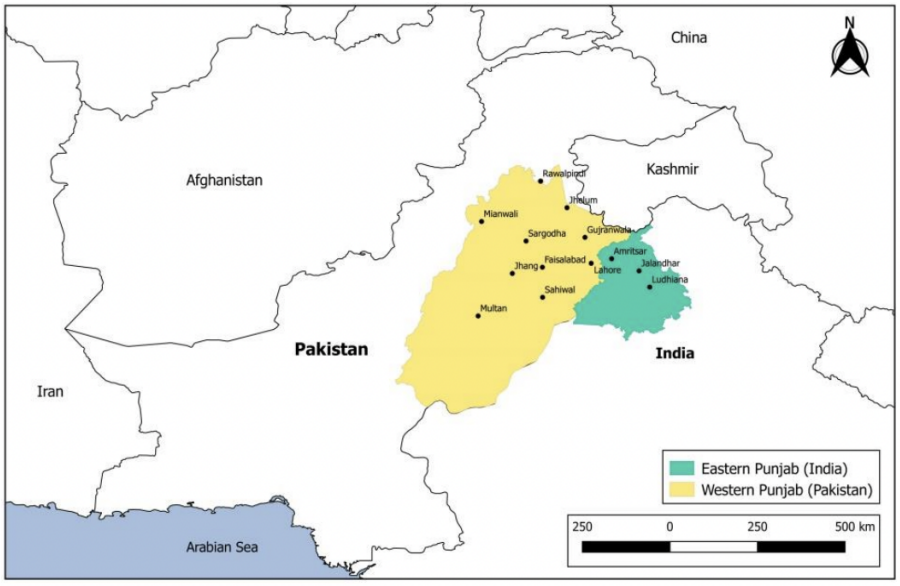


Figure 1: Map of Eastern (India) and Western (Pakistan) Punjab (Hussain 2020, 283)

A full-scale version of this map is on page 126.

The etymology of both names Punjabi and Punjab comes from the Persian words for 'five' ( پنج - panj) and 'water' (آب - ab) and refers to the five major eastern tributaries of the Indus River that flow through Punjab: the Chenab, Jhelum, Ravi, Sutlej and Beas (Ager 2023). Gurmukhi Punjabi is one of India's 22 official languages and it is the first official language in Punjab, India. It is the medium of everyday communication in the Indian state and is used in education, government, business and media. It is the religious language of the Sikhs and popular Bhangra folk dance and singing. In Pakistan, Punjabi is the second most widely spoken language but has no official status (Ager 2023). While in India, Punjabi is widely used in movies and news industries, in Pakistan, it is only used in a few radio and television programs.

As shown in Table 1, in India, Punjabi is written with the Gurmukhi (ਗੁਰਮੁਖੀ) script, while in Pakistan it is written with a version of the Urdu script known as Shahmukhi (شاہ مکھی).

Linguistically written standard for Punjabi in both India and Pakistan is known as Majhi (ਮਾਝੀ/ ماجھ), which is named after the Majha region of Punjab (Ager 2023).

Table 1: Table shows the two different scripts to write Punjabi.

| Script | Description |
|---|---|
| Gurmukhi | Within the Indian state of Punjab, Sikhs tend to use the Gurmukhi script. **Example:** ਹਾਂ ਜੀ, ਦੇਵਨਾਗਰੀ ਵਿਚ ਪੰਜਾਬੀ ਲਿਖੀ ਜਾ ਸਕਦੀ ਹੈ |
| Shahmukhi | Punjabis in Pakistan use a modified Arabic script called Shahmukhi, a modified version of the Persian script. It is written from right to left. **Example:** ہاں جی،دیوناگری وچ پنجابی لکھی جا سکدی ہے |

## 2.1 Gurmukhi script (ਗੁਰਮੁਖੀ)

The Gurmukhi script is used to write the Punjabi language in India Punjab. It is the language of Sikhs in which the holy book of Sikhism, Guru Granth Sahib, is written. Gurmukhi developed from ancient Brahmi script but was standardized by the second Guru of Sikhs, Guru Angad Dev Ji, during the 16th century (1504-1552). According to the Oxford dictionary and many other resources the nomenclature ‘Gurmukhi’ literally means "from the mouth of guru" (“Oxford University Press” 2023) but Bhardwaj (2016) claims that the name Gurmukhi comes from the Punjabi word “gurmukh”, meaning ‘guru-oriented’ or ‘pious’ and now used mostly for a devout Sikh. The word “gurmukh” is older than Sikhism and is found in writings of Hindu saints who predate Guru Nanak Dev, the first guru of Sikhs. Guru Angad Dev did not “invent” this script. All the traditional thirty-five letters and other symbols used in Gurmukhi already existed and are mentioned by Guru Nanak Dev in a hymn in Srī Gurū Granth Sāhib (Bhardwaj 2016, 15). Gurmukhi also refers to its use in

Adi Granth which includes various hymns and compositions of Sikh Gurus of fifteenth, sixteenth and twentieth centuries. Some of the hymns are composed by Sufi Muslim and Hindu saint-poets as well (Gill 1996). As mentioned in the above section, it is written from left to right. The characters are aligned below the line of writing and the words are separated by spaces.

### 2.2 Orthography Features

I am shedding light on orthographic features of Gurmukhi Punjabi in this section. I have taken all the information from Omniglot (Ager 2023) and (Bhardwaj 2016).

#### 2.2.1 Consonants

The Gurmukhi Punjabi belongs to the north-western group of the scripts of the Brahmi family and thus relates to the major Indian scripts such as Devanagari, Bangla, Gujarati, Tamil, Telugu, Kannada and Malyalam as well as to the Thai. It is neither purely alphabetic like the Latin alphabet nor purely syllabic but, is a nice synthesis of the two, known as an alpha-syllabic system. A chart of the Punjabi letters is given below with the name of each letter shown in the Punjabi script itself and in phonetic transcription. The Gurmukhi alphabet contains thirty-five base letters also known as "akkhara", (plural - akkharɑ̃). Gurmukhi is an abugida script, i.e consonants carry an inherent vowel, a sound which is used with each unmarked basic consonant symbol. To exemplify, /ə/ after a consonant is not written, but is it an inherent part of the consonant letter, so /kə/ is written by simply a consonant letter <ਕ>. Certain consonant clusters are written with special conjunct symbols which are used to combine the essential parts of each letter, similar to those of Devanagari (Gill 1996). For example, ka<ਕ> + ga<ਗ> = kàggā<ਘ>; ca<ਚ> + <ਜ>ja = càjjā<ਝ>; ṭa<ਟ>+ ḍa<ਡ>= ṭàḍḍā<ਢ>

Some characters have diacritics, which can appear above, below, before or after the consonant to which they belong (Gill 1996). Some letters have a dot diacritic to represent marginal consonants.

In Table 2, each row in the chart represents a place of articulation and each column a manner of articulation. All the consonants are basically stop consonants i.e the flow of the outgoing breath is stopped in the mouth.These technical terms are explained below.

**a. Place of articulation**

- **Velar -** Back of the tongue touches the palate and outgoing breath is stopped.
- **Palatal -** The front part of the tongue touches the hard palate behind the gum.
- **Retroflex -** The underside of the curled tongue touches the hard palate.
- **Dental -** The tip of the tongue touches the upper teeth.
- **Bilabial -** The upper and lower lip is closed to stop the flow of the air.

**b. Manner of articulation**

- **Voiceless -** The vocal cords do not vibrate.
- **Voiced -** Vocal cords vibrate and create buzzing sound.
- **Unaspirated -** The consonant is released with little or no puff of air.
- **Aspirate -** The flow of air is very strong, and the consonant is released with a strong puff of air.
- **Nasal -** The flow of air is through the nose rather than the mouth.

Table 2: Gurmukhi letters (Gill 1996; Bhardwaj 2016)

| | Vowel-bearers | | | | | | Fricatives | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Name** | **Sound [IPA][1]** | **Name** | **Sound [IPA][1]** | **Name** | **Sound [IPA][1]** | **Name** | **Sound [IPA][1]** | **Name** | **Sound [IPA][1]** |
| | ੳ ūṛā – | | ਅ aiṛā | /a/ | ੲ īṛī – | | ਸ sassā | /sa/ | ਹ hāhā | /ha/ |
| **OCCLUSIVES →** | **Voiceless** | | **Voiced** | | **Unaspirated** | | **Aspirate (Tonal)** | | **Nasals** | |
| **Velar** | ਕ kakkā | /ka/ | ਖ khakkhā | /kha/ | ਗ gaggā | /ga/ | ਘ kàggā | /kà/ | ਙ ṅaṅṅā | /ṅa/ |
| **Palatal** | ਚ caccā | /ca/ | ਛ chacchā | /cha/ | ਜ jajjā | /ja/ | ਝ càjjā | /cà/ | ਞ ñaññā | /ña/ |
| **Retroflex** | ਟ ṭaiṅkā | /ṭa/ | ਠ ṭhaṭṭhā | /ṭha/ | ਡ ḍaḍḍā | /ḍa/ | ਢ ṭàḍḍā | /ṭà/ | ਣ nāṇā | /ṇa/ |
| **Dental** | ਤ tattā | /ta/ | ਥ thatthā | /tha/ | ਦ daddā | /da/ | ਧ tàddā | /tà/ | ਨ nannā | /na/ |
| **Labial** | ਪ pappā | /pa/ | ਫ phapphā | /pha/ | ਬ babbā | /ba/ | ਭ pàbbā | /pà/ | ਮ mam mā | /ma/ |
| **SONARANTS** | | | | | | | | | | |
| Frictionless continuants → | ਯ yayyā | /ya/ | ਰ rārā | /ra/ | ਲ lallā | /la/ | ਵ vāvā | /va/ | ੜ ṛāṛā | /ṛa/ |
| **SUPPLEMENTARY CONSONANTS** | | | | | | | | | | |
| ਲ਼ lalle pair bindī | /ḷa/ | ਸ਼ sasse pair bindī | /śa/ | ਜ਼ jajje pair bindī | /za/ | ਖ਼ khakkhe pair bindī | /xa/ | ਫ਼ phapphe pair bindī | /fa/ | ਗ਼ gagge pair bindī | /ġa/ |

1 The International Phonetic Alphabet (IPA) is an alphabetic system of phonetic notation based primarily on the Latin script ("International Phonetic Alphabet (IPA) Britannica" 2023).

### 2.2.2 Vowel bearer letters

Gurmukhi Punjabi has inherent vowels. Punjabi uses 9 combining marks for the vowels, as seen in the table below, also known as 9 dependent vowel signs. Using these combining marks and three bearer characters: Ura <ੳ> phonetically corresponding to /ʊ/, Aira <ਅ> pronounced as /ə/ and Iri <ੲ> pronounced as /ɪ/, dependent vowels are created as shown in the Table 3. Except for Aira independent vowels are never used without additional vowel signs. Vowels are written as independent letters, when they appear at the beginning of a syllable. The system of the Punjabi vowels is best discussed with the help of the following topological map showing which part of the tongue is raised to what height in order to change the shape of the resonance chamber in the mouth to achieve a distinct acoustic quality for the sound of each vowel.

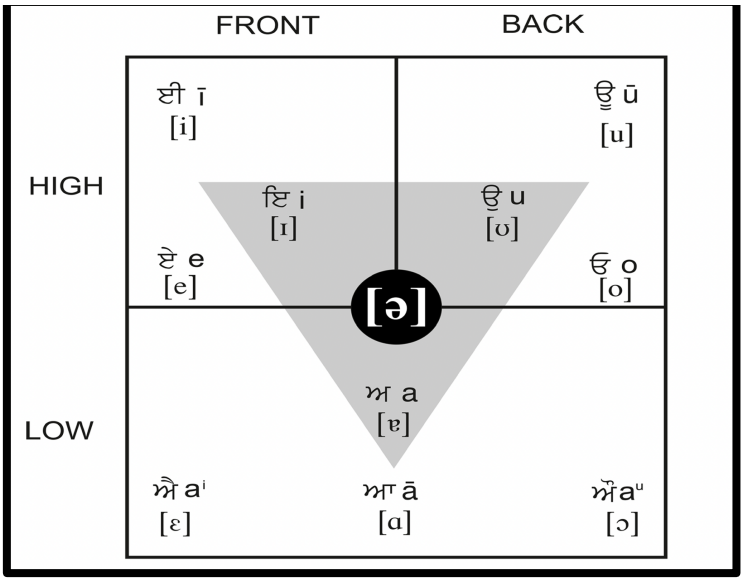


Figure 2: Gurmukhi Punjabi Vowels (Bhardwaj 2016, 49)

The Punjabi vowels can first be divided into two groups – high and low. The high vowels can be further divided into front and back vowels. The high front vowels are written by adding vowel symbols to the vowel bearer letter <ੲ>, the high back vowels are written by adding vowels symbols

to the vowel bearer letter <ੳ>, and the low vowels (irrespective of their front or back position) are written by adding vowel symbols to the vowel bearer letter <ਅ>.

An important thought to keep in mind is in English each vowel can be used with different sounds but the same does not hold true for Punjabi. In Punjabi, each vowel accent creates only one specific vowel sound. For example, in English, the symbol <a> can represent different sounds as in the word base [beis] and in the word father [faðər]. In Punjabi, each vowel accent creates only one vowel sound so two different vowel accents would create these two sounds. A consonant letter in Punjabi does not stand for a consonant sound but for the syllable "consonant+a". This is why the symbol for the Punjabi vowel <ਅ>, /a/ is regarded as invisible. The Punjabi name for this invisible vowel symbol is "Mukta", 'liberated' because it's pure liberated soul is free from all earthly blackness. The names of the visible vowel symbols of Punjabi are as follows in the table 3.

Table 3: The table represents the Punjabi vowel symbol and independent vowels.

| **Vowel symbol** | ਅ *mukaṯā* | ◌ਾ *kannā* | ◌ਿ *siâṛī* | ◌ੀ *biâṛī* | ◌ੁ *auṅkaṛ* | ◌ੂ *dulaiṅkaṛ* | ◌ੇ *lā̃/lāvā̃* | ◌ੈ *dulāvā̃* | ◌ੋ *hoṛā* | ◌ੌ *kanauṛā* |
|---|---|---|---|---|---|---|---|---|---|---|
| **Punjabi Independent vowel** | ੳ | ਆ | ਇ | ਈ | ਉ | ਊ | ਏ | ਐ | ਓ | ਔ |

The following list shows where vowel signs are positioned around a base consonant to produce vowels, and how many instances of that pattern there are.

- **Pre-base** comes before the base consonant, eg. ਵਿ
- **Post-base** comes after the base consonant, eg. ਕੀ
- **Superscript** comes at the top position of base consonant, eg. ਕੇ

- **Subscript** come at the bottom of base consonant, eg. ਕ੍ਰ

At maximum, vowel components can occur concurrently on 1 side of the base.

### 2.2.3 Nasalisation

Gurmukhi script uses two different diacritics to indicate nasalisation.

- ◌ੰ [Tippi] is used with vowels *a*, *i*, *u*, and with final *ū*, eg. banda <**ਬੰਦਾ**>.
- ◌ਂ [Bindi] is used for all other vowels, eg. Śānta <**ਸ਼ਾਂਤ**>.

The consonants can be geminated (doubled). The above diacritics can be used to indicate doubling of m or n, eg. Lammā <**ਲੰਮਾ**>, here the tippi diacritic and the consonant m are used to signal gemination.

### 2.2.4 Tone

Punjabi is tonal language which uses alteration in pitch to convey different words rather than just using consonants and has three tones high falling, low rising and level. The letters haਹ>, ghaਘ>, jhaਝ>, ḍhaਢ>, dhaਧ> and bhaਭ> have a level tone when at the beginning or a word of syllable, and a high falling tone when elsewhere. The conjuncts ghaਗ੍ਹ>, jhaਜ੍ਹ>, ḍhaਡ੍ਹ>, dhaਦ੍ਹ> and bhaਬ੍ਹ> have a level tone when at the beginning or a word of syllable, and a low rising tone when elsewhere. The different levels of tones explained are as follow:

- **High falling**: - This tone is marked by the letters ਘ /g'à/ ਝ /j'à/, ਢ /t.'à/ ਧ /t'à/ ਭ/p'à/. If these consonants are placed in front of the word they are pronounced as tonal forms of the first consonant but if they occur in the middle or at the end of a consonant they are pronounced as voiceless unaspirated stops for example: -ਘ is /k'à/ in the front but /g'à/ in the middle or at the end.

- **Low rising**: - This tone is indicated by the letter ਹ /haha/ which increases the pitch.
- **Level(medium) tone**: - It is the normal tone which every non tonal language uses.
- Some example of high falling (HFT) v/s low level tone (LLT).:- ਕਰ (do)(LLT)/ਘਰ (HFT)(home), ਕੜਾ(bracelet)(LLT)/ਘੜਾ( pot) (HFT). Some examples of low rising (LRT) v/s level tone (LLT) :- ਕਿੰਨਾ(how much)(LLT)/ ਕਿੰਨ੍ਹਾਂ(who (plural))(LRT). There is another set of examples that signifies the importance of tone in punjabi, eg,. -- ਝਾ /chá/ "peep", ਚਾ /cha/ "desire", ਚਾਹ /chà/ "tea" and ਘੜੀ /káṛī/ "watch", ਕ਼ੜੀ /kaṛī/ "link of a chain", ਕ਼ੜ੍ਹੀ /kàṛī/ "curry".

Table 4: Example of one word having different tones.

| **Low Tone** | | | **Level Tone** | | | **High Tone** | | |
|---|---|---|---|---|---|---|---|---|
| ਘ਼ੜੀ | Ghadi | Watch | ਕ਼ੜੀ | kadi | Link of chain | ਕ਼ੜੀ | kadhi | Turmeric curry |

### 2.2.5 Glyph shaping and positioning.

Glyph is the specific shape, design or representation of the character (Strizver 2011). The orthography has no case distinction. There are no capital letters in Punjabi (Jain et al. 2021, 808). For example, in English, in proper names - Peter else peter but in Punjabi its only written as “ਪੀਟਰ”. Also, there is no apostrophe or other symbols in Punjabi to identify possessive nouns (Jain et al. 2021). Within a Gurmukhi word, spacing glyphs are joined together at the top bar (shirorekha). The top bar extends across or through most spacing letters, including both consonants and vowels, but some letters create a gap in the line (while still joining at either side). Also, Punjabi

language has postpositions rather than prepositions and paraphrases (Jain et al. 2021), For example., “Naśē dī lata lagaṇā” which means “addiction to drugs” vs. “Naśēṛī” which translates as “drug addict”.

### 2.3 Dialects of Punjabi

Punjabi expatriates around the world speak a creolized form of the language that is increasingly deviating from the norms of Punjabi spoken in India and Pakistan. Commonly recognised Punjabi dialects include Majhi (the standard), Doabi and Malwai (“Punjabi Dialects and Languages” 2023). These dialects are represented in four different regions of Punjab, that is, Majha, Malwa, Doaba respectively.


Figure 3: The figure shows the geographical map of Punjab and dialects used in Punjab.

### 2.4 Where do Punjabi speakers live?

In some cases, community members still all reside in their same traditional territory. Many speakers and learners are found far from their traditional territory, having migrated to urban centers in search of better economic opportunities.

As foreign dreams have always fascinated Punjabis, according to figures presented in the India Lok Sabha by the minister of state for external affairs V Muraleedharan approximately 4780 thousand people from Punjab have moved to foreign countries from January 2016 till April 2022 and out of which 2.62 are students (S. Verma 2021). According to Christopher Kerr, director of Operations for Immigration, Refugees and Citizenship Canada (IRCC), there had been an increase of 400 percent in the number of youth migrating to Canada in the year 2021 and 2022 itself. He said Punjabis who made 2.3 per cent of the country's population constituted 60 per cent of the migration to Canada ("Punjabis Contributed to 60% Migration to Canada" 2020). Punjabi is third most common language in Canada. Recent statistics ("Punjabi Becomes Third Language in Canada's House of Commons" 2015) state that there are approximately more than 400,000 Canadians that identify Punjabi as their mother tongue.

### 2.5 To what extent does a standard and agreed-upon written form of the language exist?

Since the 10th century, in both Pakistan and India, the most common dialect used is Majhi. But from the mid-19th century, Punjabi has spread around the world and has incorporated local vocabulary from the regions where Punjabi emigrants have established themselves. While the language borrows heavily from Urdu, Hindi, Sanskrit, Persian and English, there are loanwords

from Spanish and Dutch in the evolving Modern Punjabi. Some similar words are depicted in the following table -

Table 5: The table represents examples of Punjabi loanwords from other languages.

| **English** | **Spanish** | **Hindi** | **Sanskrit** | **Punjabi** |
|---|---|---|---|---|
| Soap | Jabón (Habon) | Sabun/साबुन | Fenak/फेनक | Saban/ਸਾਬਣ |
| Young | Joven (Hoven) | Jawan/युवा | Yauvana/युवा | Jawan/ਜਵਾਨ |

# CHAPTER 3: WHAT IS COMPUTATIONAL LINGUISTICS?

For a language to be used digitally, it must have computational resources. This chapter will introduce Natural Language Processing and Computational Linguistics. I will discuss the important resources and tools required for the development of digital language.

## 3.1 Natural Language Processing and Computational Linguistics

Natural Language Processing (NLP) is the branch of computer science, more specifically, the branch of artificial intelligence or AI, that strives to build tools that give the machine the ability to understand text and spoken words and respond with text or speech of their own just like the humans do. NLP combines computational linguistics with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to 'understand' its full meaning, complete with the speaker or writer's intent and sentiment (Tsujii 2021).

With the aid of NLP, computer programs translate text from one language to another, respond to spoken commands, and summarize large volumes of text rapidly—even in real time. Several NLP tools break down human text and voice data so that it can help the computer make sense of the input fed to it. Since, in NLP, resources and tools are inextricably interlinked, in the sections below I will discuss both resources as well as tools. While an exposition of all possible NLP tools is beyond the scope of this thesis, I would go into depth with only some of the important resources and tools (Tsujii 2021).

Before I dive into further details of language resources and NLP tools, I want to point out the fact that developing NLP tools for any language is not an easy task nor is it a matter of developing a finite set of tools. Richard Littauer (2018) figured out one difficulty in this process is

finding out what has been done before, to avoid duplicated work. It is a continuous process that involves consistent development, generally involving dozens of linguistics and software developers working on various parts of the process. It is to solve this need that resource aggregators exist.

## 3.2 Resource Aggregators

For a linguist, linguistic practitioner, or language activist, it is a tedious task to find language resources. To solve this issue there are many organizations and databases where it is easier to find resources such as dictionaries, academic references, and software for low resource languages. Some of the major examples are as follows.

### 3.2.1 The Unicode Common Local Data Repository (CLDR)

This project provides the largest and most extensive standard repository of locale data available. Organizations like Google, IBM, Apple use this data for their software internationalization and localization, adapting software to the conventions of different languages for such common software tasks (“Unicode CLDR”).

### 3.2.2 Ethnologue

Ethnologue, which is both a book (Lewis et al., 2009) and an online resource, is the most comprehensive resource describing the world’s languages, such as population, size and the general geographic locations of speakers. It is published by SIL International, an evangelical Christian non-profit organization. Ethnologue gives your insight into each of the world’s nearly 7,500 known languages — whether used daily by people or existing only as a memory of cultural heritage. The

documented number is in constant flux because languages are living and dynamic. It is considered the most authoritative source on the languages of the world.

### 3.2.3 Glottolog

Glottolog is an open-source alternative to Ethnologue, developed at the Max Planck Institute for Evolutionary Anthropology. It has over 180,000 references, with information on over eight thousand languages (Hammarström et al. 2015). Sebastian Nordhoff and Harald Hammarström created the Glottolog/Langdoc project in 2011 (Nordhoff and Hammarstrom 2012). The creation of Glottolog was motivated by a lack of comprehensive language bibliography. Glottolog provides a catalog of the world's languages and language families and a bibliography on the world's less-spoken languages. It tries to accept only those languages that the editors have been able to confirm both exist and are distinct. Bibliographic information is provided, especially for lesser-known languages.

### 3.2.4 Omniglot

Omniglot was set up in 1998 by Simon Ager who has been maintaining and developing the site since then. Many other people have made contributions of new material, corrections and suggestions. It is the online encyclopedia of writing systems and languages that contains information on writing systems for around a thousand languages. It provides details about 345 writing systems, including Abjads, Alphabets, Abugidas, Syllabaries and Semanto-phonetic scripts; and has information of about 1800 languages (Ager 2023).

### 3.2.5 The Online Database of Interlinear Text (ODIN)

The Online Database of Interlinear Text (ODIN) is a multilingual repository of annotated language data for 1274 languages. The database is formed by crawling scholarly linguistics articles on the web and looking for Interlinear Glossed Text (IGT) examples. The repository is broad coverage as it contains data for a variety of the world's languages and is limited only by what data is available and what has been discovered. Currently, ODIN houses over 14,000 instances of IGT drawn from 903 linguistic documents with over 400 languages represented (Lewis and Xia 2010).

### 3.2.6 The Open Language Archives Community (OLAC)

The Open Language Archives Community (OLAC) is a worldwide virtual library of language resources (Simons and Bird 2003) and contains thousands of cross-references to resources both on the web and in print. The information about resources is stored in XML format for easy searching. OLAC was founded in 2000 and is hosted at the Linguistic Data Consortium webserver at the University of Pennsylvania.

### 3.2.7 Wikipedia

Wikipedia is the largest and most popular general reference work on the Internet. The Wikipedia has a nontrivial number of articles on low-resource languages, many of which have references themselves to scholarly work. Kornai (2013), among others, notes that Wikipedia is one of the first ports-of-call for new language communities, and while it is not a precondition for having corpora on the web, it is a sine qua non for digital vitalisation.

### 3.2.8 The World Atlas of Language Structures (WALS)

The World Atlas of Language Structures (Dryer and Haspelmath 2013) is a large database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials by a team of 55 authors. WALS Online is a publication of the Max Planck Institute for Evolutionary Anthropology. The first version was released in 2013 and sets of corrections have been published periodically.

### 3.2.9 CLARIN

CLARIN is an acronym for Common Language Resources and Technology Infrastructure (CLARIN). It is a digital infrastructure which provides easy and sustainable access to a broad range of language data and tools to support research in the humanities and social sciences, and beyond. CLARIN provides access to multimodal digital language data (written, spoken or multimodal form) and advanced tools with which to explore, analyze or combine these datasets (Hinrichs and Krauwer 2014).

### 3.2.10 ELSNET and ELRA

ELSNET, the European Network of Excellence in Language and Speech, came into existence in 1991, with an aim to facilitate, support and coordinate the creation of language and speech systems. ELRA, The European Language Resources Association was founded in 1995 with the mission of providing language resources (LR) to European research institutions and companies. The ELRA goal is to serve as a focal point for the collection and distribution of language resources in both speech, text and terminology (Krauwer 1998). Since ELSNET and ELRA work on the same

grounds, which is to create resources for lesser resource languages, I think organizations like these can make remarkable contributions in the field of language resources.

### 3.2.11 Language Data Consortium (LDC)

The Language Data Consortium (LDC) is an open consortium of universities, companies and government research laboratories. The host institution of LDC is University of Pennsylvania and was granted from the Defense Advanced Research Projects Agency (DARPA). While some of its data creation is funded by grants provided by the Information, Robotics and Intelligent Systems division of the National Science Foundation (NSF), most of its publications and distribution activities are self-supporting. The core principle of LDC is to support pre-competitive research and development in speech and language technology, along with the idea to support other language-related (Liberman and Cieri 1998).

Now that it is clear what resource aggregators are, and that it is possible to at least get a basic idea of what resources there are for a given language, what resources are relevant to low resource languages, I would like to discuss about the type of language resources and Natural Language Processing (NLP) tools required for a language.

## 3.3 Types of Language Resources

### 3.3.1 Corpora

Richard Littauer (2018) described, “All language resources ultimately depend upon corpora; without data, an algorithm does nothing.”. A corpus is a collection of linguistic data, either written texts or transcription of recorded speech or samples of spoken and written language, which can be

used for linguistic description. Historically, most of the corpora have been written corpora, due to difficulty in gathering audio, video files. But with the advent of social media and platforms like youTube, where people can upload content in their own language, audio and visual corpora are also becoming prevalent (Richard 2018). Both types of corpora are useful for Language Resource (LR); written corpora are useful for setting up fonts and characters and implementing them in Unicode, while the latter can also be used to begin extracting phonemes (Kempton and Moore 2014; Mueller et al. 2017), or for speech-to-text systems (Laurent et al. 2016).

In the past few years, the term corpus has been increasingly applied to a body of language which exists in electronic form, and which may be processed by computer for linguistic research or language processing (Garside, Geoffrey, and Tony 2013, 4). It provides the basis for computation linguistics. With the increase in computer power, corpora have also increased dramatically in size, variety and ease of access. Large corpora, especially for English, have been compiled since the 1980s and are used in the development of natural language processing software (Crystal 2003). One of the examples of English corpora is The British National Corpus (BNC) which consists of 100 million words of English written texts and spoken transcriptions, sampled from a comprehensive range of text types (Leech, Garside, and Bryant 1994). Not all corpora are same, it is further classified into different categories which are defined as follows.

### 3.3.2 Annotated Corpora

Corpora can be more useful if we can extract knowledge or information from them, which can only be possible if we feed in the information in the corpora by adding annotations. The information can vary in nature, it could be semantic annotation or prosodic or adding grammatical tags in the corpus. The raw corpus, with its orthographic form, gives no direct information like grammatical markup,

parts-of-speech, linguistic phenomenon etc which can hinder many applications in which the corpora can be put in (Garside, Geoffrey, and Tony 2013, 4). To exemplify, the word spelt as “lead” in English, can be either a noun, pronounced as /lɛd/ or a verb pronounced as /li:d/. “Lead”, having different meanings and usages, it is hard to detect the correct meaning from orthographic form. But if the corpus is annotated, each occurrence of lead would be tagged with a world class label, indicating the correct usage of the word.

The Brown Corpus (Francis and Kucera 1979), the LOB Corpus (Johansson, Leech, and Goodluck 1978 ) And the British National Corpus (BNC) (Burnard 2000) are examples of grammatically annotated corpora; the words have been assigned a word class label (part-of-speech tag). The LLC Corpus (Greenbaum and Svartvik 1990) has been prosodically annotated. The Susanne Corpus (Sampson 1994) is an example of a parsed corpus, a corpus that has been syntactically analyzed and annotated.

### 3.3.3 Unannotated Corpora

As the name suggests, it is a corpus that does not have annotated documents, it is in the raw state of plain text with no additional information. The raw text is further used by linguistics to design more specific corpus by adding annotation or tags in the text. It can range from small unannotated corpora to large unannotated corpora.

### 3.3.4 Monolingual Corpora

A monolingual corpus contains texts in one language only. The corpus is tagged for parts of speech and is used for checking usage of a word or looking up the most natural word combinations. A specialized monolingual corpus can be used as a translational resource (Bowker 2002).

Enabling Minority Language Engineering (EMILLE) Corpus, a collaborative work by Lancaster University and the Central Institute of Indian Languages (CIIL), India, contains monolingual corpora for fourteen South Asian Languages: Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Oriya, Punjabi, Sinhala, Tamil, Telugu and Urdu. The corpora contain 93 million words including 2.6 million words of transcribed spoken data for Bengali, Gujarati, Hindi, Punjabi and Urdu (P. Baker, Hardie, and McEnery 2006).

### 3.3.5 Multilingual Corpora

A corpus that contains text in more than one language. Like monolingual corpora, multilingual corpora are also useful in translation studies (M. Baker 1993). Since, multilingual corpora are based on translations, it can help the researchers to compare the languages and at the same time they can use their first language as a tool to learn the second language (Frankenberg-Garcia 2005, 190–92). Multilingual corpora is a more specified type that is known as parallel corpus which defines a relationship between the texts in the different languages, i.e. the texts are direct translations of one another (P. Baker, Hardie, and McEnery 2006) .

The European Corpus Initiative (ECI) was founded to create a large multilingual corpus for scientific research. The corpus (ECI/MCI) contains texts in a range of languages including German, French, Spanish, Dutch, Albanian, Chinese, Czech and Malay as well as some parallel texts (P. Baker, Hardie, and McEnery 2006).

### 3.3.6 Multimodal Corpora

In Multimodal corpora, annotations coexist alongside other types of data such as transcriptions and video streams (P. Baker, Hardie, and McEnery 2006). The corpora include gesture annotations as

well and sometimes textual data is supplemented with images too. A Corpus is used to study how two or more modalities interface with one another in human communication. Such corpora can be used for, “the exploration of a range of lexical, prosodic and gestural features of conversation, and for investigations of the ways in which these features interact in real, everyday speech” (Abuczki and Ghazaleh 2013, 88).

The CLARIN infrastructure offers 17 multimodal corpora, 13 of which are monolingual (English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Slovenian, and Zulu). These corpora are richly annotated for various verbal and non-verbal elements of communication, such as body gesture, gaze direction, and head, eye, and lip movement (P. Baker, Hardie, and McEnery 2006).

### 3.3.7 Monolingual/Multilingual Lexicon

In linguistics, a lexicon is the inventory of lexemes (a set of words that are related through inflection) for a particular language or, in simple words, it is a catalog of words of language (Frankenberg-Garcia 2005). Lexicons can be used as a basis for dictionary creation. Besides, it is an important resource for automated corpus analysis. Part-of-speech-tagging and semantic tagging usually depend on extensive computer lexicons containing lists of word forms and the possible tags they can have (P. Baker, Hardie, and McEnery 2006).

Alvey Natural Language Tools (ANLT), is a set of tools for use in natural language processing research, created at the Universities of Cambridge, Edinburgh and Lancaster between the late 1980s and early 1990s. These include a morphological analyser, a grammar, two parsers and a lexicon containing 63,000 entries (P. Baker, Hardie, and McEnery 2006).

### 3.3.8 Thesaurus, Wordnet, Ontologies

A Thesaurus is a book that lists antonyms and synonyms of a word. WordNet is a lexical database of semantic relations (structural similarities) between words that links words into semantic relations including synonyms, hyponyms, and meronyms. Natural Language Ontology, as defined by Stanford Encyclopedia of Philosophy (Moltmann 2022), is a sub-discipline of both philosophy and linguistics, more specifically, of metaphysics and natural language semantics.
NLP tasks which could benefit from a high-quality thesaurus include parsing, anaphor resolution, establishing text coherence and word sense disambiguation (Kilgarriff 2003).

Thesaurus Linguae Graecae (TLG), An on-line archive of texts written in ancient, classical and medieval Greek, contains more than 90 million words, representing nearly all the surviving literature in Greek prior to ad1453 (P. Baker, Hardie, and McEnery 2006).

## 3.4 Natural Languages Processing (NLP) Resources

Natural Language Processing (NLP) is a very vast field and as I have already mentioned, discussing the various NLP applications and resources is beyond the scope of this thesis but it is worth exploring what types of computational resources are common for natural language processing (NLP) work. In the following section, I have discussed some of the examples.

### 3.4.1 Morphological analyser

As described by Beesley, “Creating a morphological analyzer/generator is just one step in starting natural language processing for any language; but especially for minority, emerging or generally lesser-studied languages.” (Beesley and Karttunen 2003). Morphological analysis is a field of linguistics that studies the structure of words. It studies how a word is formed by using a morpheme

(smallest element of a word that has grammatical meaning and cannot be divided further). In inflected languages like English, words are formed by adding suffixes. For example, by adding the suffix '-s' to the verb *to dance*, we form the third person singular *dances*.

A morphological analyzer evaluates inflection, a given word has undergone and then assigns attributes to the word. If you give it the word 'crying' in English, it will tell you it is the present participle of the verb 'cry'.

If a language has a corpus of million words and has sound language resources, working on morphological analysis and generation leads directly to the other lexical applications such as tokenization, indexing, base for reduction, spell checker. Morphological analyzers can also serve as reusable enabling components in larger systems that perform disambiguation, syntactic parsing, speech generation, speech recognition, etc (Beesley and Karttunen 2003).

### 3.4.2 Parts-Of-Speech tagger

Also known as grammatical tagging, POS tagging is the process of automatically tagging the words in a text corresponding to a particular part of speech, based on both its definition and its context, that is identification of words as nouns, verbs, adjectives or adverbs, etc (Van Halteren 1999).

Example of POS tagging -

Table 6: Example of POS tagging

| **Word Tokenizer** | I | like | to | read | books |
|---|---|---|---|---|---|
| **Part-of-speech** | pronoun | verb | preposition | verb | noun |

In figure 7, each word of a sentence is categorized with lexical terms, however, taking into consideration that POS tagging will be performed on a large corpus of any language, it's

complicated to write these full terms while doing text analysis. Hence, short representations referred to as “tags” are used to represent the categories. The set of predefined tags is called a tag set. In The British National Corpus (BNC) World Edition each word is tagged according to its word-class with POS tags. Contractions and possessives written with apostrophes are tokenized as separate units and receive tags as well (Leech, Garside, and Bryant 1994). Depending on some general principle of tag set design strategy, a number of POS tag sets have been developed. Geoffrey Leech and Nicholas Smith's Manual to accompany The British National Corpus (Version 2) with Improved Word-class Tagging describes the Constituent Likelihood Automatic Word-tagging System (CLAWS) parser (Leech, Garside, and Bryant 1994) and explains the PoS codes in greater detail.

POS tagging is mainly used to extract information from that corpus to create better dictionaries or grammars of the language from real language data. In order to extract information from corpora it is important to add explicit linguistic information to the texts and that is where POS tagging comes into the picture (Mitkov 2014).

Table 7: The table depicts some of the POS tags used for BNC (Leech, Garside, and Bryant 1994) along with their example.

| **Lexical Term** | **Tag** | **Example** |
|---|---|---|
| Common Noun | NN1 | Aircraft |
| Plural Noun | NN2 | Pencils, geese |
| Proper Noun | NP0 | London, Michael |
| Personal Noun | PNP | I, you, them, ours |
| Preposition | PREF | of |
| Adjective | AJ0 | Good, old |
| Comparative Adjective | AJC | Better, older |
| Verb | VBB | Live, read |
| Cardinal Number | CRD | One, 3, fifty-five |
| … | … | … |

Table 8: Reconsidering the above POS tag sets, example in the Table 8 can be restructured as follows.

| **Word Tokenizer** | I | like | to | read | books |
|---|---|---|---|---|---|
| **Part-of-speech** | PNP | VBB | PREF | VBB | NN2 |

### 3.4.3 Word Sense disambiguation

Word sense disambiguation (WSD) is the process of deciding the senses of the words in the context. WSD has been a research area in the field of Natural Language Processing if the field exists (Mitkov 2014). In everyday life, when an ambiguous word is spoken, we somehow sense the correct meaning of the word in the context but in an application where a computer has to process natural language processing, ambiguity can be a problem. For example, if a language translation system encountered the word 'bat' in a sentence, should the translator regard the word as meaning: an implement used in sports to hit balls; or a furry, flying mammal? (Sanderson 1994)

Word Sense Disambiguation is applied in almost every application of language technologies like machine translation, information retrieval and lexicography.

### 3.4.4 Named entity recognition.

Named entity recognition, or NER, identifies proper nouns like the name of the person, organization or location. It detects the boundaries of the sentences by analyzing capital letters. It is mainly used for information extraction, question answering and machine translation. NER systems need extensive gazetteers - lists of names of people, organizations, locations and other named entities (Mikheev, Moens, and Grover 1999).

NER software serves as an important preprocessing tool for tasks such as information extraction, information retrieval and other text processing applications. It depends on the application that makes use of annotations. For example, Named Entity annotation allows you to search for all texts that mention the company "Philip Morris", ignoring documents about a possibly unrelated person by the same name (Mikheev, Moens, and Grover 1999).

Different named entity recognition systems are introduced by different researchers for the languages like English, French, etc. One of the major advantages of these languages is the capitalization feature which is the main clue to identify the named entities. Secondly, for these languages, there are sufficient annotated datasets and language resources like morphological analyzer, POS tagger available (Mikheev, Moens, and Grover 1999).

### 3.4.5 Syntactic analysis

Syntactic analysis is the process of understanding the syntactic organisation of the sentence which is made up of a sequence of words (Redd 2014). It analyzes the relationship between words and grammatical structure of sentences in order to derive the internal relationships of the constituents.

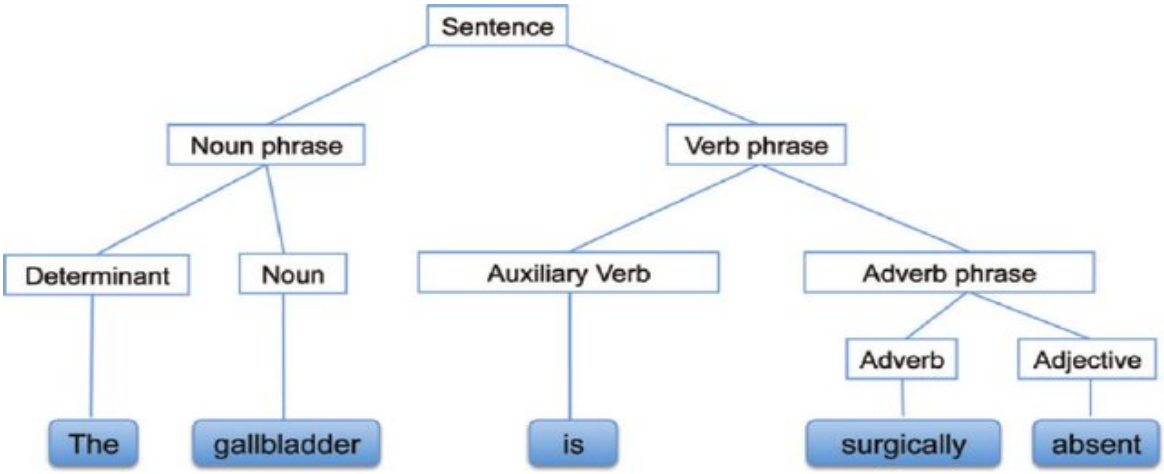


Figure 4: The diagram illustrates the syntactic analysis of a sentence (Cai et al. 2016, 180).

The above illustration demonstrates the syntactic analysis of the sentence, 'The gallbladder is surgically absent'. Each word is assigned a part-of-speech tag using grammatical rules. In natural language processing, syntactic analysis is an extremely important aspect in natural language understanding because it assists in figuring out the grammatical meaning of any sentence.

### 3.4.6 Semantic analysis

Semantics is to study the meaning of the language. With the different linguistic levels of phonology, lexicon and syntax, a language exhibits a meaningful message because of the semantic interaction. In NLP, syntactic and semantic analysis goes hand in hand. Once the computer has arrived at an analysis of the syntactic structure of the input sentence, a semantic analysis is needed to ascertain the meaning of the sentence (Redd 2014).

Since we want NLP applications to produce interpretable sentence, semantic analysis plays an important. For example, *colorless green ideas sleep furiously***,** the sentence follows syntactic rules, but semantically the sentence is meaningless. Hence, semantic analysis is important for NLP applications like Information retrieval.

### 3.4.7 Shallow Parsing

Shallow parsing is the process of assigning partial syntactic structure to sentences. It is also known as chunking; it basically aims to identify and extract phrases or chunks from a sentence. It focuses on identifying phrases or constituents, such as noun phrases, verb phrases, and prepositional phrases. Shallow parsing is an essential component of many NLP tasks, including information extraction, text classification, and sentiment analysis (Federici, Montemagni, and Pirrelli 1996).

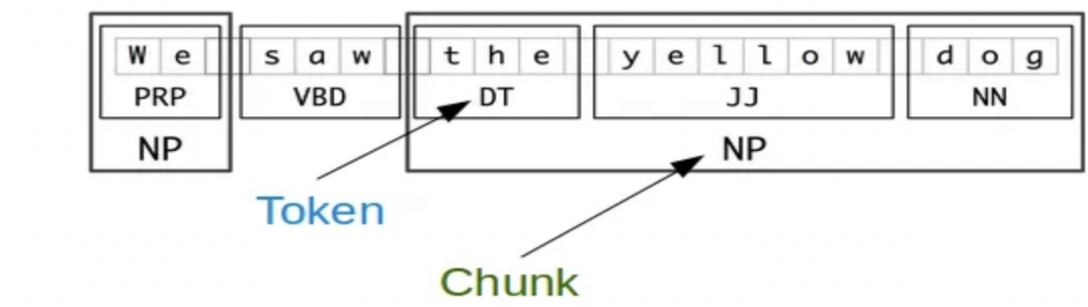


Figure 5: The illustration depicts the example of shallow parsing of sentence (Everton 2023).

As shown in the above illustration, one of the most common types of shallow parsing is noun phrase chunking, which identifies and extracts all the noun phrases in a sentence. Noun phrases typically consist of a noun and any associated adjectives, determiners, or modifiers. For example, in the sentence, *'We saw the yellow dog',* the noun phrase *'the yellow dog'* can be identified and extracted using noun phrase chunking.

### 3.4.8 OCR (Optical character Recognition)

Optical character recognition (OCR) method has been used in converting printed text into editable text. It converts scanned or printed text images, handwritten text into editable text for further processing. As quoted by Patel, "It is like a combination of the eye and mind of the human body. An eye can view the text from the images but actually the brain processes as well as interprets that extracted text read by eye." (Patel, Patel, and Patel 2012, 50).

In 1955, the first commercial system was installed at the reader's digest, which used OCR to input sales reports into a computer and then after OCR method has become very helpful in computerizing the physical office documents (Patel, Patel, and Patel 2012, 50). There are many applications of OCR, which includes: License plate recognition, image text extraction from natural scene images,

extracting text from scanned documents etc. In 2008, a system named “Geometric Rectification of Camera-Captured Document Images” was proposed to rectify the text retrieved from camera captured images (Liang, DeMenthon, and Doermann 2008).

### 3.4.9 Speech Recognition

Speech recognition, also called speech-to-text, is the task of reliably converting voice data into text data. Speech recognition is required for any application that follows voice commands or answers spoken questions. It was designed and developed to achieve systems to provide significant help so people can share information by operating a computer using voice input. It enables the system to recognize speech and convert input audio into text; it also enables a user to perform file operations like Save, Open, or Exit from voice-only input (Vashisht, Pandey, and Yadav 2021).

# CHAPTER 4: HOW DO I UNDERSTAND IN A SYSTEMATIC WAY THE DEGREE TO WHICH A LANGUAGE IS UNDER-RESOURCED?

Unless a language improves its visibility in the digital world, it is heading for extinction.

—- Edmond Kachale

This chapter is the premise of this thesis that Punjabi is a low resource language. In the previous chapter, I have reviewed some of the major components required to support natural language processing in any language and noted in passing when these tools and resources are available for Punjabi. This raises the question of how one identifies what is missing in support for a language: once you recognise that a language is not as well supported as a major language like English, how to quantify or identify what needs to change in order to improve its support. One approach might be to look at development activity: how much research and development is going on in support of NLP in each language. Another might be a simple count: how many tools and resources are there. But none of this is fully satisfactory, because resources are interdependent: in order to develop support for a language, certain resources and tools must be developed first; and others can follow. We require a roadmap for determining the state of support for a language and especially recognising the interdependencies.

## 4.1 What do you mean by a roadmap?

As Krauwer describes, “There exists no single universal definition of what a roadmap is and how it should be constructed or presented.” (Krauwer 2003). He defines that roadmap will help researchers, educators, developers, service providers, and funders decide where to concentrate their efforts in order to give a maximal push to the development of the field. It identifies the main

challenges, the intermediate goals that may help in measuring progress or reconsidering long term goals if we are not getting expected results. Krauwer further explains the ELSNET approach to the roadmap takes several objects (challenges and milestones) into consideration with different categories (like resources, technologies and applications), and other information such as expected year of completion, definition, etc.

I will discuss the roadmap that results in making language and speech technology for Punjabi. Punjabi, being a small language, has a small national economy and hence has less opportunities for private or public funding to produce technologies. My thesis does not solve the financial conditions of Punjabi, but it aims to contribute to the creation of the starting conditions for language and speech technology that is to educate language and speech researchers and engineers, to design proper curriculum for the creation of required language resources like corpora, dictionaries, parsers, tools, etc.

More for some and less for others, language resources exist for many languages, and they exist for Punjabi as well but what is the standard count of language resources to know that this count is sufficient in terms of NLP standards. In the next section I will outline how one can arrive at such a definition by introducing the concept of a BLARK: a Basic Language Resource Kit.

### 4.2 The Basic Language Resource kit (BLARK)

The Basic Language Resource Kit (BLARK) defines the number of resources required for a language to do any pre-competitive research (research conducted jointly by usually competing companies for the purpose of developing new commercially applicable technologies) or any speech technology research. ‘BLARK’ is a concept coined by Steven Krauwer in a 1998 paper ‘ELSNET and ELRA: A common past and a common future” (Krauwer 1998).

ELSNET, the European Network in Human Language Technologies, which was created with the objective of bringing together the language and speech communities and promoting research and development in language and speech technology, has been actively working on Language Resources. Language resources have always been a prominent point on ELSNET'S agenda, a task group was organized who were responsible to take new initiatives in research and development of Language resources.

Krauwer figured out that Central and Eastern Europe were not intellectually in position to compete with Western Europe with respect to language and speech technology. Hence, the idea manifested in the paper was to define a minimal set of resources that should be available for precompetitive research of any language. Arppe et al. explains that based on collective experience gained with many different languages, it has been concluded that is in principle language-independent (Arppe et al. n.d.). Every language has specific requirements, for instance, Chinese uses tone to distinguish words so it is expected that Chinese language data must have tone markup and Chinese BLARK should have included tone-detection parameter, but the same does not hold true for English. Hence, there might be some variations in the parameter depending on the language under consideration. For some languages a BLARK model has been constructed like Dutch (cf. section 4.3), Arabic (cf. section 4.4), Swedish (ongoing, cf. (Elenius, Forsbom, & Megyesi 2008)) --, for other languages the ideas behind BLARK will have been considered without adopting a full-fledged BLARK. Since, Dutch and Arabic are the langauges for BLARK has been maximally impactful, I am using Dutch and Arabic as reference throughout in my thesis.

### 4.3 The Dutch BLARK

A first BLARK was constructed in the Flemish-Dutch action plan "Dutch HLT resources" (Strik et al. 2002). The main objective of the paper was to stimulate the cooperation between industry and scientific institutes and to provide an infrastructure that will make it possible to develop, maintain and distribute HLT resources for Dutch. The work carried out in the research was organized along four action plans, i.e

- To establish a coordinating office.
- To define a set of basic HLT resources for Dutch.
- To determine if the resources are available or not.
- To define a blueprint for management, maintenance and distribution.

The important step, while defining basic Dutch HLT resources, was to define the BLARK. HLT resources were divided into three categories, i,e applications (class of applications that make use of HLT), modules (software components that are required to develop HLT applications) and data (datasets and electronic descriptions that can be used to build, improve or evaluate modules).
In order to make the model generic, irrespective of the language type, a matrix was drawn up which described (1) the modules required for different applications (2) data required for different modules (3) relative importance of module and data. This matrix was subdivided in language and speech technology, where "+" means important and "++" means very important.The Dutch BLARK was proposed consisting of the two components - Language and Speech Technology (Strik et al. 2002).

#### 4.3.1 Modules for Language Technology

- Robust modular text pre-processing (tokenization and named entity recognition)

- Morphological analysis and morpho-syntactic disambiguation
- Syntactic analysis
- Semantic analysis

### 4.3.2 Data for Language Technology:

- Mono-lingual lexicon
- Annotated corpus (a treebank with syntactic, morphological, and semantic structures)
- Benchmarks for evaluation

### 4.3.3 Modules for Speech Technology:

- Automatic speech recognition (including tools for robust speech recognition, recognition of non-natives, adaptation, and prosody recognition)
- Speech synthesis (including tools for unit selection)
- Tools for calculating confidence measures.
- Tools for identification (speaker identification as well as language and dialect identification) Tools for (semi-) automatic annotation of speech corpora

### 4.3.4 Data for Speech Technology:

- Speech corpora for specific applications, such as Computer Assisted Language Learning (CALL), directory assistance, etc.
- Multi-modal speech corpora
- Multimedia speech corpora
- Multilingual speech corpora

- Benchmarks for evaluation

The next step in the action plan was to investigate which of these data and tools are already available for Dutch and which resources can be reused. All the information was collected in a matrix, as shown in the table below, which was then sent to experts in both industry and academia to get a more complete picture. Based on the availability matrix, a priority list was formulated which played an important role in the setup of the joint Flemish-Dutch STEVIN programme (Arppe et al. 2010).

After creating the initial definition of BLARK for a language, then a case study was conducted to analyze the components or features that were already available for Dutch and the ones that were missing. The gaps were identified, priorities were assigned to the components to be produced, so that a realistic plan for the gradual completion of the BLARK becomes feasible.

| Modules | Data | | | | | | | | | Applications | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | monoling lex | multilin lex | thesauri | anno corp | unanno corp | speech corp | multi ling corp | multi mod corp | multi media cor | CALL | access control | speech input | speech output | dialog systems | doc prod | info access | transla-tion |
| **Language Technology** | | | | | | | | | | | | | | | | | |
| Grapheme-phoneme conv. | ++ | | | ++ | | | | | | + | | | ++ | ++ | + | + | |
| Token detection | ++ | | | + | ++ | | | | | + | | + | | + | + | + | + |
| Sent boundary detection | + | | | ++ | ++ | | | | | + | | ++ | ++ | + | ++ | ++ | ++ |
| Name recognition | + | + | + | ++ | ++ | ++ | | | | + | | ++ | ++ | + | ++ | ++ | ++ |
| Spelling correction | | | | | | | | | | + | | | | | | | |
| Lemmatising | ++ | | | ++ | + | | | | | + | | + | + | + | + | + | + |
| Morphological analysis | ++ | | | ++ | + | | | | | + | | + | ++ | + | ++ | ++ | ++ |
| Morphological synthesis | ++ | | | ++ | + | | | | | + | | | ++ | + | ++ | | ++ |
| Word sort disambig. | ++ | | | ++ | + | | | | | + | | ++ | + | ++ | ++ | ++ | ++ |
| Parsers and grammars | ++ | | | ++ | | | | | | + | | ++ | ++ | ++ | ++ | ++ | ++ |
| Shallow parsing | ++ | | | ++ | ++ | | | | | + | | ++ | ++ | ++ | ++ | ++ | ++ |
| Constituent recognition | ++ | | | ++ | + | | | | | + | | ++ | ++ | ++ | ++ | ++ | ++ |
| Semantic analysis | ++ | | ++ | ++ | | | | ++ | ++ | + | | ++ | ++ | ++ | | ++ | ++ |
| Referent resolution | + | | ++ | ++ | + | | | | | + | | ++ | | ++ | ++ | ++ | ++ |
| Word meaning disambig. | + | | ++ | ++ | + | | | | | + | | ++ | + | + | + | ++ | ++ |
| Pragmatic analysis | + | | + | ++ | | | | ++ | ++ | + | | ++ | ++ | ++ | | + | ++ |
| Text generation | ++ | | ++ | ++ | | | | ++ | ++ | + | | | ++ | ++ | ++ | | ++ |
| Lang. dep. translation | | ++ | ++ | ++ | | | ++ | | | + | | | | | | ++ | ++ |
| **Speech Technology** | | | | | | | | | | | | | | | | | |
| Complete speech recog. | ++ | + | | ++ | + | ++ | + | ++ | ++ | ++ | ++ | ++ | | ++ | ++ | ++ | ++ |
| Acoustic models | ++ | + | | ++ | + | ++ | + | + | + | ++ | + | ++ | | ++ | + | + | + |
| Language models | + | | | ++ | + | + | + | + | + | ++ | + | ++ | | ++ | ++ | ++ | ++ |
| Pronunciation lexicon | ++ | + | | + | | ++ | + | + | + | ++ | + | ++ | + | ++ | + | ++ | ++ |
| Robust speech recognition | + | | | + | + | + | + | + | ++ | + | + | ++ | | ++ | + | + | + |
| Non-native speech recog. | + | ++ | | + | | ++ | ++ | + | + | ++ | + | + | | + | | + | + |
| Speaker adaptation | + | | | + | + | ++ | + | + | ++ | + | + | ++ | | + | + | ++ | + |
| Lexicon adaptation | ++ | + | | + | | ++ | + | + | + | ++ | + | ++ | + | ++ | + | ++ | ++ |
| Prosody recognition | + | + | | ++ | + | ++ | + | + | + | ++ | + | ++ | | ++ | ++ | ++ | ++ |
| Complete speech synth. | ++ | + | | + | | + | | + | | + | | | ++ | ++ | + | + | ++ |
| Allophone synthesis | + | + | | + | | + | | + | | + | | | + | | + | + | + |
| Di-phone synthesis | ++ | + | | + | | + | | + | | + | | | ++ | ++ | + | + | + |
| Unit selection | ++ | + | | + | | + | | + | | + | | | ++ | ++ | + | + | + |
| Prosody prediction for Text-to-Speech | ++ | + | | + | | + | | + | + | ++ | | | ++ | ++ | | + | ++ |
| Aut. phon. transcription | ++ | ++ | | + | + | ++ | + | + | + | ++ | + | + | + | + | + | + | + |
| Aut. phon. segmentation | ++ | ++ | | + | + | ++ | + | + | + | ++ | + | + | + | + | + | + | + |
| Phoneme alignment | + | + | | + | | ++ | + | + | + | ++ | + | + | | + | | | + |
| Distance calc. phonemes | + | + | | + | | ++ | + | + | + | ++ | + | + | | + | | | + |
| Speaker identification | + | | | ++ | ++ | ++ | + | ++ | + | + | ++ | + | | + | | + | + |
| Speaker verification | + | | | ++ | ++ | ++ | + | ++ | | + | ++ | + | | + | | + | + |
| Speaker tracking | + | | | ++ | | ++ | | | ++ | + | ++ | + | | + | + | + | + |
| Language identification | + | ++ | | + | + | ++ | ++ | + | + | + | + | + | | + | | + | + |
| Dialect identification | + | ++ | | + | + | ++ | ++ | + | + | + | + | + | | + | | + | + |
| Confidence measures | + | | | + | + | ++ | + | ++ | + | ++ | ++ | ++ | | ++ | + | + | + |
| Utterance verification | + | | | + | + | ++ | + | + | + | + | + | ++ | | ++ | + | + | + |

Figure 6: The figure depicts the matrix showing the relative importance of data, modules and applications.

### 4.4 The Arabic BLARK

BLARK being language independent, a BLARK for Arabic was published in 2006 within the European Union, Network for Euro-Mediterranean LAnguage Resource (NEMLAR) project (Maegaard et al. 2006). Although, it followed the model of Dutch BLARK, but it differs in some minor points. Dutch BLARK explained the relation between the resources, tools and modules but it lacked the definition of the notions of availability, quality, quantity and standards which are well explained in the Arabic BLARK (Arppe et al. 2010).

#### 4.4.1 Resource Availability

In Binnenpoorte et al (2002), the availability of resources for Dutch was expressed by numerical values ranging from "1" (not available) to "10" (easily obtainable). However, there was no underlying report that could explain how the numbers were assigned or what the numbers meant, therefore, NEMLAR purposed different approach to availability. Three factors played an important role: accessibility, affordability and customizability (Maegaard et al. 2006).

- Accessibility:
    - **3** existents but only company-internal,
    - **2** existent and freely usable for pre-competitive research,
    - **1** existent and freely usable for both pre-competitive research and product development.
- Cost (in Euros) :
    - **4** more than 10,000,
    - **3** 1,000 – 10,000 ,

- **2** 100 – 1,000,
- **1** less than 100 or free.

- Adaptability:
  - **3** black boxes (you can neither see it nor change it, e.g. object code),
  - **2** glass boxes (you can see but not change it)
  - **1** freely manipulable, e.g., source code.

### 4.4.2 Quality of Resource

Binnenpoorte et al. considered the quality of LR but did not provide a specific way it could be measured or expressed. NEMLAR suggests taking the following attributes with corresponding criteria into account (Arppe et al. 2010):

- Standard compliance (Is the resource based on a common standard?)
- Soundness (Internal consistency, i.e. is the resource based on well-defined specs?)
- Task-relevance (Is the resource suited for a specific task X?)
- Environment-relevance (Is the resource interoperable with other resources?)

### 4.4.3 Quantity of Resource

The Dutch BLARK does not provide any quantitative figures for the required resources i.e how many words should a corpus have or how many hours of speech etc. Given the nature of the BLARK model, concerning technologically less well covered languages, it should include very clear guidelines, in figures, for what counts as a sufficiently large corpus or lexicon (Maegaard et al. 2006). In the BLARK for Arabic, figures are mentioned in the description of the respective resources (Arppe et al. 2010).

### 4.4.4 Resource Standards

According to Arppe et al.(2010) adoption of standards is crucial for the longevity of resources. There are only a few official standards for LST, although some de facto standards are emerging.

## 4.5 Why BLARK is important?

The ideal idea is to make a generic BLARK definition which is applicable to all the languages. The common model will save time and effort which otherwise would be used to reinvent the same language resources. It will impart the knowledge gained from one model to researchers and educators working on multilingual or cross lingual application areas. It will ensure interoperability and interconnectivity and also will assist in making estimates of costs and resources required to build the tools. In addition, the BLARK is useful for the humanities researchers who are working on less resource languages and can help train students for their research experiments and application pilots.

Taking this as a point of departure, I want to shed light on BLARK specification for Punjabi. Adopting and implementing Punjabi BLARK is not within the scope of the thesis, so I want to define the basic concepts of BLARK for Punjabi to provide a comprehensive overview of the existing resources, tools, and techniques for Punjabi NLP, as well to identify the gaps and opportunities for future research.

# CHAPTER 5: THE BLARK DEFINITION FOR THE PUNJABI LANGUAGE

As defined by Maegaard et al. (2006), the BLARK definition refers to the proposals for the items to be incorporated. The following section will propose a model that would tell which LR, tools or applications are already available for Punjabi and which ones still have to be developed. As can be seen, I have followed Binnenpoorte et al. (2002) and Maegaard et al. (2006) quite closely. The main contributions lies in making statements on the availability of resources, changing the applications and modules which are not required for Punjabi. As mentioned in chapter 4, BLARK is divided into two components: written language and spoken language. Therefore, two pairs of tables were made, one pair for written and one for spoken language. I will define both, written and speech matrices for Punjabi.

## 5.1 Gurmukhi Punjabi BLARK content for written Resources

The 11 Human Language Technology (HLT) modules are related to 8 Language resources. One of the important LR for Punjabi is grammatical description of Punjabi. Not much importance is given to grammatical description of a language in Arabic and Dutch BLARK but it is one of the important resources required to build Punjabi HLT modules. At this level BLARK becomes language specific as HLT modules or LR may vary with the type of language. For example, Punjabi does not require diacritiser (vowelizer) which is one of the important modules for Arabic (Maegaard et al. 2006) because unlike Arabic, Punjabi does not have diacritics which provide phonetic guide or which helps in correct pronunciation.

The matrix in table 10 shows the results (+ = relevant; ++ = important; +++ = essential)

taken from Arabic BLARK (Maegaard et al. 2006). Based on this matrix one can determine which components serve most applications, and which data are most needed for most applications, i.e., which elements should be part of the BLARK. Vertical columns consist of the Language Resources like Monolingual/Bilingual Lexicon, Thesauri/Ontologies/Wordnets, Unannotated/ Annotated Corpora, Parallel Multilingual Corpora, Multimodal Corpora and grammatical description for Punjabi. The horizontal rows consist of 11 HLT modules which include: Morphological Analyser, POS disambiguator/tagger, Named Entity Recognition, Word Sense Disambiguation, Term extraction, Shallow parsing, Syntactic analysis, Semantic Analysis, Sentence synthesis and generation, Sentence Alignment and OCR (Optical Character Recognition).

Table 9: Written language resources and corresponding HLT modules, marked with importance (Maegaard et al. 2006).

| **Language Resource** | Monolingual Lexicon | Multi/Bilingual Lexicon | Thesauri, Ontologies, Wordnet | Unannotated Corpora | Annotated Corpora | Parallel Multi-lingual Corpora | Multimodal Corpora | Grammatical Description of Punjabi |
|---|---|---|---|---|---|---|---|---|
| **HLT Modules** | | | | | | | | |
| Morphological Analyser | +++ | +++ | | | ++ | | | +++ |
| POS disambiguator/tagger | ++ | ++ | | | | | | ++ |
| Named Entity Recognition | +++ | ++ | | | + | | | |
| Word Sense Disambig. | ++ | | +++ | ++ | ++ | | | |
| Term extraction | +++ | | ++ | | | | | |
| Shallow parsing | ++ | | ++ | | | | | |
| Syntactic analysis | ++ | | | | + | | | ++ |
| Semantic Analysis | +++ | | +++ | | | | | |
| Sentence synthesis and generation | +++ | | | + | ++ | | ++ | |
| Sentence Alignment | | | | | | +++ | | +++ |
| OCR (Optical Character Recognition) | ++ | | | +++ | + | | | |

The above table gives the definition for Punjabi BLARK and describes the type of resources needed but as mentioned by Maegaard et al. (2006) it does not give an indication of the size or any other characteristic of each type of resource. As cited by Maegaard et al. (2006), in a paper presented at the ELSNET ENABLER Workshop in Paris, Cieri et al. (2003) suggested that there must be a written language corpus of at least 100,000 words for a language as a core resource. Although there is no statistical evidence given as to why and how they set this word limit, but the number is not unrealistic especially if it involves core vocabulary which means it is feasible to create written language corpus for Punjabi of 100,000 words. In the section below, I tried to represent reasonable figures for each of the Punjabi language resources using the Arabic model as reference; the important thing to be noted is that there is no statistical and methodological evidence represented by Maegaard et al. (2006) that could justify the accuracy of the figures used in Arabic BLARK but since if the numbers are the minimum estimated value, I will use the figures as benchmark for Punjabi LRs.

### 5.1.1 Monolingual Lexicon

As mentioned in Maegaard et al. (2006), there should be a monolingual lexicon that includes a minimum of 40,000 stems with POS and morphology which could be used for semantic analysis; At least a list of 50,000 proper names is required for named entities (Maegaard et al. 2006). It can be inferred from the above table that monolingual lexicon is essential to build most of the HLT modules, but Punjabi does not have a monolingual lexicon, even Glottolog does not have any specific paper that could confirm the existence of a monolingual Punjabi dictionary. The main advantage of monolingual lexicon is that it would provide more comprehensive information about the Punjabi which can be used to include additional meanings of the word. The additional

information can be further embedded in the various NLP applications to get the grammatical behavior of the word. Absence of a monolingual lexicon can be considered one of the major gaps that has hindered the progress of NLP tools for Punjabi. It is good motivation to have a monolingual lexicon for Punjabi that could contribute to the advancement of above-mentioned applications.

### 5.1.2 Multi/Bilingual Lexicon

Maegaard et al. (2006) recommends the same size as monolingual lexicon, depending on application. As shown in the below table 11, there are multi bilingual dictionaries that cover a wide range of words from 7000 to 40,000 which can be used for building morphological analyzers, POS tagger and Named Entity Recognition for Punjabi. Given the large number of multilingual dictionaries available, covering a vast range of Punjabi words, and meeting up the minimum requirement set according to Maegaard et al. (2006), it is advocated to carry forward the research with the aim to build machine learning tools.

Table 10: Language resources for Multi/Bilingual Lexicon

| Name of Lexicon | Provider | Size | Other Information | Availability, price, manipulable |
|---|---|---|---|---|
| **Punjabi Kosh** (Shabdkosh.com) | Online | >7000 Words | Allows keyboard entry and there are also number of learning games and lessons | 1,1,2 |
| **Punjabi Shabdkosh** | Online | Comprehensive | | 1,1, R |
| **Sampadak** ("Sampadak") | Centre for Development of Advanced Computing (CDAC) | 40,000 words with GIST typing tools | | 1,2,2 |
| **GurShabad Ratanakar Mahankosh** (Singh Nabha) | Online | Comprehensive | | 1,1, R |
| **Punjabi University Students English-Punjabi Dictionary** (Rayall 2002) | Punjabi University | 31,000 entries | With grammatical information | 1,1, C |
| **English-Punjabi Topic Dictionary by Punjabi University** (Lehal 2009) | Advanced Centre for Technical Development of Punjabi Language Literature and Culture, Punjabi University | 3100 entries | Organized into more than eighty categories such as adjectives, nouns, food, fruits, animals, months etc. | |
| **Punjabi-English and English-Punjabi Dictionary by Jasjit Singh Thind** (Lehal 2009) | Online | Comprehensive | | 1,1, R |
| **Punjabi Encyclopedia and** | Online | Comprehensive | Online search engine used for searching for | 1,1, R |

| | | | | |
|---|---|---|---|---|
| **Gurbani Dictionary by Dr. Kulbir Singh Thind** (Lehal 2009) | | | any Punjabi word in Mahan Kosh Encyclopedia, Gurbani Dictionaries and Punjabi/English Dictionaries | |

### 5.1.3 Thesauri, Ontologies, Wordnets

The dictionary meaning of thesaurus is a book or electronic resource that lists words in groups of synonyms and related concepts. In NLP, it is useful to have lexical knowledge to know relations between the words; one of the ways to represent this knowledge is by using thesauri (Kilgarriff 2003). All types of NLP (Natural Language Processing) tasks need a thesaurus (D. V. Sharma 2011). When a user works on documents by using natural language like Punjabi, Hindi, Oriya, Bengali etc, the system needs to pick the best word to convey a specific nuance of meaning. Consider a example, the words: ਉਸਤਤ (Usatata), ਉਪਮਾ (Upamā), ਪ੍ਸ਼ਸਾ (Pśasā), ਸ਼ਲਾਘਾ (Śalāghā), ਵਿਡਆਈ (Viḍa'ā'ī) all means “praise”. Users searching for information or resources commonly define the same query using differing terms, if there is no control placed on subject keyword from a set of synonymous terms through the use of a thesaurus, users may not be able to locate all the resources that are relevant to their search (D. V. Sharma 2011). Ontology, in AI, is a specification of the meanings of the symbols in an information system. That is, an ontology is a description (like a formal specification of a program) of the concepts and relationships (Gruber 1993). Wordnet, on the other hand, is a large lexical database in which words are grouped together according to their similarity of meanings and is widely used by the NLP community. Major applications of WordNet include Word Sense Disambiguation (WSD), text categorization and generic text summarization etc

(Narang, Sharma, and Kumar 2013). All the three resources described above play an important role in building HLT modules like Word Sense Disambiguation, Term extraction, Semantic Analysis and shallow parsing.

One of the big challenges is to collect a huge collection of words along with their synonyms and antonyms from various sources (D. V. Sharma 2011). According to Maegaard et al. (2006), the minimum requirement for building a thesauri is a subject tree with 200-300 nodes for each domain and ontologies and wordnets should ideally be the same size as the lexicon. As can be seen in Table 12, there are two wordnets available for Punjabi. Punjabi University and Thapar University have been working under the active guidance of the WordNet team of IIT-Bombay for building IndoWordNet (Dash, Bhattacharyya, and Pawar 2017).

Table 11: Language resources for Thesauri, Ontologies, Wordnets

| Resource | Provider | Size | Other Information | Availability, price, manipulable |
|---|---|---|---|---|
| **Punjabi Wordnet** (Dash, Bhattacharyya, and Pawar 2017) | Punjabi University and Thapar University (Punjabi) | 30,682 words | Synset generation statistics of IndoWordNet Consortium. (Dash, Bhattacharyya, and Pawar 2017) | 1,1,2 |
| **Punjabi WordNet** (Puri, Bedi, and Goyal 2015) | CFILT (Centre For Indian Language Technology), IIT BomBay | 53,902 words | This database is used for checking the presence of stemmed words in a dictionary. (Puri, Bedi, and Goyal 2015) | 1,1,2 |

### 5.1.4 Unannotated Corpora

According to Maegaard et al. (2006), for an unannotated corpus, at least 100 million words are required. Unannotated corpora are used in HLT modules like Word sense disambiguation, sentence synthesis and generation and OCR. In table 13, it can be seen that the maximum word count for the Punjabi corpora is 15 million which is approximately one-tenth of the minimum requirement set by Maegaard et al. (2006). If I go along with the minimum numbers presented in Arabic BLARK, then there is still a lot of work that needs to be done for Punjabi to build a corpus with large datasets. Punjabi news websites like Punjabi Tribune (“Punjabi Tribune”), Daily Ajit (“Ajit - Punjab Di Awaaz”) should be referred to collect data for creating the Punjabi corpus.

Table 12: Language resources for Unannotated Corpora

| Resource | Provider | Size | Other Information | Availability, price, manipulable |
|---|---|---|---|---|
| **A Gold Standard Raw Text Corpus** ("A Gold Standard Punjabi Raw Text Corpus") | Central Institute of Indian Languages (CIIL) | 10,125,770 Words | Covering data from 5 different domains - Aesthetics, Science & Technology, Social Science, Commerce and Mass Media. | 3,2,3 |
| **Guru Granth Sahib** | Holy book of sikhism | 1430 pages with 511,874 words, 1,720,345 characters, and 28,534 lines. (D. S. Kaur and Singh 2015) | It contains hymns of 36 composers written in twenty-two languages in Gurmukhi script. (D. S. Kaur and Singh 2015) | Free for Research Use |
| **The EMILLE-CIIL Monolingual Written Corpora** ("The EMILLE Corpus") | ELRA | It contains a Punjabi written corpus of approximately 15,600,000 words. ("EMILLE Corpus Documentation") | For commercial use | 7500.00 € |

### 5.1.5 Annotated corpora

For the annotated corpora, as per Maegaard et al. (2006), 2 million words will meet requirement for most of the applications like POS tagger, Sentence boundary, Named Entity, Term Extraction and Word sense disambiguation. As far as I searched, I found only one annotated corpus which is used to detect emotion in Punjabi poems. There is a need to focus on the lack of annotated corpus which is one of the important language resources for any language.

Table 13: Language resources for Annotated Corpora

| Resource | Provider | Size | Other Information | Availability, price, manipulable |
|---|---|---|---|---|
| **Kāvi: An Annotated Corpus of Punjabi Poetry with Emotion Detection Based on 'Navrasa'** (J. R. Saini and Kaur 2020) | | 948 poems, written in Gurmukhi script, were classified into 9 emotion states. | This corpus was manually annotated | |

### 5.1.6 Parallel Multi Ling Corpora

As shown in table 9, parallel multilingual corpora are used for sentence alignment. Table 15 shows the current existence of parallel corpus available; English-Punjabi and Hindi-Punjabi. The development of Hindi-Punjabi parallel corpus has been misinterpreted as an alignment problem (Pardeep Kumar and Goyal 2010). The alignment problem is the next step of the research problem, if the parallel corpus is available of a particular language pair, automatic alignment for that parallel corpus can be done. To address the alignment issue, each text unit (sentence, clause or phrase) in the parallel corpus is represented by tags, which are fed into a programming algorithm that computes the alignment of units (Papageorgiou, Cranias, and Piperidis 1994).

Parallel corpus can also be used to build transliteration tools. As we know that in the Punjab area most of the government departments use Punjabi language to store their data, so the applications like transliteration or spell checker or translators will help them a lot to save the data on a click of a button. Given the availability of parallel corpora for Punjabi, it can be used to

develop such kinds of applications. Table 15 shows the current existence of parallel corpora available for Punjabi which could help researchers to carry forward their work in the related field to develop tools for Punjabi.

Table 14: Language resources for Parallel Corpora

| Resource | Provider | Size | Other Information | Availability, price, manipulable |
|---|---|---|---|---|
| **English-Punjabi Code-Mixed Social Media Content** ("English-Punjabi Code-Mixed Social Media Content – ELRA Catalogue") | ELRA | 893,615 parallel sentences of English-Punjabi | Covering data from 10 different domains - Agriculture, Culture, Health, Religion, Sports, Technology, Tourism, Education and Entertainment. | 1,1, R |
| **Hindi-Punjabi Parallel Corpus** (Pardeep Kumar and Goyal 2010) | Punjabi University | 50,000 sentences | Developed Existing Hindi-Punjabi Machine Translation System and Using Sentence Alignments | 3 |
| **Samanantar** (Ramesh et al. 2022) | IIT Madrass | 688 Sentence pairs | The collection contains a total of 49.7 million sentence pairs between English and 11 Indic languages, including Punjabi. | |

### 5.1.7 Multi modal corpora

In the given table, though multi modal data is available for Punjabi but still there are shortcomings when it comes to related research approaches and methodologies. There are only audio datasets, and lack of synchronized video and textual records may hinder the specific research needs. Standardized

procedures for conventions which can be used to record, code, annotate and interrogate are yet to be developed. Also, the datasets are limited in size, multi-million-word multimodal corpora do not exist as yet and are domain specific that is recorded in one discourse context.

Table 15: Language resources for Multimodal corpora

| Resource | Provider | Size | Other Information | Availability, price, manipulable |
|---|---|---|---|---|
| **General Conversation Speech Data in Punjabi** ("Pre-Labelled Datasets \| Structured Datasets for AI/ML") | FutureBeeAI | 30 speech hours | Audio dataset consists of different general conversations between native Punjabi people from Punjab state of India. | 3 |
| **BFSI Call Centre Speech Data in Punjabi** ("Pre-Labelled Datasets \| Structured Datasets for AI/ML") | FutureBeeAI | 10 speech hours | Audio dataset consists of different call centre conversations in the BFSI domain between native Punjabi people from Punjab. | 3 |

### 5.2 Gurmukhi Punjabi HLT Modules corresponding to written resources.

Table 17 classifies HLT modules that have been developed so far. Morphological analyzer and Morphological generator can help to build applications like Machine translation, spell checker and search engine (Zarkar 2022). A Punjabi Morphological Analyzer and Generator has been developed by the Advanced Centre for Technical Development of Punjabi Language, Literature and Culture, Punjabi University by Dr Mandeep Singh. The software has been included in the language CD launched by MCIT. Earlier IIIT Hyderabad had also developed a Punjabi Morph under the

Anusarka project (Lehal 2009). Beta version of online Punjabi Morphological Analyzer and Generator is also available by Aglsoft.

POS tagging which helps in text classification, information extraction, machine translation, named entity recognition and natural language generation (Ekbal and Saha 2013), has been developed for Punjabi by Dr Mandeep Singh Gill and Dr Gurpreet Singh Lehal at Advanced Centre for Technical Development of Punjabi Language, Literature and Culture, Punjabi University and available for free use on the University website (Lehal 2009). It is developed at IIIT, Hyderabad. Like, morphological analyser, beta version of online Punjabi POS tagger is available at Aglsoft.

Named Entity Recognition which is used for NLP tasks like machine translation, question-answering systems, indexing for information retrieval and automatic summarization is also available for Punjabi (A. Kaur, Josan, and Kaur 2009). The training set has been manually annotated with a NE tagset of 12 tags. Unlike English, Punjabi does not have capitalization which makes it challenging to develop such tools. Hence, Vishal and Gurpreet (2011) explains that they developed various Punjabi resources like list of prefix names, suffix names, proper names, middle names to implement NER for Punjabi. Further research can be carried forward to implement other NLP tools for Punjabi.

Moving forward to OCR, a robust, multi-font Gurmukhi OCR has been developed by Dr Gurpreet Singh Lehal and his team at Punjabi University Patiala. The OCR can recognize Gurmukhi text printed in any of the common Gurmukhi fonts and has more than 97% recognition accuracy at character level (Lehal 2009). Dr. Chandan Singh, Dr. Indu Chhabra and Dr. Renu Dhir have also been working on different phases related to development of Gurmukhi OCR (cited in Lehal 2009, 19). Lehal has also been working on the development of Gurmukhi OCR capable of processing bad quality documents (Lehal 2009, 20). The team of Anuj Sharma, Rajesh Kumar and

R.K. Sharma has been working at Thapar University for development of online Gurmukhi OCR system (Lehal 2009, 20). Dharamveer Sharma and his team at Punjabi University has developed a form recognition system for Gurmukhi, which can automatically detect and recognize the handwritten Gurmukhi text written on preprinted forms (Lehal 2009, 20).

Since there is significant work done in developing Punjabi HLT modules, but these modules are just prototypes which has been tested on small amount of data. To the best of my knowledge, I could not find any significant work done for modules like semantic analysis, term extraction, syntactic analysis, word sense disambiguation, shallow parsing and sentence alignment in Punjabi. It is a good motivation for researchers or scholars to carry forward the research and development to create future NLP tools and applications for Punjabi.

Table 16: HLT Modules

| HLT Modules | Name of the module | Provider | Other Information/References | Availability, price, manipulable |
|---|---|---|---|---|
| Morphological Analyser | Punjabi Morphological Analyzer and Generator (Lehal 2009) | Punjabi University | It has been developed by the Advanced Centre for Technical Development of Punjabi Language, Literature and Culture, Punjabi University by Dr Mandeep Singh. | **3** |
| | Punjabi Morphological Analyzer ("Online Punjabi Morphological Analyzer") | Aglsoft | The system is in BETA release and new words are being added. | 3 |
| | Punjabi Morph (Lehal 2009) | IIT Hyderabad | It was built under the Anusarka Project. | |
| POS disambiguator/tagger | Punjabi Part of Speech Tagger ("Online Punjabi Part of Speech Tagger") | Aglsoft | | 3 |
| | Speech tagger for Punjabi (Lehal 2009) | Punjabi University | It has been developed by the Advanced Centre for Technical Development of Punjabi Language, Literature and Culture, Punjabi University by Dr Mandeep Singh and Dr. Gurpreet Lehal | |
| Named Entity Recognition | Named entity recognition for punjabi | | A manually NE tagged Punjabi news corpus is used for evaluation which | |

| | | | | |
|---|---|---|---|---|
| | (A. Kaur, Josan, and Kaur 2009) | | has been developed from Punjabi newspapers available online. | |
| | **Named Entity Recognition for Punjabi Language Text Summarization[2] (Vishal and Gurpreet 2011) | | The paper explains the Named Entity Recognition System for Punjabi language text summarization. | |
| Spell Checker | Punjabi Grammar Checker ("Punjabi Grammar Checker") | Advanced Centre for Technical Development of Punjabi Language, Literature and Culture | The Punjabi grammar checker can detect and suggest rectifications for a few grammatical errors, resulting from the lack of agreement, order of words in various phrases etc., in literary style Punjabi texts. | 3 |
| | Spell Checker for Punjabi Language Using Deep Neural Network (G. Kaur, Kaur, and Singh 2019) | | A novel approach for spell checker for Punjabi language. | |
| Optical Character Recognition (OCR) | Online OCR Converter ("Converter") | | Online OCR Converter services that convert images, scans doc images text into editable files by using OCR Technology (Optical Character Recognition). | |
| | Multi Font Gurmukhi OCR by Dr. Gurpreet Singh | Punjabi University | The OCR can recognize Gurmukhi text printed in any of the common | |

2 **The HLT Module is not available but the concept of developing the HLT module has been proposed in the paper.

| | | | | |
|---|---|---|---|---|
| | Lehal (Lehal 2009) | | Gurmukhi fonts with more than 97% recognition accuracy at character level. | |
| | Online Gurmukhi OCR system. (Lehal 2009) | The team of Anuj Sharma, Rajesh Kumar and R.K. Sharma have been working at Thapar University | An online OCR system recognizes the handwritten text as it is being written on a tablet. | |

### 5.3 Gap Analysis for written resources along with their corresponding HLT modules

After having compared, the existing modules and Language Resources available for Punjabi, it can be analyzed that though the Punjabi corpora is scarce a significant amount of work has been done and significant amount of work is in progress, taking that into consideration it is good motivation for researchers and developers to create basic modules required for HLT. As depicted in the table 1, it clearly indicates 10 out of 12 modules are dependent on monolingual lexicon, it is the call of hour to create that resource. Word sense disambiguation is one of the important modules in NLP which requires mono lexicon as well as thesauri, and Punjabi lacks both language resources. Here, we have observed that Punjabi does not have monolingual lexicon and thesauri but instead of starting from scratch to build this language resources, BLARK as a model should be used to examine interdependent language resource. There already exists a bilingual lexicon for Punjabi with almost 40k words. Also, there are existed printed dictionaries of Punjabi, their digital originals could be turned into a monolingual lexicon, or OCR could be used to that end. In addition, there is Punjabi

Wordnet with 53,902 words which could be another starting point to devise at least single monolingual lexicon. Likewise, Punjabi Wordnets can be used to build thesauri. Thus, one could use some existing BLARK technologies to fill in the current gaps.

As shown in table 17, there are some of the modules available but there is no quality assessment done on these modules to know if they are reliable and could be used as a foundation to build other tools that are dependent on the existing tools. Machine learning requires annotated datasets for specific tasks such as detecting hate speech but again the challenge is there are no annotated datasets available for Punjabi. For a good language model, it requires a large collection of text (Howcroft and Gkatzia 2022), which is not available in abundance if I talk about Punjabi in particular. Let's consider GPT3, for instance, it is trained on 45TB of text data in English (Howcroft and Gkatzia 2022). Training and designing such a model would require a large amount of trained data.

It can be inferred that data resources play a crucial role in the development of any NLP tool, the root problem is not the lack of applications and tools but the lack of reliable and authenticated data resources, which creates the gap between a language and NLP for that language. As stated in the previous section, there must be a written language corpus of at least 100,000 words for a language as a core resource (Strassel, Maxwell, and Cieri 2003), and the same needs to be done for Punjabi. Punjabi has corpora but the data is not parsed and processed systematically to sustain larger applications and tools.

As the data resources are available for mostly all the languages, even ELRA inventory contains more resources than the tools and applications. The other fact is the variety of applications and modules is so much greater, given to the languages like Hindi, Marathi, Tamil or Bengali even if the applications are build, they might or might not come in usage, hence, there is no point

spending money and time in development of applications or modules which don't have practical application within the community of that language. Even if the resources are available, they are not easily and readily available within a legal framework for the researchers and scholars. Most importantly, the data in the above tables do not address the coverage in terms of quantity, quality and adequacy to technological purposes.

### 5.4 BLARK content for Spoken Resources

The section particularly focuses on audio and visual data required by various speech applications like voice dictation, transcriptions, speech recognition, telephony servers etc. The main idea is to provide a minimum basic kit to researchers and developers that can help them to build such applications and techniques.

I am referring to matrix charts mentioned in the Arabic BLARK (Maegaard et al. 2006) in which speech technologies and corresponding HLT modules are marked with importance. The matrix in table 18 shows the results (+ = relevant; ++ = important; +++ = essential) taken from Arabic BLARK (Maegaard et al. 2006). Based on this matrix one can determine which components serve most applications, and which data are most needed for most applications, i.e., which elements should be part of the BLARK.

Following the chart, I would explain what the minimum resources are required as mentioned in the Arabic BLARK to build the basic tools and applications and what is already available for Punjabi.

| | Dictation | Telephony speech applications | Embedded speech recognition | Transcription of broadcast News | Transcription of conversational speech | Speaker recognition | Dialect / language identification | "Emotion" Identification | Speaker Adaptation | Lips movement reading : | 'topic' detection, segmentation, topic boundaries | Speaker 2 speaker mapping | "Emotion/ Prosody" output | – Text to Speech (inc. formatted data e.g. databases) | – Synthesis by Concatenation : | – Customization to different voices | – Generation Lips Movement |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | | | | |
| Acoustic models | +++ | +++ | +++ | +++ | +++ | ++ | +++ | +++ | +++ | +++ | +++ | ++ | +++ | +++ | +++ | +++ | +++ |
| Language models | +++ | ++ | ++ | +++ | +++ | | ++ | | | | | | ++ | +++ | | | |
| Pronunciation lexicon | +++ | +++ | +++ | +++ | +++ | | | | | | | ++ | | +++ | | | |
| Lexicon Adaptation | + | + | + | + | + | | | | | | | ++ | | +++ | | | |
| Phoneme Alignment | + | + | + | + | + | + | ++ | | | | | ++ | | | | | |
| Prosody recognition | + | + | + | + | + | + | + | +++ | + | | | ++ | | | | | |
| Speech Units Selection | | | | | | | | | | | | | +++ | +++ | | | |
| Prosody prediction | | | | | | | | | | | | | +++ | +++ | | | |
| segmenter Speech / Silence: | ++ | + | ++ | ++ | ++ | + | ++ | ++ | + | + | + | | + | | | | |
| Sentence boundary detection: | + | + | + | + | + | + | + | ++ | + | + | + | | ++ | +++ | | | |
| Dialect / language identification | + | + | + | + | + | + | + | + | + | + | + | | | + | | | |
| (word) Boundary identification, | + | + | + | + | + | + | + | + | + | + | + | | ++ | | | | |
| Speech /Non-speech (music) detection: | + | + | + | + | + | + | + | ++ | + | + | + | | | | | | |
| Speaker recognition/identification | + | + | + | + | + | + | + | + | + | + | + | ++ | | | | | |
| "Emotion" Identification | + | + | + | + | + | + | + | | + | + | + | ++ | ++ | | | | |
| Speaker Adaptation | ++ | + | ++ | + | ++ | + | + | + | + | + | + | ++ | | + | | | |
| Lips movement reading | | | | | | | | | | +++ | | | | | | | |
| Morphological comp.(infl, deriv., stemm., diacritic,...) | ++ | + | + | ++ | ++ | | | | | | | | | +++ | | | |
| POS disambiguator/tagger | ++ | + | + | ++ | ++ | | + | | | | | | | +++ | | | |
| Diacritizer | | | | | | | | | | | | | | +++ | | | |
| Named Entity Recognition | ++ | + | ++ | ++ | ++ | | | | | | | | | ++ | | | |
| Word Sense Disambig. | | | | | | | | | | | | | | ++ | | | |
| Shallow parsing | ++ | + | + | ++ | ++ | | | | | | | | | ++ | | | |
| Syntactic analysis comp. | ++ | + | + | ++ | ++ | | | | | | | | | ++ | | | |
| Sentence synthesis and generation | | | | | | | | | | | | | + | ++ | | | |
| Semantic Analysis | + | | | + | + | | | | | | + | | + | + | | | |

Figure 7: Speech language applications and corresponding HLT modules, marked with importance.

### 5.4.1 Voice dictation

Voice dictation is another term for voice to text or speech to text. Speech-to-Text Recognition (STR) technology synchronously transcribes text streams from speech input (Trivedi et al. 2018). As described by Maegaard et al (2006), the minimum requirement is transcribed recordings of 50-100 speakers, speaking 20 minutes each and 10 additional speakers for testing purposes and a written corpus of a few million words.

The models trained on less data, between 50 hours to 160 hours, still predict words with relatively high precision (S. Bansal et al. 2018). With at least 50 hours of data, it gives 50% accuracy.

### 5.4.2 Telephony speech applications

The goal of telephony speech applications is to make the telephonic conversations as natural as possible. For example, a user can dial a speech banking application and can make a request to transfer money from one account to another, the voice request should be able to activate database query and on successful completion of the request it should convert back the request to voice message to deliver the information to the user (Cochrane 2004). To develop such a kind of speech application like voice assistant or speech recognition software, there is a need for speech recognition data for Punjabi language to train machine learning algorithms.

According to (Maegaard et al. 2006), preferably data about 500-1000 speakers uttering around 50 different sentences based on the type of application, if it is for banking or education or for some other domain, would be enough to serve the purpose. The speech material can be restricted to a list of common words or phrases or numeric numbers, but it can also contain sentences and paragraphs or continuous speech for interactive speech dialogues.

### 5.4.3 Speaker recognition

Speaker recognition is used to recognize a person from a spoken phrase (Campbell 1997). It is either used to identify a particular person or to verify a person's claimed identity. Due to the low cost, speaker recognition systems are gaining increasing importance in the various fields, such as in forensics, remote banking services, biometrics solutions etc.

Speech corpora is the basic requirement for the development and evaluation of speaker recognition systems (Campbell 1997). Considering Punjabi a less resource language, it has limited speech corpora which is not sufficient for development of robust speaker recognition systems. Punjabi doesn't have a large speech corpora and after examining the Arabic BLARK model (Maegaard et al. 2006) and (Campbell 1997), it could be concluded that at least an audio corpus of about 500 speakers, amount of speech 3 minute per speaker, along with 100 speakers for testing, is required.

### 5.4.4 Embedded speech recognition

With the development of embedded systems, speech recognition technology based on embedded systems has become a new and important direction in this field (Yehui, Enliang, and Liqin 2016). One may use the desktop data from voice dictation resources (Maegaard et al. 2006), for embedded speech recognition systems.

### 5.4.5 Broadcast News Speech Corpus

Applications related to transcription of Broadcast News require transcribed audio and video data. Analyzing language models of Arabic (Maegaard et al. 2006), it can be inferred that transcribed

audio data of about 50-100 hours and non-transcribed data of about 1000 hours can be useful along with the written corpus from newspapers, press releases and podcasts.

### 5.4.6 Conversational speech

Maegaard et al (2006) did mention that language should have telephonic speech data similar to that of CallHome American English (Speech developed by LDC) but the report of Arabic BLARK did not mention how this data could be useful for NLP speech applications or tools. A Conversational speech dataset is considered to be a powerful tool in NLP as it can help in training machine learning algorithms to understand, process and generate human language. Such types of datasets can be used to improve the performance and scalability of NLP models like machine translation and chatbot systems ("The Benefits Of A Conversational Speech Dataset - Way With Words").

Just like LDC focused on providing resources needed for telephonic applications and speech recognition research (Godfrey 1994), following which LDC managed to create three types of telephone collection projects: CallHome, CallFriend and Switchboard-2. CallHome supports large vocabulary conversational speech recognition with transcribed collections of audio and lexical resources (Cieri and Liberman 2002). Similar type of speech dataset can be helpful for Punjabi as well.

### 5.4.6 Dialect / language identification

As mentioned above, CallFriend corpus consists of telephone conversations for language identification research.The corpus is documented describing the information of the speaker and call information (channel quality and number of speakers). It contains un-transcribed audio in Arabic, Canadian French, English, Farsi, German, Japanese, Korean, Mandarin, Russian, Spanish and

Vietnamese. It does support some of the Indian languages like Hindi and Tamil (Cieri and Liberman 2002). Although most of these calls have not yet been transcribed, there is a growing body of transcription for Spanish, Mandarin, Farsi, Korean and Russian to support speech recognition. Similarly, speech corpus for Punjabi can be created with a set of varieties of Punjabi dialects.

### 5.4.7 Speech Synthesis Corpus

Speech synthesis is the artificial production of human speech. Developing text to speech (TTS) systems for a particular language requires a specialized speech corpus for that language. Training speech synthesis methods requires about 30 hours of good quality audio recordings from a single speaker in a noiseless environment, and an accurate transcription of these recordings. For less resourced languages, obtaining such a corpus is the major obstacle to development of TTS solutions (Veaux, Yamagishi, and MacDonald 2017). The other possible way to get high quality data can be to transcribe existing audio recordings, but that requires a lot of manual work. Approximately 10 hours of work are required to transcribe one hour of raw audio from scratch (Veaux, Yamagishi, and MacDonald 2017).

Text to speech (TTS) systems are necessary for any language to ensure accessibility and availability of digital language services (Veaux, Yamagishi, and MacDonald 2017). Synthetic speech may be used in several applications such as announcement or warning systems, reading machines for the blind or electronic-mail readers (Lemmetty 1999, 47-50). Hence, it is good motivation to develop such corpora for Punjabi language.

### 5.4.8 Written corpus for speech technologies

Written corpus is used in order to derive phonetic lexicon and language models, same as for written technologies (Maegaard et al. 2006). The primary goal for written corpus is to build phonetically rich corpus that could provide a certain linguistic richness and could assure that speech synthesis system produces at least a natural intelligible sound (Amrouche et al. 2021). It also ensures that all the sentences or words are recorded by the speaker while designing or developing any speech corpus (Amrouche et al. 2021). When creating Arab speech synthesis, an initial corpus was designed which consisted of prose and poetry, scientific texts, text from newspapers, words from The Quran and Hadith (Amrouche et al. 2021). A similar corpus could be designed for Punjabi as well, mainly consisting of an unannotated corpus of minimum of 300 million words, depending on the requirement of the application (Maegaard et al. 2006).

Table 17: Available speech resources for Punjabi

| Name of Resource | Provider | Size | Other Information | Applications Possible | Availability, cost, Adaptability |
|---|---|---|---|---|---|
| Punjabi Raw Speech Corpus<br><br>("Punjabi Raw Speech Corpus") | Central Institute of Indian Languages | 101:09:28 Hours \| 65.5 GB | 467 Speakers and 76,230 Audio Segments | | > 3000 |
| Speech Corpus for Punjabi<br><br>("ELRA - ELRA-U-S 0062 : Speech Corpus for Punjabi") | ELRA | | Comprises recordings of syllables, frequent words, digits, phonetically rich sentences, prosody rich sentences, domain-specific texts and news texts. | Speech synthesis | Freely available for research and development |
| Punjabi Speech data<br><br>(Dhanjal and Singh 2022) | | Corpus consists of 401 daily used words and a total of 16,793 audio files are prepared, each file containing 5 words. | Punjabi speech data is recorded using SmartRecorder software. Both males and females were selected from different age groups. | ASR | |
| OTS Dataset<br><br>("Pre-Labelled Datasets \| Structured Datasets for AI/ML") | FutureBee AI | Contains 30 speech hours of audio/speech data. | The participants in the collection are males and females from the age group of 18 to 70 years, with different accents to keep the speech dataset diverse and unbiased. | ASR, Chatbot, TTS, Speech Analytics, Language modeling, Conversational AI | Freely available for research and development |

Table 18: HLT Modules for spoken resources

| HLT Modules | Name of the module | Provider | Other Information/References | Availability, price, manipulable | |
|---|---|---|---|---|---|
| ASR | Automatic Speech Recognition (Dev, Sharma, and Agarwal 2021) | 3 Speakers, 200 words | The efforts made by various researchers to develop Punjabi ASR for Indian languages have been analyzed and then their applicability to Punjabi language has been discussed so that concrete work can be initiated for Punjabi language. | ASR | 3 |
| | Kaur and Singh (2016a)<br><br>(Dev, Sharma, and Agarwal 2021) | 3 Speakers,43 words (noise free) 3 speakers,53 words (noisy) | Optimizing feature extraction techniques using phone based modeling on connected words for Punjabi ASR. | ASR | 3 |
| | Kumar and Singh (2016a)<br><br>(Dev, Sharma, and Agarwal 2021) | Punjabi live speech 462 sentences, 1213 words | The focus is to develop Spontaneous Speech Recognition. | Automatic Spontaneous Speech Recognition | 3 |
| | Kadyan et al. (2017a)<br><br>(Dev, Sharma, and Agarwal 2021) | 25 speakers, 5000 words, 32 hrs | | | 3 |

| | | | | | |
|---|---|---|---|---|---|
| | Kadyan et al. (2017b)<br><br>(Dev, Sharma, and Agarwal 2021) | Training 24 speakers ,58700 words,46 hrs 40 min Testing 13 speakers, 6100 words, 5hrs30 min | | | 3 |
| | Kadyan et al. (2018)<br><br>(Dev, Sharma, and Agarwal 2021) | Continuous speech corpus 3611 sentences, 13 speakers | | | 3 |
| | Guglani and Mishra (2018)<br><br>(Dev, Sharma, and Agarwal 2021) | 24 speakers, 0-9 digits in Malwai dialect,240 hrs | Punjabi speech recognition model | | 3 |

**5.4 Gap Analysis for spoken resources along with their corresponding HLT modules.**

As shown in the table 18, Punjabi has a raw speech corpus of almost 100 hours of speech. Given the amount of speech data available, it is a pathway for researchers to direct their research and development towards creating and designing speech technologies for Punjabi.

If we compare English or Chinese, those are languages that are trained with massive amounts of speech and text information, however, as it can be inferred from the table 19 that Punjabi does not have such resources or stable orthography. There is a need for realistic Punjabi

corpora to fuel reproducible and replicable language studies at the phonetic, lexical and syntactic levels. Most of the speech-based tools work on intensive data and depend on good quality data to be provided to these models to ensure good output is generated. Usually, researchers are focusing on speech recognition or speech synthesis, but these are not the only two types of speech resources that a researcher might be interested in.

There are already a 100 hours of spoken Punjabi recordings, which is quite a good starting point, but it needs to be properly transcribed and aligned. Hence, the priority is to improve speech corpora in order to develop speech technologies for Punjabi.

### 5.5 Challenges for creating data resources.

Although, it has been analyzed that lack of resources is one the main reasons why Gurmukhi Punjabi is behind when it comes to NLP tools, what are significant reasons or challenges that Punjabi is facing when it comes to the development of resources or corpus. In the following sections I will discuss some of the major challenges.

#### 5.5.1 No capitalization and no contractions

There are no capital letters in Punjabi. For example, in English, in proper names- Peter else peter but in Punjabi its only written as “ਪੀਟਰ”. It can be a problem while tokenizing as it would be difficult to distinguish nouns and proper names from the other parts of speech. NER, one of the important concepts in NLP, detects the boundaries of the sentences by analyzing capital letters. Also**,** there is no apostrophe or other contractions in Punjabi to identify possessive nouns. Lack of capital letters and apostrophe makes it challenging to build such modules. I won’t say it is difficult just because there is a way it could be done for English, which is the dominant language, and a lot

of work has been done for it. Every language has some unique features and so does Punjabi. All that needs to be done is to have different approaches and methods to achieve the desired goal of creating linguistics resources for Punjabi.

### 5.5.2 Cost and deployment of Linguistic Resources

Capital and infrastructure are another factor why it is hard to implement the given directions and ideas in real-time environments. It's a long-term project and not a change that occurs overnight. Keeping this in mind, one needs to be prepared to handle the time, resources, and capital involved in building, testing, and deploying the system in the market. Another drawback can be that training language models takes considerable time and expertise. Collecting enough language resources or effectively making do with the available ones may not be cheap. All in all, manual development would be really expensive. According to the report published by the Central Institute of Indian Languages in September 2018 (Choudhary 2018), it approximately cost around 10 million rupees to create a labeled corpora for Hindi. Hindi being the national language of the country, it is not that hard to get funds from the central government but when it comes to languages like Punjabi, the same does not hold true for it. Then there is the additional cost of storage and management of corpus. For creating speech corpora particularly, the cost of text preparation time is included as well. A text is first prepared and then recorded. After the data is recorded, then there is a process of transferring the data on to computers, segmenting and storing the data. With a revenue deficit of 9%, as per the financial year report of 2022, Punjab government can hardly grant funds for Punjabi Linguistic resources.

### 5.5.3 Lack of linguistic expertise

Punjabi is not a universal language, so expecting users from different geographies to have the same level of proficiency is unrealistic. For labeled corpora or annotated corpora, there is a need for an annotator who is trained in linguistics and having an expert knowledge in the language concerned. With drift in the STEM field, students are more into engineering or medical studies, thus, in the longer run it would be hard to find someone who has linguistic expertise of a language.

# CHAPTER 6: LANGUAGES RESOURCES AND TOOLS AVAILABLE FOR OTHER INDIAN LANGUAGES

According to Ethnologue, India is the fourth most linguistically diverse country in the world after Papua New Guinea, Indonesia and Nigeria and is home to 1,300 million (Eberhard, Gary F., and Charles D. 2023) people which is almost 18% of the world population, but still the information technology advancement is lagging by decades compared to other languages of the world like English, Dutch and Arabic. This may be due to various factors but one of the primary factors is fewer language resources required for the development of such technology. In the following section, a subset of major Indian languages has been chosen to evaluate the number of LRs for each language. And since Arabic and Dutch have been able to create more NLP resources with the use of BLARK, I am comparing the Indian languages with these two languages to analyze to what degree the Indian languages lag digitally. The languages covered by this report are Arabic, Dutch, Hindi, Marathi, Bengali, Tamil and Telugu.

## 6.1 Existing Resources for the Individual languages

The following overview is divided into two parts: Table 21 and 22. Table 21 shows the coverage of LRs for the chosen languages and 22 shows the available modules of those languages. Since requirements regarding resources and tools may differ depending on the field, availability is stated in terms of degree. I have distinguished between three levels:

+ a single resource/tool is available.

++ various resources/tools are available.

+++ various resources /tools are available; cover different fields/language pairs/language features.

Table 19: Available language resources for these languages

| **Resource** | **Arabic (ELRA/ MEDAR)** | **Dutch (ELRA/CLARIN)** | **Hindi (ELRA/ Other)** | **Tamil (ELRA/ Other)** | **Telugu (ELRA/ Other)** | **Marathi (ELRA/ Other)** | **Bengali (ELRA/ Other)** | **Punjabi (ELRA/ Other)** |
|---|---|---|---|---|---|---|---|---|
| Thesaurus/Wordnet | -/+++ | +/++ | -/++ | +/++ | -/+ | -/+ | -/+ | -/++ |
| Unannotated corpus | +++/+++ | -/++ | -/+++ | -/+++ | -/+ | -/++ | -/+++ | +/+++ |
| Multimedia/Multimodal corpus | +/+++ | -/++ | -/++ | –/+ | – | – | – | +/+ |
| Multilingual lexicon | +++/+++ | +++/+++ | +++/+ | +/+++ | -/+++ | -/+++ | -/+++ | –/+++ |
| Annotated corpus | ++/+++ | +/+++ | ++/+++ | -/+++ | – | -/++ | -/+++ | –/+ |
| Speech corpus | +++/++ | +++/++ | +++/+++ | +++/+++ | +/+ | +/+++ | +++/+++ | +/+++ |
| Speech corpus (Transcript) | +++/+ | +++/++ | +++/++ | +++/+ | -/+ | -/+ | +/+ | ++/-- |
| Monolingual lexicon | +++/+++ | +++/+ | -/++ | –/+++ | – | – | -/++ | – |
| Terminology Databases | +++/+++ | +/- | +++/+ | –/+++ | – | – | -/+ | – |
| Monolingual corpora | +++/- | +++/+ | ++/+++ | +++/++ | +/+ | +/+++ | ++/+ | ++/– |
| Multilingual corpus | ++/- | +++/+++ | +/+ | + | +/- | – | -/++ | – |
| Multilingual Parallel/comparable corpora | +++/+++ | ++/+++ | ++/+++ | +/+ | +/++ | -/+++ | +/+++ | +++/++ |

[1] Coverage for Arabic includes elements that are available on ELRA as well as MEDAR.

[2] Other Sources for Thesaurus/Wordnet are: **Dutch** - (Spyns and Odijk 2013); **Hindi** - ("Hindi WordNet"), (Bhattacharyya, Pushpak, Pande, Prabhakar, and Lupu, Laxmi 2008); **Tamil** -

(Rajendran, 2014), (ThaniThamizhAkarathiKalanjiyam [2020] 2020); **Telugu** - ("TDIL-DC :Indo Wordnet"); **Marathi** – ("TDIL-DC :Indo Wordnet"); **Bengali** – ("TDIL-DC :Indo Wordnet"); **Punjabi** - (Dash, Bhattacharyya, and Pawar 2017), (Puri, Bedi, and Goyal 2015).

[3] Other Sources for Unannotated Corpus are: **Hindi** - (Kapoor et al. 2022), (A. Verma et al. 2021), (Choudhary 2021); **Tamil -** (Choudhary 2021), (Neechalkaran [2019] 2022), ("Tamil News Dataset"); **Telugu -** (Choudhary 2021); **Marathi -** (Choudhary 2021), ("Digital Corpus of Old Marathi (DCOMA)"); **Bengali -** (Choudhary 2021)**,** (Mithun Biswas et al. 2017)**,** (Mridha et al. 2021)**; Punjabi -** (Choudhary 2021)**,** ("A Gold Standard Punjabi Raw Text Corpus")**,** Guru Granth Sahib.

[4] Other Sources for Multimedia/Multimodal Corpus are: **Hindi** - (Parida, Bojar, and Dash 2019), (Meetei et al. 2023); **Tamil** - (Chakravarthi et al. 2021); **Punjabi** - ("Pre-Labelled Datasets | Structured Datasets for AI/ML").

[5] Other Sources for Multilingual lexicon are: **Hindi** - ("Shabdkosh.com"), ("Collins Hindi Dictionary | Translations, Definitions and Pronunciations"), ("Hindi Dictionary Online Translation LEXILOGOS"); **Telugu** - (Brown 1903), ("Shabdkosh.com"), ("Hindi – Hindi -Telugu Dictionary – Neelkamal Publications Pvt. Ltd"); **Tamil** - (Fabricius 1972), ("Tamil Hindi Dictionary | Lexicool"), (Percival 1993); **Bengali** - (S. Biswas 2004), (Mitra et al. 2017), (M.A 2009); **Punjabi** - ("Shabdkosh.com"), ("Sampadak"), (Singh Nabha).

[6] Other Sources for Annotated Corpus are: **Arabic** - (Zeman et al. 2012); **Dutch** - (Zeman et al. 2012); **Hindi** - (Zeman et al. 2012), (Aggarwal, Ghosh, and Mamidi 2020), (A. Prasad and Sharma 2020); **Tamil** - (Zeman et al. 2012), (Chakravarthi et al. 2021); **Marathi** - (Kulkarni et al., n.d.); **Bengali -** (Zeman et al. 2012)**,** (Chaudhury et al. 2017)**,** (Monisha Biswas and Hoque 2019)**; Punjabi -** (J. R. Saini and Kaur 2020).

[7] Other Sources for Speech Corpus are: **Hindi** - Choudhary 2021), (Yadavally [2019] 2021), ("Openslr.Org"); **Tamil -** (Choudhary 2021), (A, Pilar and G 2022), ("Tamil Speech Recognition Corpus (Desktop)-Speechocean"); **Telugu -** (Choudhary 2021), ; **Marathi -** (Choudhary 2021), (Lahoti, Mittal, and Singh 2023); **Bengali -** (Choudhary 2021)**,** (Mantoro et al. 2021)**,** (Sodimana et al. 2018)**; Punjabi -** (Choudhary 2021), ("Punjabi Raw Speech Corpus"), (Dhanjal and Singh 2022).

[8] Other Sources for Speech Corpus (Transcript) are: **Hindi** - (Yadavally [2019] 2021), ("Openslr.Org"); **Telugu** - ("Openslr.Org"), **Bengali** - (Sodimana et al. 2018).

[9] Other Sources for Monolingual Lexicon are: **Hindi** - ("Hindi to Hindi Dictionary - Apps on Google Play"), ("Hindi Dictionary - Online Hindi to Hindi Devanagari Words Dictionary"); **Telugu** - ("Telugu Monolingual PoS Tagged Text Corpus ILCI"); **Tamil** - (R [2013] 2023), (Madras 1924), ("Tamil to Tamil Dictionary | Tamil Translation Services | Tamil English Dictionary - Ilearntamil"); **Bengali** - ("Offline Bangla Dictionary (E2B & B2E)" 2022), ("Bengali Dictionary Online Translation (Bangla) LEXILOGOS").

[10] Other Sources for Terminology Databases are: **Hindi** - ("Hindi Stop Words and Sentiment Lexicons"); **Tamil** - ("All_tamil_nouns" [2020] 2023), ("Open-Tamil/Solthiruthi/Data at 5eb9fb1447fe021ca47e2cc4605f7111e6b1088f · Ezhil-Language-Foundation/Open-Tamil"); **Bengali** - ("Stopwords Bengali (BN)" [2016] 2023).

[11] Other Sources for Monolingual Corpus are: **Hindi** - ("IIT Bombay English-Hindi Parallel Corpus"), ("Hindi Monolingual Chunked Text Corpus ILCI"), ("Download Corpora Hindi"); **Telugu** - ("A Gold Standard Telugu Raw Text Corpus"); **Tamil** - ("A Gold Standard Tamil Raw Text Corpus" n.d.), ("Tamil - Language Corpus for NLP" n.d.); **Marathi** - (Lahoti, Mittal, and Singh 2023); **Bengali** - (Mumin et al. 2014)

[12] Other Sources for Multilingual Corpus are: **Hindi** - (J. Kumar, Rakhra, and Dubey 2022); **Bengali** - (Mumin et al. 2014), (Publications 2019)

[13] Other Sources for Multilingual Parallel/comparable Corpora are: **Hindi** - (Ramesh et al. 2022), (Pardeep Kumar and Goyal 2010), (P. Singh [2020] 2022); **Tamil -** (Ramesh et al. 2022), ("English-Tamil Parallel Corpora"); **Telugu -** (Ramesh et al. 2022), ("Telugu-English Translated Parallel Corpora for Religious Domain"); **Marathi -** (Ramesh et al. 2022), (Lahoti, Mittal, and Singh 2023); **Bengali -** (Ramesh et al. 2022)**,** (Nowshin, Sultana Ritu, and Ismail 2018)**,** ("Bengali-English Translated Parallel Corpora for Tourism Domain")**; Punjabi -** (Ramesh et al. 2022), (Pardeep Kumar and Goyal 2010).

Table 20: Available modules for these languages

| Module | Arabic (MEDAR/ Other) | Dutch | Hindi | Tamil | Telugu | Marathi | Bengali | Punjabi |
|---|---|---|---|---|---|---|---|---|
| Morphological analyser | +++ | + | +++ | +++ | +++ | +++ | +++ | +++ |
| POS tagger | +++ | +++ | +++ | +++ | +++ | +++ | +++ | ++ |
| Diacritizer | -/+ | – | – | – | – | – | – | – |
| Stemmer | ++/- | +++ | +++ | +++ | ++ | +++ | ++ | +++ |
| Sentence Boundary Detection | +/- | ++ | + | — | — | ++ | +++ | — |
| Named Entity Recognition | -/+++ | +++ | +++ | +++ | +++ | +++ | +++ | ++ |
| Shallow parser | -/++ | ++ | ++ | + | + | + | + | + |
| Spell checker | -/+++ | ++ | +++ | +++ | + | +++ | ++ | ++ |
| Parsers and grammars | ++ | ++ | +++ | +++ | — | ++ | ++ | — |
| Font converters | -/+ | +++ | — | — | — | —- | ++ | + |
| Sentence Alignment | -/+++ | ++ | +++ | + | — | —- | + | —- |
| Text generation | -/+ | – | — | — | — | —- | —- | —- |
| Text annotation tools | +++/++ | – | — | — | — | —- | —- | —- |
| Machine translators | -/+++ | +++ | +++ | ++ | +++ | +++ | +++ | ++ |
| Token detection | -/+ | ++ | ++ | ++ | — | —- | +++ | —- |
| OCR | -/+++ | +++ | +++ | +++ | +++ | +++ | | +++ |

Coverage for Arabic includes elements that are available on ELRA as well as MEDAR.

[2] Other Sources for Morphological Analysers are: **Dutch** - ("ELG - Morphology Analyzer and Generator")**; Hindi -** ("Downloads | Siva Reddy"), ("Hindi Morphological Tagger"), (Goyal and Lehal 2008); **Tamil** - (Sarveswaran, Dias, and Butt 2021), (Akilan 2012), (Kumar, Jayashree, and Manimegalai 2020, 801–809); **Telugu** - (Burla and Tech 2019), (Sai Kiranmai et al. 2010), (D. S. Ghosh 2014); **Marathi** - (Lahoti, Mittal, and Singh 2023), (Gawade et al. 2013); **Bengali** - (Faridee 2009), (P. Das and Das 2013), (Gelbukh 2023, 13451:595–607); **Punjabi -** (Lehal 2009), ("Online Punjabi Morphological Analyzer").

[3] Other Sources for POS tagger are: **Dutch** - (Poel and Boschman 2008)**,** (Miltenburg [2016] 2022)**,** (ivdnt.org n.d.)**; Hindi -** (Shrivastava and Bhattacharyya 2008)**,** (Garg, Goyal, and Preet 2012)**,**

(Joshi, Darbari, and Mathur 2013)**; Tamil -** (Sarveswaran and Dias 2021), (Ramanathan, Chidambaram, and Patro, n.d.), (V. et al. 2009); **Telugu** - (Badugu 2014), (Suresh, n.d.), (Ganesh 2006), **Marathi** - (Lahoti, Mittal, and Singh 2023), (Jyoti Singh, Joshi, and Mathur 2013); **Bengali -** (Raha et al. 2020), (Hoque and Seddiqui 2015), ("Awesome-Bangla" [2017] 2023); **Punjabi -** ("Online Punjabi Part of Speech Tagger"), (Sanjeev Kumar Sharma and Gurpreet Singh Lehal 2011).

[4] Other Sources for Diacritizer are: **Arabic** - (Darwish, Mubarak, and Abdelali 2017)**.**

[5] Other Sources for Stemmer are: **Dutch** - (Jonker, 2020), ("Dutch Stemmer" 2006), ("Dutch Keyword Stemmer - SearchWP" 2015); **Hindi** - (Mishra and Tech 2012), (sainimohit23 [2019] 2022), (A. K. Pandey and Siddiqui 2008); **Tamil** - (Rajalingam [2010] 2023), (Thangarasu and Manavalan 2013), (Pan, Chen, and Nguyen 2012, 7198:197–205); **Telugu** - (Raju and Sreenivasulu 2022), (Department of CSE, CMR college of Engineering & Technology Hyderabad, India and Swapna* 2019), **Marathi** - (Lahoti, Mittal, and Singh 2023), (P. Pandey, Amin, and Govilkar 2015); **Bengali -** (S. Das, Pandit, and Naskar 2020), ("BanglaKit Bengali Stemmer" [2017] 2023); **Punjabi -** (Puri, Bedi, and Goyal 2015), (Tiwari et al. 2022, 318:493–503), (V. Gupta and Lehal 2009).

[6] Other Sources for Sentence Boundary Detection are: **Dutch** - (Wong, Chao, and Zeng 2014), **Hindi** - (Sa et al. 2018), **Marathi** - (Wanjari, Dhopavkar, and Zungre 2016); **Bengali -** (Bhowmik and Das Mandal 2020), ("BanglaKit Sentence Boundary Detector (SBD)" [2019] 2019).

[7] Other Sources for Named Entity Recognition are: **Arabic** - (Benajiba and Rosso 2023)**,** (Hossam [2022] 2023)**,** (Qu et al. 2023)**; Dutch -** (Schraagen et al. 2017)**,** (Van Toledo, Van Dijk, and Spruit 2020)**; Hindi -** (V. Singh et al. 2018), (Murthy et al. 2022), (R. Sharma, Morwal, and Agarwal 2022); **Tamil** - (H. S. Saini et al. 2019, 74:91–99), (Kathiravan and Saranya, n.d.), (Anbukkarasi et

al. 2022); **Telugu** - (Gorla et al. 2018), (Rajalakshmi et al. 2022), (Gorla et al. 2022); **Marathi** - (Dongare and Mhaiskar 2023), (Litake et al. 2022), (Hemanth et al. 2019, 26:371–77); **Bengali -** (Ekbal, Haque, and Bandyopadhyay 2008), (Rifat [2019] 2023), (Drovo et al. 2019); **Punjabi -** (Jaspreet Singh and Lehal 2015), (Chopra and Morwal 2012).

[8] Other Sources for Shallow parser are: **Arabic** - (Green and Manning 2010)**,** (Castro, Gelbukh, and González 2013)**, Dutch -** (Oostdijk and van Halteren 2013)**,** (Canisius 2003)**; Hindi -** (A. Sharma et al. 2016), ("Hindi Shallow Parser"); **Tamil** - (Ariaratnam, Weerasinghe, and Liyanage 2014); **Telugu** - ("Telugu Shallow Parser"), **Marathi** - (Lahoti, Mittal, and Singh 2023); **Bengali -** ("Bengali Shallow Parser"); **Punjabi -** ("Punjabi Shallow Parser").

[9] Other Sources for Spell checker are: **Arabic** - ("Free Online Arabic Spell Checker"), ("Arabic Spell Checker" [2018] 2023), **Dutch** - ("Free Online Dutch Spell Checker"), ("Multilingual Spell and Grammar Text Corrector"); **Hindi** - (S. Singh and Singh 2021), (R. Kaur and Sharma 2016), (Vinitha and Jawahar 2016); **Tamil** - (Pawan Kumar, Kannan, and Goel 2020), (Murugan, Bakthavatchalam, and Sankarasubbu 2020), (P. Gupta 2019); **Telugu** - (Etoori, Chinnakotla, and Mamidi 2018), **Marathi** - (P. Gupta 2019), (Dixit, Dethe, and Joshi 2007); **Bengali -** (Akash [2017] 2021), (Rahman et al. 2022); **Punjabi -** ("Punjabi Grammar Checker"), (G. Kaur, Kaur, and Singh 2019).

[10] Other Sources for Parser and grammars are: **Dutch** - ("NedGram Constraint Grammar Parser for Dutch – META-SHARE"), ("Generation Grammars - ACL Wiki"); **Hindi** - (Ramteke, Ramteke, and Dongare 2014), (Bhat et al. 2018), (Makwana and Vegda 2015); **Tamil** - (Rajendran 2006), (Sarveswaran and Dias 2021), (Saravanan, Parthasarathi, and Geetha 2003); **Marathi** - (Umale 2016), (Zhang, n.d.); **Bengali -** (Azharul Hasan et al. 2011), (Dhar et al. 2012).

[11] Other Sources for Font converters are: **Arabic** - ("Arabic and Arab-Culture Fonts | Font & Text Generator"); **Dutch** - ("Dutch Voice Typing - Unicode Font Converter"), ("Dutch Fonts"); **Bengali -** ("Bijoy - Unicode Converter | বিজয় - ইউনিকোড কনভার্টার"); **Punjabi -** ("Punjabi Gurmukhi Unicode Font Converter | Free Font Converter").

[12] Other Sources for Sentence Alignment are: **Arabic** - (Semmar and Fluhr 2007), (Fattah, Ren, and Kuroiwa 2006), (Ellouze, Neifar, and Belguith 2018); **Dutch** - (Trushkina, Macken, and Paulussen 2008), (Tiedemann 2007); **Hindi** - (Venkatapathy and Joshi 2007), (Aswani and Gaizauskas 2005), (Nigam and Jaiswal 2006); **Tamil** - (Harshawardhan and Augustine 2008); **Bengali -** (Ahmed, Hasan, and Selim 2018), ("Generating Parallel Translation Corpora in Indian Languages: Cultivating Bilingual Texts for Cross Lingual Fertilization" 2016).

[13] Other Sources for Text generation are: **Arabic** - (Hejazi et al. 2021); **Bengali**

[14] Other Sources for Text annotation tools are: **Arabic** - (Al-Twairesh et al. 2016), (Bouziane et al. 2021).

[15] Other Sources for Machine translators are: **Arabic** - (Zakraoui et al. 2021), ("Tarjama AMT - The Best Arabic Translation Tool"), (Akhter et al. 2017); **Dutch** - ("Dutch"); **Hindi** - (Sitender et al. 2023), (Goyal and Lehal 2011), (SHUKLA [2018] 2023); **Tamil** - (Pathak, Choudhary, and Shah 2018), (Sankaravelayuthan 2019); **Telugu** - (Rao, n.d.), (Lingam, Lakshmi, and Theja 2014), (Laskar et al., n.d.); **Marathi** - (Shirsath et al. 2021), (Jadhav 2020), (Aishwarya [2015] 2015); **Bengali -** (Menon [2020] 2023), (East West University, Dhaka-1212, Bangladesh et al. 2020), (Ohidujjaman et al. 2021); **Punjabi -** (M. Kaur 2018), (N. Bansal and Kumar 2020)

[16] Other Sources for Token detection are: **Arabic** - (Attia 2007); **Dutch** - (van Es et al. 2022), (Van Hee et al. 2018); **Hindi** - (K Panchapagesan et al. 2004), (M. Das et al. 2020); **Tamil** -

(Subramanian, Vadivel, and Shibani 2022), (Rajalakshmi et al. 2022); **Bengali -** (Alam, Khan, and Alam, n.d.), (Jahan et al. 2022), (Karim et al. 2022).

[17] Other Sources for OCR are: **Arabic** - (Alghamdi, Alkhazi, and Teahan 2016), (Nashwan et al. 2017), (R. Prasad et al. 2008); **Dutch** - ("Dutch; Flemish OCR for PDFs and Images"), ("Dutch Online OCR"), (Gheselle 2021); **Hindi** - (Mathew, Singh, and Jawahar 2016), (Jawahar, Pavan Kumar, and Ravi Kiran 2003), (Yadav, Sanchez-Cuadrado, and Morato 2013); **Tamil** - (Krishnamoorthy 2002), ("Search - Tag - Tamil OCR" n.d.), (Vasantharajan, Tharmalingam, and Thayasivam 2021)**; Telugu -** (Achanta and Hastie 2017), (Jawahar, Pavan Kumar, and Ravi Kiran 2003), (Chandra Prakash et al. 2018); **Marathi** - ("Marathi OCR"), (Ghodekar [2020] 2023), ("Marathi OCR")**; Bengali -** (Dipu, Shohan, and Salam 2021)**,** ("Bengali OCR Online (Image to Text) - Unicode Font Converter")**,** ("I2OCR - Free Online Bengali OCR")**; Punjabi -** ("Gurmukhi OCR :: Robust Document Analysis and Recognition System")**,** ("Indian Language OCR")**,** (Govindaraju and Setlur 2010, 43–71)

### 6.2 Analysis

The tables show differences in coverage for the Arabic, Dutch and major Indian languages. It can clearly be seen that Arabic and Dutch have more applications and modules as compared to the Indian languages. Considering ELRA alone, it has 71 Arabic LRs, 42 Dutch LRs, 9 Hindi LRs, 7 Tamil LRs, 4 Bengali LRs, 4 Punjabi LRs and 2 Telugu LRs. More research has been done on Arabic in the last decade due to widespread use of communication and information technology applications (Alansary, Nagi, and Adly 2013). Also, European countries have been funding national programs that accelerate measurable research, ELSNET (European Network in language and speech) is one of the examples which is sponsored by the Human Language Technologies

programme of the European commission. NEMLAR (Network for Euro-Mediterranean Language Resources) is another example of a European Commission project which carried out surveys on the availability of Arabic LRs in the region, this led to the identification of the needs and the availability of resources and tools in the BLARK "Basic Language Resources Kit" (Maegaard et al. 2006). Also, the ELRA distributes useful resources for the development of Arabic speech and language processing technologies and applications (Alansary, Nagi, and Adly 2013). But the same is not true for most languages in the world.

In 2007, The Central Institute of Indian Languages and Government of India became sensitized about this situation and thus they established LDC-IL (Linguistic Data Consortium for Indian Languages) ("LDC-IL"). Almost for a decade, LDC worked for 22 Indian scheduled languages (Languages with official status in India) and created large to medium-sized language resources (Choudhary 2018). It works on several types of linguistic resources that include text corpora, speech corpora, annotated speech corpora and image corpora for OCR but still Indian languages lags far behind other languages with respect to language resources. Though, more effort has been put on Hindi being the national language; it has more LRs and applications, but it is still a long way for the other languages to enjoy the luxury of having advanced NLP tools and applications.

### 6.3 What are the most pressing needs to be addressed and a set of recommendations?

It can be concluded from BLARK model that for any NLP module or application, Language Resource is a critical component. Critical mass of Language Resource (LR) can make advancement in research and technology development possible and quicker. To exemplify, giant IT companies

like Google or Microsoft are delivering extensive LT solutions as they access a huge amount of data in many different languages.

If building BLARK model for Arabic helped to enrich Arabic language, the same can be done for Punjabi too. Like European Commision established organizations and projects like ELRA and NEMLAR to reinforce and to create a cooperation roadmap for Human Language Technologies for Arabic, a similar roadmap can be followed for Punjabi as well. The purpose of such organizations should be to extend this collaboration to all persons and institutions who share the goal of promoting Punjabi language technology in a collaborative framework.

# CHAPTER 7: CONCLUSION

## 7.1 Summary of Contributions

Though the internet continues to reach more of the world and gain speakers of more languages, NLP technologies have only been deployed for a small fraction of those languages. Tools such as parsers and POS taggers are traditionally built by running supervised machine learning methods on a large, annotated corpus. There are usually not any such corpora for low-resource languages like Gurmukhi Punjabi, so instead of typically ignoring low-resource languages, the research should be directed towards what can be done without these annotated corpora like non-machine-learning, i.e. rule-based approaches - cf. EL-BlaRK (Arppe et al. 2016). low-resource languages are typically ignored. The main question we explored in this thesis is "What can be practically done to build NLP tools for Gurmukhi Punjabi with existing resources?"

The chapters of the thesis move through the stages, starting off with a brief introduction of Gurmukhi Punjabi, its history and orthographic features. In Chapter III, I introduced a few words of NLP and Computational Linguistics. I explained about the resource aggregators and types of language resources which could be used to build the NLP tools and applications for a language. The main purpose of Chapter I and Chapter II is to create a background required to understand the idea of the thesis.

Chapter IV transitions to discover how we can know that a language is a low resource language. It introduces a BLARK model which provides the basic parameters to investigate if the language is low resource or not. I gave examples of Dutch and Arabic BLARK models to strengthen my argument in this chapter.

In Chapter V, I defined BLARK for Punjabi. The chapter is divided into two sections - BLARK for written resources and BLARK for spoken resources. In both sections, I have investigated the currently available Punjabi LRs and basic NLP tools and applications. Since, BLARK outlines the Basic Resource Kit, so I have focused on basic NLP tools and applications for Punjabi. I have mentioned in detail what are the different resources and HLT modules required for Punjabi, what are currently available and what needs to be done. This chapter illustrates the gap analysis between what is available and what is required for Punjabi. The gap analysis helped in investigating the problem that lack of important Punjabi language resources and minimal amount of data in available resources proves that Punjabi is a low resources language.

In Chapter VI, I have advanced my research to other major Indian languages to show that it's not only Punjabi but other Indian languages as well that lack LRs and have limited NLP tools and applications. I defined BLARK by finding out available resources and NLP modules for Hindi, Marathi, Bengali, Tamil, Bengali, Telugu and Punjabi along with Arabic and Dutch. I compared Arabic and Dutch because those are two languages which use BLARK to deal with the problem of lack of language resources. Consequently, Dutch and Arabic languages are more powerful with respect to NLP applications. Arabic BLARK is one of the classic examples to showcase the effectiveness of the BLARK model. Having said that, this strengthens the thesis argument that defining and implementing BLARK for Gurmukhi Punjabi can help Punjabi researchers and scholars to work more effectively on creating Punjabi LRs.

### 7.2 Future Directions

Further research should focus on not only using existing NLP technologies such as the ones described in this paper for language description, but also on extending methods to be appropriate for

larger numbers of languages. Finally, there is still considerable room for improvement on the tasks studied in this thesis. All the techniques described here are rather naïve linguistically. Certainly, much more can be gained by making use of linguistic knowledge, for example, building into a cross lingual POS tagging system linguistic knowledge about why certain concepts are expressed differently syntactically in different languages. Such approaches are greatly lacking in NLP research today.

## WORKS CITED


"A Gold Standard Punjabi Raw Text Corpus." n.d. Accessed May 30, 2023. https://data.ldcil.org/a-gold-standard-punjabi-raw-text-corpus.

"A Gold Standard Tamil Raw Text Corpus." n.d. Accessed June 22, 2023. https://data.ldcil.org/a-gold-standard-tamil-raw-text-corpus.

"A Gold Standard Telugu Raw Text Corpus." n.d. Accessed June 23, 2023. https://data.ldcil.org/a-gold-standard-telugu-raw-text-corpus.

"Ajit - Punjab Di Awaaz." n.d. Accessed May 31, 2023. https://www.ajitjalandhar.com/.

"All_tamil_nouns." (2020) 2023. Python. Kaniyam Foundation. https://github.com/KaniyamFoundation/all_tamil_nouns.

"Applications of Synthetic Speech." n.d. Accessed June 7, 2023. http://research.spa.aalto.fi/publications/theses/lemmetty_mst/chap6.html.

"Arabic and Arab-Culture Fonts | Font & Text Generator." n.d. Accessed June 24, 2023. https://www.font-generator.com/arabic/#google_vignette.

"Arabic Spell Checker." (2018) 2023. Java. QCRI. https://github.com/qcri/ArabicSpellChecker.

"Awesome-Bangla." (2017) 2023. BanglaKit. https://github.com/banglakit/awesome-bangla.

"BanglaKit Bengali Stemmer." (2017) 2023. Python. BanglaKit. https://github.com/banglakit/bengali-stemmer.

"BanglaKit Sentence Boundary Detector (SBD)." (2019) 2019. Python. BanglaKit. https://github.com/banglakit/bengali-sbd.

"Bengali Dictionary Online Translation (Bangla) LEXILOGOS." n.d. Accessed June 24, 2023. https://www.lexilogos.com/english/bengali_dictionary.htm.

"Bengali OCR Online (Image to Text) - Unicode Font Converter." n.d. Accessed June 28, 2023. https://unicodefont.in/bengali-ocr-online.

"Bengali Shallow Parser." n.d. Accessed June 28, 2023. http://ltrc.iiit.ac.in/analyzer/bengali/.

"Bengali-English Translated Parallel Corpora for Tourism Domain." n.d. Accessed June 24, 2023. https://www.futurebeeai.com/dataset/parallel-corpora/bengali-english-translated-parallel-corpus-for-tourism-domain.

"Bijoy - Unicode Converter | বিজয় - ইউনিকোড কনভার্টার." n.d. Accessed June 28, 2023.

https://bsbk.portal.gov.bd/apps/bangla-converter/index.html.

"Collins Hindi Dictionary | Translations, Definitions and Pronunciations." n.d. Accessed June 21, 2023. https://www.collinsdictionary.com/dictionary/english-hindi.

"Digital Corpus of Old Marathi (DCOMA)." n.d. Accessed June 23, 2023. http://marathi.osu.edu/.

"Download Corpora Hindi." n.d. Accessed June 21, 2023. https://wortschatz.uni-leipzig.de/en/download/Hindi.

"Downloads | Siva Reddy." n.d. Accessed June 25, 2023. https://sivareddy.in/downloads/.

"Dutch Fonts." n.d. Fontspace. Accessed June 25, 2023. https://www.fontspace.com/category/dutch.

"Dutch Keyword Stemmer - SearchWP." 2015. April 28, 2015. https://searchwp.com/extensions/dutch-keyword-stemmer/.

"Dutch Online OCR." n.d. Accessed June 25, 2023. https://keyboardingonline.net/online-ocr-en-nl/.

"Dutch Stemmer." 2006. Drupal.Org. July 15, 2006. https://www.drupal.org/project/dutchstemmer.

"Dutch Voice Typing - Unicode Font Converter." n.d. Accessed June 25, 2023. https://unicodefont.in/dutch-voice-typing.

"Dutch; Flemish OCR for PDFs and Images." n.d. Accessed June 25, 2023. https://ocrbear.com/ocr-dutch;-flemish.

"Dutch." n.d. Machine Translate. Accessed June 25, 2023. https://machinetranslate.org/dutch.

"ELG - Morphology Analyzer and Generator." n.d. Accessed June 24, 2023. https://live.european-language-grid.eu/catalogue/tool-service/7007.

"ELRA - ELRA-U-S 0062 : Speech Corpus for Punjabi." n.d. Accessed June 8, 2023. http://universal.elra.info/product_info.php?cPath=37_39&products_id=1745.

"EMILLE Corpus Documentation." n.d. Accessed May 31, 2023. https://www.lancaster.ac.uk/fass/projects/corpus/emille/MANUAL.htm.

"English-Punjabi Code-Mixed Social Media Content – ELRA Catalogue." n.d. Accessed May 31, 2023. https://catalog.elra.info/en-us/repository/browse/ELRA-W0319/.

"English-Tamil Parallel Corpora." n.d. Accessed June 22, 2023. https://ufal.mff.cuni.cz/~ramasamy/parallel/html/.

"Free Online Arabic Spell Checker." n.d. Accessed June 24, 2023.

https://www.stars21.com/spelling/arabic/.

"Free Online Dutch Spell Checker." n.d. Accessed June 25, 2023. https://www.stars21.com/spelling/dutch/.

"Generating Parallel Translation Corpora in Indian Languages: Cultivating Bilingual Texts for Cross Lingual Fertilization." 2016. Translation Today 10 (1). https://doi.org/10.46623/tt/2016.10.1.ar5.

"Generation Grammars - ACL Wiki." n.d. Accessed June 25, 2023. https://aclweb.org/aclwiki/Generation_grammars.

"Government of India." 2011. 2011. https://censusindia.gov.in/census.website/data/census-tables.

"Gurmukhi OCR :: Robust Document Analysis and Recognition System." n.d. Accessed June 28, 2023. https://www.learnpunjabi.org/ocr/project.html.

"Hindi – Hindi -Telugu Dictionary – Neelkamal Publications Pvt. Ltd." n.d. Accessed June 23, 2023. https://www.neelkamalbooks.com/books/hindi-hindi-telugu-dictionary/.

"Hindi Dictionary - Online Hindi to Hindi Devanagari Words Dictionary." n.d. Accessed June 21, 2023. https://www.hindi2dictionary.com/.

"Hindi Dictionary Online Translation LEXILOGOS." n.d. Accessed June 21, 2023. https://www.lexilogos.com/english/hindi_dictionary.htm.

"Hindi Monolingual Chunked Text Corpus ILCI." n.d. Accessed June 21, 2023. https://nplt.in/demo/index.php?route=product/product&product_id=2192.

"Hindi Morphological Tagger." n.d. Accessed June 25, 2023. http://ccat.sas.upenn.edu/plc/tamilweb/hindi.html.

"Hindi Shallow Parser." n.d. Accessed June 25, 2023. http://ltrc.iiit.ac.in/analyzer/hindi/.

"Hindi Stop Words and Sentiment Lexicons." n.d. Accessed June 21, 2023. https://www.kaggle.com/datasets/ruchi798/hindi-stopwords.

"Hindi to Hindi Dictionary - Apps on Google Play." n.d. Accessed June 21, 2023. https://play.google.com/store/apps/details?id=com.sachi.hindi2hindi.dictionary&hl=en_CA.

"Hindi WordNet." n.d. Accessed June 21, 2023. https://www.cfilt.iitb.ac.in/wordnet/webhwn/wn.php.

"I2OCR - Free Online Bengali OCR." n.d. Accessed June 28, 2023. https://www.i2ocr.com/free-online-bengali-ocr.

"IIT Bombay English-Hindi Parallel Corpus." n.d. Accessed June 21, 2023. https://www.cfilt.iitb.ac.in/iitb_parallel/.

"Indian Language OCR." n.d. Accessed June 28, 2023. https://tdil-dc.in/eocr/index.html.
"LDC-IL." n.d. Accessed June 20, 2023. https://www.ldcil.org/.

"Marathi OCR." n.d. Marathi Typing. Accessed June 27, 2023. https://marathi.indiatyping.com/index.php/apps/ocr-marathi.

"Multilingual Spell and Grammar Text Corrector." n.d. WebSpellChecker (blog). Accessed June 25, 2023. https://webspellchecker.com/supported-languages/.

"NedGram Constraint Grammar Parser for Dutch – META-SHARE." n.d. Accessed June 25, 2023. http://metashare.tilde.com/repository/browse/nedgram-constraint-grammar-parser-for-dutch/bc027bb858da11e2b78a0050569b000027f73467bcc341fa815ec2db26a0c556/.

"Offline Bangla Dictionary (E2B & B2E)." 2022. SourceForge. December 27, 2022. https://sourceforge.net/projects/aparajeyo-bangla-dictionary/.

"Online Punjabi Morphological Analyzer." n.d. Accessed May 31, 2023. http://punjabi.aglsoft.com/?show=morph.

"Online Punjabi Part of Speech Tagger." n.d. Accessed May 31, 2023. http://punjabi.aglsoft.com/?show=tagger.

"Open-Tamil/Solthiruthi/Data at 5eb9fb1447fe021ca47e2cc4605f7111e6b1088f · Ezhil-Language-Foundation/Open-Tamil." n.d. GitHub. Accessed June 22, 2023. https://github.com/Ezhil-Language-Foundation/open-tamil.

"Openslr.Org." n.d. Accessed June 21, 2023. https://www.openslr.org/118/.

"Oxford University Press." 2023. In OED Online. Oxford University Press. https://www.oed.com/view/Entry/82651.

"Pakistan Bureau of Statistics." 2017. Pakistan Bureau of Statistics. 2017. https://www.pbs.gov.pk/content/final-results-census-2017-0.

"Pre-Labelled Datasets | Structured Datasets for AI/ML." n.d. FutureBeeAI. Accessed May 31, 2023. https://www.futurebeeai.com/dataset.

"Punjabi Becomes Third Language in Canada's House of Commons." n.d. Accessed June 28, 2023. https://www.ndtv.com/world-news/punjabi-becomes-third-language-in-canadas-house-of-commons-1239259.

"Punjabi Dialects and Languages." 2023. In Wikipedia. https://en.wikipedia.org/w/index.php?title=Punjabi_dialects_and_languages&oldid=115108

7500.

“Punjabi Grammar Checker.” n.d. Accessed June 28, 2023. https://pgc.learnpunjabi.org/.

“Punjabi Gurmukhi Unicode Font Converter | Free Font Converter.” n.d. Sikh Siyasat News (blog). Accessed June 28, 2023. https://sikhsiyasat.net/free-punjabi-gurmukhi-unicode-font-converter/.

“Punjabi Raw Speech Corpus.” n.d. Accessed June 8, 2023. https://data.ldcil.org/punjabi-raw-speech-corpus.

“Punjabi Shallow Parser.” n.d. Accessed June 28, 2023. http://ltrc.iiit.ac.in/analyzer/punjabi/.

“Punjabi Tribune.” n.d. Accessed May 31, 2023. https://www.punjabitribuneonline.com/.

“Punjabis Contributed to 60% Migration to Canada.” 2020. The Tribune India. 2020. https://www.tribuneindia.com/news/jalandhar/punjabis-contributed-to-60-migration-to-canada-44325.

“Sampadak.” n.d. Accessed May 26, 2023. https://www.cdac.in/index.aspx?id=mlc_pr_P_Sampadak.

“Search - Tag - Tamil OCR.” n.d. Accessed June 26, 2023. https://nplt.in/demo/index.php?route=product/search&tag=Tamil%20OCR.

“Stopwords Bengali (BN).” (2016) 2023. Stopwords ISO. https://github.com/stopwords-iso/stopwords-bn.

“Tamil - Language Corpus for NLP.” n.d. Accessed June 22, 2023. https://www.kaggle.com/datasets/praveengovi/tamil-language-corpus-for-nlp.

“Tamil Hindi Dictionary | Lexicool.” n.d. Accessed June 22, 2023. https://www.lexicool.com/online-dictionary.asp?FSP=A31B44.

“Tamil News Dataset.” n.d. Accessed June 22, 2023. https://www.kaggle.com/datasets/disisbig/tamil-news-dataset.

“Tamil Speech Recognition Corpus(Desktop)-Speechocean.” n.d. Accessed June 22, 2023. https://en.speechocean.com/datacenter/details/1612.html.

“Tamil to Tamil Dictionary | Tamil Translation Services | Tamil English Dictionary - Ilearntamil.” n.d. Learn Tamil Online-Best Tamil School Online (blog). Accessed June 22, 2023. https://ilearntamil.com/tamil-to-tamil-dictionary/.

“Tarjama AMT - The Best Arabic Translation Tool.” n.d. Accessed June 24, 2023. https://translate.tarjama.com/.

"TDIL-DC :Indo Wordnet." n.d. Accessed June 22, 2023. https://tdil-dc.in/index.php?option=com_vertical&parentid=90&lang=en.

"Telugu Monolingual PoS Tagged Text Corpus ILCI." n.d. Accessed June 23, 2023. https://nplt.in/demo/index.php?route=product/product&product_id=2176.

"Telugu Shallow Parser." n.d. Accessed June 27, 2023. http://ltrc.iiit.ac.in/analyzer/telugu/.

"Telugu-English Translated Parallel Corpora for Religious Domain." n.d. FutureBeeAI. Accessed June 23, 2023. https://www.futurebeeai.com/dataset/parallel-corpora/telugu-english-translated-parallel-corpus-for-religious-domain.

"The Benefits Of A Conversational Speech Dataset - Way With Words." n.d. Accessed June 7, 2023. https://waywithwords.net/landing/conversational-speech-dataset-2/.

"The EMILLE Corpus." n.d. Accessed May 31, 2023. https://www.lancaster.ac.uk/fass/projects/corpus/emille/.

"Unicode CLDR." n.d. Accessed April 27, 2023. https://cldr.unicode.org/.

A, Madhavaraj, Bharathi Pilar, and Ramakrishnan A G. 2022. "Knowledge-Driven Subword Grammar Modeling for Automatic Speech Recognition in Tamil and Kannada." https://doi.org/10.48550/ARXIV.2207.13333.

Abuczki, Ágnes, and Esfandiari Baiat Ghazaleh. 2013. "An Overview of Multimodal Corpora, Annotation Tools and Schemes."

Achanta, Rakesh, and Trevor Hastie. 2017. "Telugu OCR Framework Using Deep Learning." arXiv. http://arxiv.org/abs/1509.05962.

Ager, Simon. 2023. "Punjabi Language, Alphabets and Pronunciation." 2023. https://omniglot.com/writing/punjabi.htm.

Aggarwal, Salil, Abhigyan Ghosh, and Radhika Mamidi. n.d. "SUKHAN: Corpus of Hindi Shayaris Annotated with Sentiment Polarity Information."

Ahmed, Raihan, Mehedi Al Hasan, and Mohammad Reza Selim. 2018. "Aligning Sentences in English-Bengali Corpora." In 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), 1–5. Rajshahi: IEEE. https://doi.org/10.1109/IC4ME2.2018.8465608.

Aishwarya. (2015) 2015. "Aishward/Marathi-to-English-Machine-Translation-of-Simple-Sentences." Java. https://github.com/aishward/Marathi-to-English-Machine-Translation-of-Simple-Sentences.

Akash, Pritom Saha. (2017) 2021. "Context-Sensitive-Bangla-Spell-Checker." Python.

https://github.com/pritomsaha/Context-sensitive-Bangla-spell-checker.

Akhter, Maheen, Sahar Noor, Muhammad Ramzan, and Hikmat Ullah. 2017. “Evaluating Urdu to Arabic Machine Translation Tools.” International Journal of Advanced Computer Science and Applications 8 (10). https://doi.org/10.14569/IJACSA.2017.081012.

Akilan, R. n.d. “Morphological Analyzer for Classical Tamil Texts: A Rule-Based Approach for Case Marker.”

Al-Twairesh, Nora, Abeer Al-Dayel, Hend Al-Khalifa, Maha Al-Yahya, Sinaa Alageel, Nora Abanmy, and Nouf Al-Shenaifi. n.d. “MADAD: A Readability Annotation Tool for Arabic Text.”

Alam, Tanvirul, Akib Khan, and Firoj Alam. n.d. “Bangla Text Classification Using Transformers.”

Alansary, Sameh, Magdy Nagi, and Noha Adly. 2013. “A Suite of Tools for Arabic Natural Language Processing: A UNL Approach.” In 2013 1st International Conference on Communications, Signal Processing, and Their Applications (ICCSPA), 1–6. Sharjah: IEEE. https://doi.org/10.1109/ICCSPA.2013.6487236.

Alghamdi, Mansoor A, Ibrahim S Alkhazi, and William J. Teahan. 2016. “Arabic OCR Evaluation Tool.” In 2016 7th International Conference on Computer Science and Information Technology (CSIT), 1–6. Amman, Jordan: IEEE. https://doi.org/10.1109/CSIT.2016.7549460.

Amrouche, Aissa, Ahcène Abed, Kamel Ferrat, Khadidja Nesrine Boubakeur, Youssouf Bentrcia, and Leila Falek. 2021. “Balanced Arabic Corpus Design for Speech Synthesis.” International Journal of Speech Technology 24 (3): 747–59. https://doi.org/10.1007/s10772-021-09846-8.

Anbukkarasi, S., S. Varadhaganapathy, S. Jeevapriya, A. Kaaviyaa, T. Lawvanyapriya, and S. Monisha. 2022. “Named Entity Recognition for Tamil Text Using Deep Learning.” In 2022 International Conference on Computer Communication and Informatics (ICCCI), 1–5. Coimbatore, India: IEEE. https://doi.org/10.1109/ICCCI54379.2022.9740745.

Ariaratnam, I., A. R. Weerasinghe, and C. Liyanage. 2014. “A Shallow Parser for Tamil.” In 2014 14th International Conference on Advances in ICT for Emerging Regions (ICTer), 197–203. Colombo, Sri Lanka: IEEE. https://doi.org/10.1109/ICTER.2014.7083901.

Arppe, Antti, Kathrin Beck, Antonio Branco, Vanessa Camilleri, Tommaso Caselli, Dan Cristea, Erhard Hinrichs, et al. n.d. “Description of the BLARK, the Situation of Individual Languages D5C-4.”

Aswani, Niraj, and Robert Gaizauskas. 2005. “A Hybrid Approach to Align Sentences and Words in English-Hindi Parallel Corpora.” In Proceedings of the ACL Workshop on Building and Using Parallel Texts - ParaText ’05, 57. Ann Arbor, Michigan: Association for

Computational Linguistics. https://doi.org/10.3115/1654449.1654458.

Attia, Mohammed A. 2007. “Arabic Tokenization System.” In Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages Common Issues and Resources - Semitic ’07, 65. Prague, Czech Republic: Association for Computational Linguistics. https://doi.org/10.3115/1654576.1654588.

Azharul Hasan, K.M., Al Mahmud, Amit Mondal, and Amit Saha. 2011. “Recognizing Bangla Grammar Using Predictive Parser.” International Journal of Computer Science and Information Technology 3 (6): 61–73. https://doi.org/10.5121/ijcsit.2011.3605.

Badugu, Srinivasu. 2014. “Morphology Based POS Tagging on Telugu” 11 (1).

Baker, Mona. 1993. “Corpus Linguistics and Translation Studies — Implications and Applications.” In Text and Technology, edited by Mona Baker, Gill Francis, and Elena Tognini-Bonelli, 233. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/z.64.15bak.

Baker, Paul, Andrew Hardie, and Tony McEnery. 2006. A Glossary of Corpus Linguistics. Edinburgh: Edinburgh University Press.

Bansal, Nitin, and Ajit Kumar. n.d. “Punjabi to Urdu Machine Translation System.”

Bansal, Sameer, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018. “Low-Resource Speech-to-Text Translation.” In Interspeech 2018, 1298–1302. ISCA. https://doi.org/10.21437/Interspeech.2018-1326.

Beesley, Kenneth R, and Lauri Karttunen. 2003. “Finite-State Morphology: Xerox Tools and Techniques.”

Benajiba, Yassine, and Paolo Rosso. 2023. “Arabic Named Entity Recognition Using Conditional Random Fields,” April.

Bhardwaj, Mangat. 2016. Panjabi: A Comprehensive Grammar. 0 ed. Milton Park, Abingdon, Oxon ; New York, NY : Routledge, [2016] |: Routledge. https://doi.org/10.4324/9781315760803.

Bhat, Irshad, Riyaz A. Bhat, Manish Shrivastava, and Dipti Sharma. 2018. “Universal Dependency Parsing for Hindi-English Code-Switching.” In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 987–98. New Orleans, Louisiana: Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1090.

Bhattacharyya, Pushpak, Pande, Prabhakar, and Lupu, Laxmi. 2008. “Hindi WordNet.” Linguistic Data Consortium. https://doi.org/10.35111/S81S-5N27.

Bhowmik, Tanmay, and Shyamal Kumar Das Mandal. 2020. “Prosodic Word Boundary Detection

from Bengali Continuous Speech." Language Resources and Evaluation 54 (3): 747–65. https://doi.org/10.1007/s10579-019-09478-0.

Binnenpoorte, D, F De Vriend, J Sturm, W Daelemans, H Strik, and C Cucchiarini. n.d. "A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch."

Biswas, Mithun, Rafiqul Islam, Gautam Kumar Shom, Md Shopon, Nabeel Mohammed, Sifat Momen, and Anowarul Abedin. 2017. "BanglaLekha-Isolated: A Multi-Purpose Comprehensive Dataset of Handwritten Bangla Isolated Characters." Data in Brief 12: 103–7. https://doi.org/10.1016/j.dib.2017.03.035.

Biswas, Monisha, and Mohammed Moshiul Hoque. 2019. "Development of a Bangla Sense Annotated Corpus for Word Sense Disambiguation." In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), 1–6. Sylhet, Bangladesh: IEEE. https://doi.org/10.1109/ICBSLP47725.2019.201516.

Biswas, Sailendra. 2004. "Samsad Bengali-English Dictionary." Dictionary. Kalikata : Sahitya Samsad. 2004. https://dsal.uchicago.edu/dictionaries/biswas-bengali/.

Bouziane, Abdelghani, Djelloul Bouchiha, Redha Rebhi, Giulio Lorenzini, Noureddine Doumi, Younes Menni, and Hijaz Ahmad. 2021. "ARALD: Arabic Annotation Using Linked Data." Ingénierie Des Systèmes d Information 26 (2): 143–49. https://doi.org/10.18280/isi.260201.

Bowker, Lynne. 2002. "Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study." Meta 43 (4): 631–51. https://doi.org/10.7202/002134ar.

Brown, Charles Philip. 1903. "A Telugu-English Dictionary. New Ed., Thoroughly Rev. and Brought up to Date...2nd Ed." Dictionary. Madras, Promoting Christian Knowledge. 1903. https://dsal.uchicago.edu/dictionaries/brown/.

Burla, Rajesh, and B Tech. n.d. "Lexicon Based Sentiment Analyzer for the Telugu Language."

Burnard, Lou. 2000. "British National Corpus." Text. 2000. http://www.natcorp.ox.ac.uk/.

Cai, Tianrun, Andreas A. Giannopoulos, Sheng Yu, Tatiana Kelil, Beth Ripley, Kanako K. Kumamaru, Frank J. Rybicki, and Dimitrios Mitsouras. 2016. "Natural Language Processing Technologies in Radiology Research and Clinical Applications." RadioGraphics 36 (1): 176–91. https://doi.org/10.1148/rg.2016150080.

Campbell, J.P. 1997. "Speaker Recognition: A Tutorial." Proceedings of the IEEE 85 (9): 1437–62. https://doi.org/10.1109/5.628714.

Canisius, Sander. n.d. "A Memory-Based Shallow Parser for Spoken Dutch."

Castro, Félix, Alexander Gelbukh, and Miguel González, eds. 2013. Advances in Soft Computing and Its Applications. Vol. 8266. Lecture Notes in Computer Science. Berlin, Heidelberg:

Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-45111-9.

Chakravarthi, Bharathi Raja, Jishnu Parameswaran P. K, Premjith B, K. P. Soman, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kingston Pal Thamburaj, and John P. McCrae. 2021. "DravidianMultiModality: A Dataset for Multi-Modal Sentiment Analysis in Tamil and Malayalam." arXiv. http://arxiv.org/abs/2106.04853.

Chakravarthi, Bharathi Raja, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. n.d. "Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text."

Chandra Prakash, Konkimalla, Y. M. Srikar, Gayam Trishal, Souraj Mandal, and Sumohana S. Channappayya. 2018. "Optical Character Recognition (OCR) for Telugu: Database, Algorithm and Application." In 2018 25th IEEE International Conference on Image Processing (ICIP), 3963–67. Athens: IEEE. https://doi.org/10.1109/ICIP.2018.8451438.

Chaudhury, Tasnim Haider, Abdul Matin, M S Hossain, and Asie Uzzaman. 2017. "Annotated Bangla News Corpus and Lexicon Development with POS Tagging and Stemming."

Chopra, Deepti, and Sudha Morwal. 2012. "Named Entity Recognition in Punjabi Using Hidden Markov Model." Engineering Technology 3 (12).

Choudhary, Narayan. 2018. Cost Analysis of Linguistic Resources. Central Institute of Indian Languages, Mysore.

Choudhary, Narayan. 2021. "LDC-IL: The Indian Repository of Resources for Language Technology." Language Resources and Evaluation 55 (3): 855–67. https://doi.org/10.1007/s10579-020-09523-3.

Cieri, Christopher, and Mark Liberman. n.d. "Language Resource Creation and Distribution at the Linguistic Data Consortium: A Progress Report."

Cochrane, Shaun. 2004. "Introduction to Telephony Speech Technologies." ITWeb. July 30, 2004. https://www.itweb.co.za/content/DZQ58vV6d56vzXy2.

Converter, Online OCR. n.d. "Free OCR: Convert Image to Text Convertor | Image to Word Convertor." Online OCR Converter. Accessed May 31, 2023. https://onlineocrconverter.com/.

Crystal, David. 2003. A Dictionary of Linguistics & Phonetics / David Crystal. A Dictionary of Linguistics & Phonetics. 5th ed. The Language Library. Malden, MA: Blackwell Pub.

Darwish, Kareem, Hamdy Mubarak, and Ahmed Abdelali. 2017. "Arabic Diacritization: Stats, Rules, and Hacks." In Proceedings of the Third Arabic Natural Language Processing Workshop, 9–17. Valencia, Spain: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1302.

Das, Mithun, Punyajoy Saha, Binny Mathew, and Animesh Mukherjee. n.d. “HateCheckHIn: Evaluating Hindi Hate Speech Detection Models.”

Das, Priyanka, and Arjun Das. 2013. “Bengali Noun Morphological Analyzer.” In 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 1538–43. Mysore: IEEE. https://doi.org/10.1109/ICACCI.2013.6637408.

Das, Souvick, Rajat Pandit, and Sudip Kumar Naskar. n.d. “A Rule Based Lightweight Bengali Stemmer.”

Dash, Niladri Sekhar, Pushpak Bhattacharyya, and Jyoti D. Pawar, eds. 2017. The WordNet in Indian Languages. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-10-1909-8.

Department of CSE, CMR college of Engineering & Technology Hyderabad, India, and Dr. Narla Swapna*. 2019. “Root Based Stemmer for Telugu Script.” International Journal of Engineering and Advanced Technology 8 (6): 2565–68. https://doi.org/10.35940/ijeat.F8734.088619.

Dev, Amita, Arun Sharma, and S. S. Agarwal. 2021. Artificial Intelligence and Speech Technology: Proceedings of the 2nd International Conference on Artificial Intelligence and Speech Technology, (AIST2020), 19-20 November, 2020, Delhi, India. CRC Press.

Dhanjal, Amandeep Singh, and Williamjeet Singh. 2022. “An Optimized Machine Translation Technique for Multi-Lingual Speech to Sign Language Notation.” Multimedia Tools and Applications 81 (17): 24099–117. https://doi.org/10.1007/s11042-022-12763-w.

Dhar, Arnab, Sanjay Chatterji, Sudeshna Sarkar, and Anupam Basu. n.d. “A Hybrid Dependency Parser for Bangla.”

Dipu, Nadim Mahmud, Sifatul Alam Shohan, and K. M. A. Salam. 2021. “Bangla Optical Character Recognition (OCR) Using Deep Learning Based Image Classification Algorithms.” In 2021 24th International Conference on Computer and Information Technology (ICCIT), 1–5. Dhaka, Bangladesh: IEEE. https://doi.org/10.1109/ICCIT54785.2021.9689864.

Dixit, Veena, Satish Dethe, and Rushikesh K Joshi. n.d. “Design and Implementation of a Morphology-Based Spellchecker for Marathi, an Indian Language.”

Dongare, Pratibha, and Rahul Mhaiskar. 2023. “Named Entity Recognition For Marathi.”

Drovo, Mah Dian, Moithri Chowdhury, Saiful Islam Uday, and Amit Kumar Das. 2019. “Named Entity Recognition in Bengali Text Using Merged Hidden Markov Model and Rule Base Approach.” In 2019 7th International Conference on Smart Computing & Communications (ICSCC), 1–5. Sarawak, Malaysia, Malaysia: IEEE. https://doi.org/10.1109/ICSCC.2019.8843661.

East West University, Dhaka-1212, Bangladesh, Shaykh Siddique, Tahmid Ahmed, Md. Rifayet Azam Talukder, and Md. Mohsin Uddin. 2020. “English to Bangla Machine Translation Using Recurrent Neural Network.” International Journal of Future Computer and Communication, June, 46–51. https://doi.org/10.18178/ijfcc.2020.9.2.564.

Eberhard, David M., Simons Gary F., and Fennig (eds.) Charles D. 2023. “India | Ethnologue Free.” Ethnologue (Free All). 2023. http://www.ethnologue.com.

Ekbal, Asif, and Sriparna Saha. 2013. “Simulated Annealing Based Classifier Ensemble Techniques: Application to Part of Speech Tagging.” Information Fusion 14 (3): 288–300. https://doi.org/10.1016/j.inffus.2012.06.002.

Ekbal, Asif, Rejwanul Haque, and Sivaji Bandyopadhyay. n.d. “Named Entity Recognition in Bengali: A Conditional Random Field Approach.”

Elenius, Kjell, Eva Forsbom, and Beáta Megyesi. 2008. “Language Resources and Tools for Swedish: A Survey.”

Ellouze, Mourad, Wafa Neifar, and Lamia Hadrich Belguith. n.d. “Word Alignment Applied on English-Arabic Parallel Corpus.”

Es, Bram van, Leon C. Reteig, Sander C. Tan, Marijn Schraagen, Myrthe M. Hemker, Sebastiaan R. S. Arends, Miguel A. R. Rios, and Saskia Haitjema. 2022. “Negation Detection in Dutch Clinical Texts: An Evaluation of Rule-Based and Machine Learning Methods.” arXiv. http://arxiv.org/abs/2209.00470.

Etoori, Pravallika, Manoj Chinnakotla, and Radhika Mamidi. 2018. “Automatic Spelling Correction for Resource-Scarce Languages Using Deep Learning.” In Proceedings of ACL 2018, Student Research Workshop, 146–52. Melbourne, Australia: Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-3021.

Everton, Gomede. 2023. “The Shallow Parsing in Natural Language Processing.” Medium (blog). April 11, 2023. https://medium.com/@evertongomede/the-shallow-parsing-in-natural-language-processing-89cf3049f42c.

Fabricius, J. P. 1972. “J. P. Fabricius’s Tamil and English Dictionary. 4th Ed., Rev.and Enl.” Dictionary. Evangelical Lutheran Mission Pub. House. 1972. https://dsal.uchicago.edu/dictionaries/fabricius/.

Faridee, Abu Zaher. n.d. “Development of a Morphological Analyser for Bengali.”

Fattah, Mohamed Abdel, Fuji Ren, and Shingo Kuroiwa. n.d. “Text-Based English-Arabic Sentence Alignment.”

Federici, Stefano, Simonetta Montemagni, and Vito Pirrelli. n.d. “Shallow Parsing And Text Chunking: A View On Underspecification In Syntax.”

Federici, Stefano, Simonetta Montemagni, and Vito Pirrelli. "Shallow parsing and text chunking: a view on underspecification in syntax." Cognitive science research paper-university of Sussex CSRP (1996): 35-44.

Francis, W.N., and H. Kucera. 1979. “Brown Corpus Manual.” 1979. http://korpus.uib.no/icame/manuals/BROWN/INDEX.HTM.

Frankenberg-Garcia, Ana. 2005. “Pedagogical Uses of Monolingual and Parallel Concordances.” ELT Journal 59 (3): 189–98. https://doi.org/10.1093/elt/cci038.

Ganesh, T Sree. 2006. “Telugu Parts Of Speech Tagging In Wsd” 6.

Garg, Navneet, Vishal Goyal, and Suman Preet. n.d. “Rule Based Hindi Part of Speech Tagger.”

Garside, Roger, Leech Geoffrey, and McEnery Tony. 2013. Corpus Annotation.

Gawade, Pratiksha, Deepika Madhavi, Jayshree Gaikwad, Sharvari Jadhav, and Rahul Ambekar. 2013. “Morphological Analyzer for Marathi Using NLP.” International Journal of Engineering Research and Applications 3 (2).

Gelbukh, Alexander, ed. 2023. Computational Linguistics and Intelligent Text Processing: 20th International Conference, CICLing 2019, La Rochelle, France, April 7–13, 2019, Revised Selected Papers, Part I. Vol. 13451. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-24337-0.

Gheselle, Simon De. 2021. “OCR Correction with ByT5.” Medium. December 15, 2021. https://blog.ml6.eu/ocr-correction-with-byt5-5994d1217c07.

Ghodekar, Sayali. (2020) 2023. “Marathi-OCR-Dataset.” https://github.com/sayalighodekar/Marathi-OCR-Dataset.

Ghosh, Dr Siddhartha. 2014. “Morphological Analysis Of Telugu Language Using Apertium” 15.

Ghosh, Mohul. 2018. “India Produces 25% Of World’s Engineers, But Lacks In Researchers!” Trak.in. Trak.in - Indian Business of Tech, Mobile & Startups (blog). January 22, 2018. https://trak.in/tags/business/2018/01/22/india-produces-25-percent-world-engineers/.

Gill, H.S. 1996. “The Gurmukhi Script.” In The World’s Writing Systems. Daniels, P.T. and Bright, W. (eds.).

Godfrey, John J. 1994. “Multilingual Speech Databases at LDC.” In Proceedings of the Workshop on Human Language Technology - HLT ’94, 23. Plainsboro, NJ: Association for Computational Linguistics. https://doi.org/10.3115/1075812.1075819.

Gorla, SaiKiranmai, Sai Sharan Tangeda, Lalita Bhanu Murthy Neti, and Aruna Malapati. 2022. “Telugu Named Entity Recognition Using Bert.” International Journal of Data Science and

Analytics 14 (2): 127–40. https://doi.org/10.1007/s41060-021-00305-w.

Gorla, SaiKiranmai, Sriharshitha Velivelli, N L Bhanu Murthy, and Aruna Malapati. n.d. “Named Entity Recognition for Telugu News Articles Using Naïve Bayes Classifier.”

Govindaraju, Venu, and Srirangaraj (Ranga) Setlur, eds. 2010. Guide to OCR for Indic Scripts. Advances in Pattern Recognition. London: Springer London. https://doi.org/10.1007/978-1-84800-330-9.

Goyal, Vishal, and Gurpreet Singh Lehal. 2008. “Hindi Morphological Analyzer and Generator.” In 2008 First International Conference on Emerging Trends in Engineering and Technology, 1156–59. Nagpur, Maharashtra, India: IEEE. https://doi.org/10.1109/ICETET.2008.11.

Green, Spence, and Christopher D Manning. n.d. “Better Arabic Parsing: Baselines, Evaluations, and Analysis.”

Greenbaum, Sidney, and Jan Svartvik. 1990. “LLC Corpus.” 1990. https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/llc.html.

Gruber, Tom. 1993. What Is an Ontology.

Gupta, Prabhakar. 2019. “A Context Sensitive Real-Time Spell Checker with Language Adaptability.” arXiv. http://arxiv.org/abs/1910.11242.

Gupta, Vishal, and Gurpreet Singh Lehal. n.d. “Punjabi Language Stemmer for Nouns and Proper Names.”

Harshawardhan, R, and Mridula Sara Augustine. n.d. “A Simplified Approach To Word Alignment Algorithm For English-Tamil Translation” 2 (1).

Hejazi, Hani D., Ahmed A. Khamees, Muhammad Alshurideh, and Said A. Salloum. 2021. “Arabic Text Generation: Deep Learning for Poetry Synthesis.” In Advanced Machine Learning Technologies and Applications, edited by Aboul-Ella Hassanien, Kuo-Chi Chang, and Tang Mincong, 1339:104–16. Advances in Intelligent Systems and Computing. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-69717-4_11.

Hemanth, Jude, Xavier Fernando, Pavel Lafata, and Zubair Baig, eds. 2019. International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018. Vol. 26. Lecture Notes on Data Engineering and Communications Technologies. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-03146-6.

Hinrichs, Erhard, and Steven Krauwer. n.d. “The CLARIN Research Infrastructure: Resources and Tools for EHumanities Scholars.”

Hoque, Md. Nesarul, and Md. Hanif Seddiqui. 2015. “Bangla Parts-of-Speech Tagging Using

Bangla Stemmer and Rule Based Analyzer." In 2015 18th International Conference on Computer and Information Technology (ICCIT), 440–44. Dhaka, Bangladesh: IEEE. https://doi.org/10.1109/ICCITechn.2015.7488111.

Hossam, Mahmoud. (2022) 2023. "Arabic Named Entity Recognition." Jupyter Notebook. https://github.com/Th3Moody/Arabic-Named-Entity-Recognition.

Howcroft, David M., and Dimitra Gkatzia. "Most NLG is Low-Resource: here's what we can do about it." Association for Computational Linguistics (ACL), 2022.

Hussain, Qandeel, Michael Proctor, Mark Harvey, and Katherine Demuth. 2020. "Punjabi (Lyallpuri Variety)." Journal of the International Phonetic Association 50 (2): 282–97. https://doi.org/10.1017/S0025100319000021.

Hussain, Qandeel. 2020. "Punjabi (India and Pakistan) – Language Snapshot."

ivdnt.org. n.d. "Frog." INT Taalmaterialen (blog). Accessed June 25, 2023. https://taalmaterialen.ivdnt.org/download/tstc-frog/.

Jadhav, Swapnil Ashok. n.d. "Marathi To English Neural Machine Translation With Near Perfect Corpus And Transformers."

Jahan, Saroar, Mainul Haque, Nabil Arhab, and Mourad Oussalah. n.d. "BanglaHateBERT: BERT for Abusive Language Detection in Bengali."

Jain, Arti, Anuja Arora, Divakar Yadav, Jorge Morato, and Amanpreet Kaur. 2021. "Text Summarization Technique for Punjabi Language Using Neural Networks." The International Arab Journal of Information Technology. https://doi.org/10.34028/iajit/18/6/8.

Jawahar, C.V., M.N.S.S.K. Pavan Kumar, and S.S. Ravi Kiran. 2003. "A Bilingual OCR for Hindi-Telugu Documents and Its Applications." In Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings., 1:408–12. Edinburgh, UK: IEEE Comput. Soc. https://doi.org/10.1109/ICDAR.2003.1227699.

Johansson, Stig, Geoffrey N. Leech, And Helen Goodluck. 1978. "London/Oslo/Bergen Corpus." http://korpus.uib.no/icame/manuals/LOB/INDEX.HTM.

Jonker, Anne. n.d. "Bag & Tag'em - A New Dutch Stemmer."

Joshi, Nisheeth, Hemant Darbari, and Iti Mathur. 2013. "HMM Based POS Tagger for Hindi." In Computer Science & Information Technology ( CS & IT ), 341–49. Academy & Industry Research Collaboration Center (AIRCC). https://doi.org/10.5121/csit.2013.3639.

K Panchapagesan, Partha Pratim Talukdar, N Sridhar, Kalika Bali, and A G Ramakrishnan. 2004. "Hindi Text Normalization." https://doi.org/10.13140/RG.2.1.3839.8565.

Kapoor, Arnav, Mudit Dhawan, Anmol Goel, T. H. Arjun, Akshala Bhatnagar, Vibhu Agrawal, Amul Agrawal, Arnab Bhattacharya, Ponnurangam Kumaraguru, and Ashutosh Modi. 2022. "HLDC: Hindi Legal Documents Corpus." arXiv. http://arxiv.org/abs/2204.00806.

Karim, Md Rezaul, Sumon Kanti Dey, Tanhim Islam, Md Shajalal, and Bharathi Raja Chakravarthi. 2022. "Multimodal Hate Speech Detection from Bengali Memes and Texts." arXiv. http://arxiv.org/abs/2204.10196.

Kathiravan, P, and R Saranya. n.d. "Named Entity Recognition (NER) For Social Media Tamil Posts Using Deep Learning With Singular Value Decomposition."

Kaur, Amandeep, G Josan, and Jagroop Kaur. 2009. "Named Entity Recognition for Punjabi: A Conditional Random Field Approach." In Proceedings of 7th International Conference on Natural Language ProcessingICON-09. Macmillan Publishers, India.

Kaur, Dhanoj Sandeep, and Dr. Amitoj Singh. 2015. "Schwa Deletion: Investigating Improved Approach for Text-to-IPA System for Shiri Guru Granth Sahib." IJARCCE 4 (4): 623–25. https://doi.org/10.17148/IJARCCE.2015.44145.

Kaur, Gurjit, Kamaldeep Kaur, and Parminder Singh. 2019. "Spell Checker for Punjabi Language Using Deep Neural Network." In 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 147–51. Coimbatore, India: IEEE. https://doi.org/10.1109/ICACCS.2019.8728369.

Kaur, Manpreet. n.d. "A More Accurate Punjabi To English Machine Transliteration System For Proper Nouns."

Kaur, Rajneet, and Dr Dharam Veer Sharma. 2016. "Development of Font Independent Spell Checker for Hindi" 2 (6).

Kempton, Timothy, and Roger K. Moore. 2014. "Discovering the Phoneme Inventory of an Unwritten Language: A Machine-Assisted Approach." Speech Communication 56 (January): 152–66. https://doi.org/10.1016/j.specom.2013.02.006.

Kilgarriff, A. 2003. "Thesauruses for Natural Language Processing." In International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003, 5–13. Beijing, China: IEEE. https://doi.org/10.1109/NLPKE.2003.1275859.

Krauwer, Steven. "ELSNET and ELRA: A common past and a common future." ELRA Newsletter 3, no. 2 (1998): 4-5.

Krauwer, Steven. "The basic language resource kit (BLARK) as the first milestone for the language resources roadmap." In Proceedings of SPECOM, vol. 2003, no. 8, p. 15. 2003.

Krishnamoorthy, Dr V. 2002. "OCR Software for Printed Tamil Text."

Kulkarni, Atharva, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. n.d. "L3CubeMahaSent: A Marathi Tweet-Based Sentiment Analysis Dataset."

Kumar, Joginder, Manik Rakhra, and Preeti Dubey. 2022. "Bilingual Parallel Corpora: A Major Resource for Developing Computational Tools for Automatic Processing of Hindi-Dogri Language Pair." In 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 1–6. Noida, India: IEEE. https://doi.org/10.1109/ICRITO56286.2022.9964875.

Kumar, L. Ashok, L. S. Jayashree, and R. Manimegalai, eds. 2020. Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications: AISGSC 2019. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-24051-6.

Kumar, Pardeep, and Vishal Goyal. 2010. "Development of Hindi-Punjabi Parallel Corpus Using Existing Hindi-Punjabi Machine Translation System." In Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia, 114–18. Allahabad India: ACM. https://doi.org/10.1145/1963564.1963583.

Kumar, Pawan, Abishek Kannan, and Nikita Goel. 2020. "Design and Implementation of NLP-Based Spell Checker for the Tamil Language." In Proceedings of 1st International Electronic Conference on Applied Sciences, 7636. Sciforum.net: MDPI. https://doi.org/10.3390/ASEC2020-07636.

Lahoti, Pawan, Namita Mittal, and Girdhari Singh. 2023. "A Survey on NLP Resources, Tools, and Techniques for Marathi Language Processing." ACM Transactions on Asian and Low-Resource Language Information Processing 22 (2): 1–34. https://doi.org/10.1145/3548457.

Laskar, Sahinur Rahman, Bishwaraj Paul, Prottay Kumar Adhikary, Partha Pakray, and Sivaji Bandyopadhyay. n.d. "Neural Machine Translation for Tamil–Telugu Pair."

Laurent, Antoine, Thiago Fraga-Silva, Lori Lamel, and Jean-Luc Gauvain. 2016. "Investigating Techniques for Low Resource Conversational Speech Recognition." In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5975–79. Shanghai: IEEE. https://doi.org/10.1109/ICASSP.2016.7472824.

Leech, Geoffrey, Roger Garside, and Michael Bryant. 1994. "CLAWS4: The Tagging of the British National Corpus." In Proceedings of the 15th Conference on Computational Linguistics -, 1:622. Kyoto, Japan: Association for Computational Linguistics. https://doi.org/10.3115/991886.991996.

Lehal, Gurpreet Singh. 2009. "A Survey of the State of the Art in Punjabi Language Processing."

Liang, Jian, Daniel DeMenthon, and David Doermann. 2008. "Geometric Rectification of Camera-Captured Document Images." IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (4): 591–605. https://doi.org/10.1109/TPAMI.2007.70724.

Liberman, Mark, and Christopher Cieri. n.d. “The Creation, Distribution And Use Of Linguistic Data: The Case Of The Linguistic Data Consortium.”

Lingam, Keerthi, E. Rama Lakshmi, and L Ravi Theja. 2014. “Rule-Based Machine Translation from English to Telugu with Emphasis on Prepositions.” In 2014 First International Conference on Networks & Soft Computing (ICNSC2014), 183–87. Guntur, Andhra Pradesh, India: IEEE. https://doi.org/10.1109/CNSC.2014.6906669.

Litake, Onkar, Maithili Ravindra Sabane, Parth Sachin Patil, Aparna Abhijeet Ranade, and Raviraj Joshi. n.d. “L3Cube-MahaNER: A Marathi Named Entity Recognition Dataset and BERT Models.”

M.A, C. Sesma. 2009. Bengali Edition Word To Word Bilingual Dictionary. Bengali edition. Bilingual Dictionaries, Inc.

Madras, University of. 1924. “Tamil Lexicon.” Dictionary. Madras : University of Madras. 1936 1924. https://dsal.uchicago.edu/dictionaries/tamil-lex/.

Maegaard, Bente, Steven Krauwer, Khalid Choukri, and Lise Damsgaard Jørgensen. 2006. “The BLARK Concept and BLARK for Arabic.” In LREC, 773–78. Citeseer.

Makwana, Monika T, and Deepak C Vegda. n.d. “Survey:Natural Language Parsing For Indian Languages.”

Mantoro, Teddy, Minho Lee, Media Anugerah Ayu, Kok Wai Wong, and Achmad Nizar Hidayanto, eds. 2021. Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part II. Vol. 13109. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-92270-2.

Masica, C.P. 1991. The Indo-Aryan Languages. Cambridge Language Surveys. Cambridge University Press. https://books.google.ca/books?id=bI5fQgAACAAJ.

Mathew, Minesh, Ajeet Kumar Singh, and C. V. Jawahar. 2016. “Multilingual OCR for Indic Scripts.” In 2016 12th IAPR Workshop on Document Analysis Systems (DAS), 186–91. Santorini, Greece: IEEE. https://doi.org/10.1109/DAS.2016.68.

Meetei, Loitongbam Sanayai, Salam Michael Singh, Alok Singh, Ringki Das, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2023. “Hindi to English Multimodal Machine Translation on News Dataset in Low Resource Setting.” Procedia Computer Science 218: 2102–9. https://doi.org/10.1016/j.procs.2023.01.186.

Menon, Mehadi Hasan. (2020) 2023. “BanglaTranslator.” Python. https://github.com/menon92/BanglaTranslator.

Mikheev, Andrei, Marc Moens, and Claire Grover. 1999. “Named Entity Recognition without

Gazetteers." In Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics -, 1. Bergen, Norway: Association for Computational Linguistics. https://doi.org/10.3115/977035.977037.

Miltenburg, Emiel van. (2016) 2022. "Dutch Tagger." Python. https://github.com/evanmiltenburg/Dutch-tagger.

Mishra, Upendra, and M Tech. 2012. "MAULIK: An Effective Stemmer for Hindi Language" 4 (05).

Mitkov, Ruslan, ed. 2014. The Oxford Handbook of Computational Linguistics 2nd Edition. 2nd ed. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199573691.001.0001.

Mitra, Moitreyee, Dipendranath Mitra, Moitreyee Mitra, and Dipendranath Mitra, eds. 2017. English-English-Bengali Dictionary. Oxford, New York: Oxford University Press.

Moltmann, Friederike. 2022. "Natural Language Ontology." In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta and Uri Nodelman, Winter 2022. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2022/entries/natural-language-ontology/.

Mridha, M.F., Abu Quwsar Ohi, M. Ameer Ali, Mazedul Islam Emon, and Muhammad Mohsin Kabir. 2021. "BanglaWriting: A Multi-Purpose Offline Bangla Handwriting Dataset." Data in Brief 34 (February): 106633. https://doi.org/10.1016/j.dib.2020.106633.

Mueller, Judith E., Maxime Woringer, Souleymane Porgho, Yoann Madec, Haoua Tall, Nadège Martiny, and Brice W. Bicaba. 2017. "The Association between Respiratory Tract Infection Incidence and Localised Meningitis Epidemics: An Analysis of High-Resolution Surveillance Data from Burkina Faso." Scientific Reports 7 (1): 11570. https://doi.org/10.1038/s41598-017-11889-4.

Mumin, Abdullah Al, Abu Awal Shoeb, Mohammad Reza Selim, and M Zafar Iqbal. n.d. "SUMono: A Representative Modern Bengali Corpus."

Mumin, Abdullah Al, Abu Awal Shoeb, Reza Selim, and M Zafar Iqbal. n.d. "SUPara: A Balanced English-Bengali Parallel Corpus."

Murthy, Rudra, Pallab Bhattacharjee, Rahul Sharnagat, Jyotsana Khatri, Diptesh Kanojia, and Pushpak Bhattacharyya. 2022. "HiNER: A Large Hindi Named Entity Recognition Dataset." arXiv. http://arxiv.org/abs/2204.13743.

Murugan, Selvakumar, Tamil Arasan Bakthavatchalam, and Malaikannan Sankarasubbu. n.d. "SymSpell and LSTM Based Spell- Checkers for Tamil."

Narang, Ashish, R. K. Sharma, and Parteek Kumar. 2013. "Development of Punjabi WordNet." CSI Transactions on ICT 1 (4): 349–54. https://doi.org/10.1007/s40012-013-0034-0.

Nashwan, Farhan, Mohsen Rashwan, Hassanin Al-Barhamtoshy, Sherif Abdou, and Abdullah Moussa. 2017. “A Holistic Technique for an Arabic OCR System.” Journal of Imaging 4 (1): 6. https://doi.org/10.3390/jimaging4010006.

Neechalkaran. (2019) 2022. “Tamil-Corpus.” https://github.com/neechalkaran/Tamil-corpus.

Nigam, Nitika, and Umesh Chandra Jaiswal. n.d. “Word Alignment of English-Hindi Parallel Corpus: Relative Study.”

Nordhoff, Sebastian, and Harald Hammarstrom. n.d. “Glottolog/Langdoc: Increasing the Visibility of Grey Literature for Low-Density Languages.”Nordhoff, Sebastian, and Harald Hammarström. "Glottolog/Langdoc: Increasing the visibility of grey literature for low-density languages." In the 8th international conference on language resources and evaluation [lrec 2012], pp. 3289-3294. ELRA, 2012.

Nowshin, Nafisa, Zakia Sultana Ritu, and Sabir Ismail. 2018. “A Crowd-Source Based Corpus on Bangla to English Translation.” In 2018 21st International Conference of Computer and Information Technology (ICCIT), 1–5. Dhaka, Bangladesh: IEEE. https://doi.org/10.1109/ICCITECHN.2018.8631947.

Ohidujjaman, Fahim Faysal, Shams Sumon, and Mohammad Nurul Huda. 2021. “Automatic Machine Translation for Bangla and English Resolving Ambiguities.” In 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), 27–32. DHAKA, Bangladesh: IEEE. https://doi.org/10.1109/ICREST51555.2021.9331085.

Oostdijk, Nelleke, and Hans van Halteren. 2013. “Shallow Parsing for Recognizing Threats in Dutch Tweets.”

Oregonian/OregonLive, Special to The. 2022. “The Quest to Save Oregon’s Kalapuya: ‘You Lose a Language, You Lose a Culture.’” Oregonlive. March 6, 2022. https://www.oregonlive.com/pacific-northwest-news/2022/03/the-quest-to-save-oregons-kalapuya-you-lose-a-language-you-lose-a-culture.html.

Pan, Jeng-Shyang, Shyi-Ming Chen, and Ngoc Thanh Nguyen, eds. 2012. Intelligent Information and Database Systems. Vol. 7198. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-28493-9.

Pandey, Amaresh Kumar, and Tanveer J Siddiqui. 2008. “An Unsupervised Hindi Stemmer with Heuristic Improvements.” In Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data, 99–105. Singapore: ACM. https://doi.org/10.1145/1390749.1390765.

Pandey, Pooja, Dhiraj Amin, and Sharvari Govilkar. n.d. “Rule Based Stemmer Using Marathi WordNet for Marathi Language” 5 (10).

Parida, Shantipriya, Ondřej Bojar, and Satya Ranjan Dash. 2019. “Hindi Visual Genome: A Dataset for Multi-Modal English to Hindi Machine Translation.” Computación y Sistemas 23 (4). https://doi.org/10.13053/cys-23-4-3294.

Patel, Chirag, Atul Patel, and Dharmendra Patel. 2012. “Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study.” International Journal of Computer Applications 55 (10): 50–56. https://doi.org/10.5120/8794-2784.

Pathak, Aditya Kumar, Himanshu Choudhary, and Rajiv Ratn Shah. n.d. “NMT Based Tamil Translation.”

Percival, P. 1993. Percival’s English-Tamil Dictionary. New Delhi: Laurier Books Ltd.

Poel, Mannes, and Egwin Boschman. n.d. “A Neural Network Based Dutch Part of Speech Tagger.”

Prasad, Arpana, and Neeraj Sharma. 2020. “Syntactically and Semantically Annotated Hindi Corpus for an Opinion Mining System.” In 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 36–42. Greater Noida, India: IEEE. https://doi.org/10.1109/ICACCCN51052.2020.9362761.

Prasad, Rohit, Shirin Saleem, Matin Kamali, Ralf Meermeier, and Prem Natarajan. 2008. “Improvements in Hidden Markov Model Based Arabic OCR.” In 2008 19th International Conference on Pattern Recognition, 1–4. Tampa, FL, USA: IEEE. https://doi.org/10.1109/ICPR.2008.4761446.

Publications, Islam International. 2019. The Holy Quran with Bengali Translation. Islam International Publications.

Puri, Rajeev, R. P. S. Bedi, and Vishal Goyal. 2015. “Punjabi Stemmer Using Punjabi WordNet Database.” Indian Journal of Science and Technology 8 (27). https://doi.org/10.17485/ijst/2015/v8i27/82943.

Qu, Xiaoye, Yingjie Gu, Qingrong Xia, Zechang Li, Zhefeng Wang, and Baoxing Huai. 2023. “A Survey on Arabic Named Entity Recognition: Past, Recent Advances, and Future Trends.” arXiv. http://arxiv.org/abs/2302.03512.

R, Prabhu. (2013) 2023. “Tamil Dictionary.” JavaScript. https://github.com/rprabhu/TamilDictionary.

Raha, Tathagata, Sainik Mahata, Dipankar Das, and Sivaji Bandyopadhyay. n.d. “Development of POS Tagger for English-Bengali Code-Mixed Data.”

Rahman, Chowdhury Rafeed, MD Hasibur Rahman, Samiha Zakir, Mohammad Rafsan, and Mohammed Eunus Ali. 2022. “BSpell: A CNN-Blended BERT Based Bengali Spell Checker.” arXiv. http://arxiv.org/abs/2208.09709.

Rajalakshmi, Ratnavel, Mohit More, Bhamatipati Shrikriti, Gitansh Saharan, Hanchate Samyuktha, and Sayantan Nandy. 2022. "DLRG@TamilNLP-ACL2022: Offensive Span Identification in Tamil UsingBiLSTM-CRF Approach." In Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, 248–53. Dublin, Ireland: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.dravidianlangtech-1.38.

Rajalingam, Damodharan. (2010) 2023. "An Affix Stripping Iterative Stemming Algorithm for Tamil." HTML. https://github.com/rdamodharan/tamil-stemmer.

Rajendran, S. n.d. "Parsing in Tamil: Present State of Art."

Raju, Madri Vijaya, and M. Sreenivasulu. 2022. "A Lightweight Stemmer for Telugu Language." In 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA), 1385–88. Coimbatore, India: IEEE. https://doi.org/10.1109/ICIRCA54612.2022.9985623.

Ramanathan, Madhu, Vijay Chidambaram, and Ashish Patro. n.d. "An Attempt at Multilingual POS Tagging for Tamil."

Ramesh, Gowtham, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, et al. 2022. "Samanantar : The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages." Transactions of the Association for Computational Linguistics 10 (February): 145–62. https://doi.org/10.1162/tacl_a_00452.

Ramteke, Swati, Komal Ramteke, and Rajesh Dongare. 2014. "Lexicon Parser for Syntactic and Semantic Analysis of Devanagari Sentence Using Hindi Wordnet" 3 (4).

Rao, Uma Maheshwar. n.d. "English-Telugu Machine Translation."

Rayall, G.S. 2002. English-Punjabi Dictionary. Punjabi University.

Redd, Mallamma V. 2014. "Semantical and Syntactical Analysis of NLP" 5.

Rifat, Md Jamiur Rahman. (2019) 2023. "Bengali-NER." Python. https://github.com/Rifat1493/Bengali-NER.

Sa, Pankaj Kumar, Sambit Bakshi, Ioannis K. Hatzilygeroudis, and Manmath Narayan Sahoo, eds. 2018. Recent Findings in Intelligent Computing Techniques: Proceedings of the 5th ICACNI 2017, Volume 3. Vol. 709. Advances in Intelligent Systems and Computing. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-10-8633-5.

Sai Kiranmai, G., K. Mallika, M. Anand Kumar, V. Dhanalakshmi, and K. P. Soman. 2010. "Morphological Analyzer for Telugu Using Support Vector Machine." In Information and Communication Technologies, edited by Vinu V Das and R. Vijaykumar, 101:430–33. Communications in Computer and Information Science. Berlin, Heidelberg: Springer Berlin

Heidelberg. https://doi.org/10.1007/978-3-642-15766-0_68.

Saini, H. S., Rishi Sayal, Aliseri Govardhan, and Rajkumar Buyya, eds. 2019. Innovations in Computer Science and Engineering: Proceedings of the Sixth ICICSE 2018. Vol. 74. Lecture Notes in Networks and Systems. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-13-7082-3.

Saini, Jatinderkumar R., and Jasleen Kaur. 2020. "Kāvi: An Annotated Corpus of Punjabi Poetry with Emotion Detection Based on 'Navrasa.'" Procedia Computer Science 167: 1220–29. https://doi.org/10.1016/j.procs.2020.03.436.

sainimohit23. (2019) 2022. "Hindi-Stemmer." Python. https://github.com/sainimohit23/hindi-stemmer.

Sampson, Geoffrey. n.d. "The Susanne Corpus." Accessed July 11, 2023. https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/public/susanne.html.

Sanderson, Mark. 1994. "Word Sense Disambiguation and Information Retrieval." In SIGIR '94, edited by Bruce W. Croft and C. J. Van Rijsbergen, 142–51. London: Springer London. https://doi.org/10.1007/978-1-4471-2099-5_15.

Sanjeev Kumar Sharma and Gurpreet Singh Lehal. 2011. "Using Hidden Markov Model to Improve the Accuracy of Punjabi POS Tagger." In 2011 IEEE International Conference on Computer Science and Automation Engineering, 697–701. Shanghai, China: IEEE. https://doi.org/10.1109/CSAE.2011.5952600.

Sankaravelayuthan, Rajendran. 2019. "English To Tamil Machine Translation System Using Parallel Corpus."

Saravanan, K, Ranjani Parthasarathi, and T V Geetha. 2003. "Syntactic Parser for Tamil."

Sarveswaran, Kengatharaiyer, and Gihan Dias. 2021. "Building a Part of Speech Tagger for the Tamil Language." In 2021 International Conference on Asian Language Processing (IALP), 286–91. Singapore, Singapore: IEEE. https://doi.org/10.1109/IALP54817.2021.9675195.

Sarveswaran, Kengatharaiyer, Gihan Dias, and Miriam Butt. 2021. "ThamizhiMorph: A Morphological Parser for the Tamil Language." Machine Translation 35 (1): 37–70. https://doi.org/10.1007/s10590-021-09261-5.

Schraagen, Marijn, Matthieu Brinkhuis, Floris Bex, M P Schraagen, M J S Brinkhuis, and F J Bex. n.d. "Evaluation of Named Entity Recognition in Dutch Online Criminal Complaints."

Sciforce. 2019. "NLP for Low-Resource Settings." Sciforce (blog). October 11, 2019. https://medium.com/sciforce/nlp-for-low-resource-settings-52e199779a79.

Semmar, Nasredine, and Christian Fluhr. n.d. “Arabic to French Sentence Alignment: Exploration of a Cross-Language Information Retrieval Approach.”

Shabdkosh.com. n.d. “English Hindi Dictionary | अंग्रेज़ी हिन्दी शब्दकोश.” SHABDKOSH. Accessed June 21, 2023a. https://www.shabdkosh.com/.

Sharma, Arnav, Sakshi Gupta, Raveesh Motlani, Piyush Bansal, Manish Shrivastava, Radhika Mamidi, and Dipti M. Sharma. 2016. “Shallow Parsing Pipeline - Hindi-English Code-Mixed Social Media Text.” In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1340–45. San Diego, California: Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1159.

Sharma, Dharam Veer. 2011. “Punjabi Language Characteristics and Role of Thesaurus in Natural Language Processing” 2.

Sharma, Richa, Sudha Morwal, and Basant Agarwal. 2022. “Named Entity Recognition Using Neural Language Model and CRF for Hindi Language.” Computer Speech & Language 74 (July): 101356. https://doi.org/10.1016/j.csl.2022.101356.

Shirsath, Nilesh, Aniruddha Velankar, Ranjeet Patil, and Shilpa Shinde. 2021. “Various Approaches of Machine Translation for Marathi to English Language.” Edited by M.D. Patil and V.A. Vyawahare. ITM Web of Conferences 40: 03026. https://doi.org/10.1051/itmconf/20214003026.

Shrivastava, Manish, and Pushpak Bhattacharyya. n.d. “Hindi POS Tagger Using Naive Stemming : Harnessing Morphological Information Without Extensive Linguistic Knowledge.”

SHUKLA, SHIVAM. (2018) 2023. “Machine-Translation-Hindi-to-English-.” Jupyter Notebook. https://github.com/shvmshukla/Machine-Translation-Hindi-to-english-.

Singh, Jaspreet, and Gurpreet Singh Lehal. 2015. “NAMED ENTITY RECOGNITION FOR PUNJABI LANGUAGE USING HMM AND MEMM.”

Singh, Jyoti, Nisheeth Joshi, and Iti Mathur. 2013. “Development of Marathi Part of Speech Tagger Using Statistical Approach.” In 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 1554–59. Mysore: IEEE. https://doi.org/10.1109/ICACCI.2013.6637411.

Singh, Priyanshu. (2020) 2022. “Sanskrit-Hindi-Machine-Translation.” Jupyter Notebook. https://github.com/priyanshu2103/Sanskrit-Hindi-Machine-Translation.

Singh, Shashank, and Shailendra Singh. 2021. “HINDIA: A Deep-Learning-Based Model for Spell-Checking of Hindi Language.” Neural Computing and Applications 33 (8): 3825–40. https://doi.org/10.1007/s00521-020-05207-9.

Singh, Vinay, Deepanshu Vijay, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. “Named Entity Recognition for Hindi-English Code-Mixed Social Media Text.” In Proceedings of the Seventh Named Entities Workshop, 27–35. Melbourne, Australia: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-2405.

Sitender, Seema Bawa, Munish Kumar, and Sangeeta. 2023. “A Comprehensive Survey on Machine Translation for English, Hindi and Sanskrit Languages.” Journal of Ambient Intelligence and Humanized Computing 14 (4): 3441–74. https://doi.org/10.1007/s12652-021-03479-0.

Sodimana, Keshan, Pasindu De Silva, Supheakmungkol Sarin, Oddur Kjartansson, Martin Jansche, Knot Pipatsrisawat, and Linne Ha. 2018. “A Step-by-Step Process for Building TTS Voices Using Open Source Data and Frameworks for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese.” In 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018), 66–70. ISCA. https://doi.org/10.21437/SLTU.2018-14.

Spyns, Peter, and Jan Odijk, eds. 2013. Essential Speech and Language Technology for Dutch: Results by the STEVIN Programme. Theory and Applications of Natural Language Processing. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30910-6.

Strik, Helmer, Walter Daelemans, Diana Binnenpoorte, Janienke Sturm, F. De Vriend, and Catia Cucchiarini. 2002. “Dutch HLT Resources: From BLARK to Priority Lists.” In 7th International Conference on Spoken Language Processing (ICSLP 2002), 1549–52. ISCA. https://doi.org/10.21437/ICSLP.2002-469.

Strizver, Ilene. 2011. “Confusing (and Frequently Misused) Type Terminology, Part 1.” fonts.com.

Subramanian, Malliga, Kogilavani Shanmuga Vadivel, and Antonette Shibani. n.d. “Detection Offensive Tamil Texts Using Machine Learning and Multilingual Transformers Models.”

Suresh, Dr V. n.d. “Reduced Tagset To Improve Accuracy Of HMM Based Parts Of Speech Tagger In Telugu Language.”

Thangarasu, M, and Dr R Manavalan. 2013. “Stemmers for Tamil Language: Performance Analysis.” Engineering Technology 4 (07).

ThaniThamizhAkarathiKalanjiyam. (2020) 2020. “Twn.” https://github.com/ThaniThamizhAkarathiKalanjiyam/twn.

Tiedemann, Jorg. n.d. “Improved Sentence Alignment for Building a Parallel Subtitle Corpus.”

Tiwari, Shailesh, Munesh C. Trivedi, Mohan Lal Kolhe, K.K. Mishra, and Brajesh Kumar Singh, eds. 2022. Advances in Data and Information Sciences: Proceedings of ICDIS 2021. Vol. 318. Lecture Notes in Networks and Systems. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-16-5689-7.

Trivedi, Ayushi, Navya Pant, Pinal Shah, Simran Sonik, and Supriya Agrawal. "Speech to text and text to speech recognition systems-Areview." IOSR J. Comput. Eng 20, no. 2 (2018): 36-43.

Trushkina, Julia, Lieve Macken, and Hans Paulussen. n.d. “Sentence Alignment in DPC: Maximizing Precision, Minimizing Human Effort.”

Tsujii, Junichi. "Lifetime Achievement Award." Computational Linguistics 47, no. 4 (2021).

Tsujii, Junichi. n.d. “Natural Language Processing and Computational Linguistics.” Computational Linguistics 47 (4).

Umale, Yogesh Vijay. 2016. “Dependency Framework for Marathi Parser.”

V., Dhanalakshmi, Padmavathy P., Anand Kumar M., Soman K.P., and Rajendran S. 2009. “Chunker for Tamil.” In 2009 International Conference on Advances in Recent Technologies in Communication and Computing, 436–38. Kottayam, Kerala, India: IEEE. https://doi.org/10.1109/ARTCom.2009.191.

Van Halteren, Hans, ed. 1999. Syntactic Wordclass Tagging. Vol. 9. Text, Speech and Language Technology. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-015-9273-4.

Van Hee, Cynthia, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2018. “Automatic Detection of Cyberbullying in Social Media Text.” Edited by Hussein Suleman. PLOS ONE 13 (10): e0203794. https://doi.org/10.1371/journal.pone.0203794.

Van Toledo, Chaïm, Friso Van Dijk, and Marco Spruit. 2020. “Dutch Named Entity Recognition and De-Identification Methods for the Human Resource Domain.” International Journal on Natural Language Computing 9 (6): 23–34. https://doi.org/10.5121/ijnlc.2020.9602.

Vasantharajan, Charangan, Laksika Tharmalingam, and Uthayasanker Thayasivam. 2021. “Adapting the Tesseract Open-Source OCR Engine for Tamil and Sinhala Legacy Fonts and Creating a Parallel Corpus for Tamil-Sinhala-English.” ArXiv E-Prints, September, arXiv:2109.05952. https://doi.org/10.48550/arXiv.2109.05952.

Vashisht, Vineet, Aditya Kumar Pandey, and Satya Prakash Yadav. 2021. “Speech Recognition Using Machine Learning.” IEIE Transactions on Smart Processing & Computing 10 (3): 233–39. https://doi.org/10.5573/IEIESPC.2021.10.3.233.

Veaux, Christophe, Junichi Yamagishi, and Kirsten MacDonald. 2017. SUPERSEDED - CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit. University of Edinburgh. The Centre for Speech Technology Research (CSTR). https://doi.org/10.7488/DS/1994.

Venkatapathy, Sriram, and Aravind K. Joshi. 2007. “Discriminative Word Alignment by Learning the Alignment Structure and Syntactic Divergence between a Language Pair.” In

Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation - SSST '07, 49–56. Rochester, New York: Association for Computational Linguistics. https://doi.org/10.3115/1626281.1626288.

Verma, Ark, Vivek Sikarwar, Himanshu Yadav, Ranjith Jaganathan, and Pawan Kumar. 2021. "Shabd: A Psycholinguistic Database for Hindi." Behavior Research Methods 54 (2): 830–44. https://doi.org/10.3758/s13428-021-01625-2.

Verma, Sanjeev. 2021. "Since 2016, 4.7L People from Punjab Went Abroad for Jobs." The Times of India, March 26, 2021. https://timesofindia.indiatimes.com/city/chandigarh/since-2016-4-7l-people-from-punjab-went-abroad-for-jobs/articleshow/81698819.cms.

Vinitha, V. S., and C. V. Jawahar. 2016. "Error Detection in Indic OCRs." In 2016 12th IAPR Workshop on Document Analysis Systems (DAS), 180–85. Santorini, Greece: IEEE. https://doi.org/10.1109/DAS.2016.31.

Vishal, G., and Singh Gurpreet. 2011. "Named Entity Recognition for Punjabi Language Text Summarization." International Journal of Computer Applications 33: 28–32.

Wanjari, Nagmani, G.M. Dhopavkar, and Nutan B. Zungre. 2016. "Sentence Boundary Detection For Marathi Language." Procedia Computer Science 78: 550–55. https://doi.org/10.1016/j.procs.2016.02.101.

William D. Lewis and Fei Xia. 2010. Developing ODIN: A multilingual repository of annotated language data for hundreds of the world's languages. Literary and Linguistic Computing, 25(3):303–319.

Yadav, Divakar, Sonia Sanchez-Cuadrado, and Jorge Morato. 2013. "Optical Character Recognition for Hindi Language Using a Neural-Network Approach." Journal of Information Processing Systems 9 (1): 117–40. https://doi.org/10.3745/JIPS.2013.9.1.117.

Yadavally, Aashish. (2019) 2021. "Hindi-Speech-Corpus." Python. https://github.com/aashishyadavally/Hindi-Speech-Corpus.

Yehui, Chen, Wang Enliang, and Ji Liqin. 2016. "Embedded Speech Recognition System Design and Optimization." In 2016 Eighth International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), 266–69. Macau, China: IEEE. https://doi.org/10.1109/ICMTMA.2016.72.

Zakraoui, Jezia, Moutaz Saleh, Somaya Al-Maadeed, and Jihad Mohamed Alja'am. 2021. "Arabic Machine Translation: A Survey With Challenges and Future Directions." IEEE Access 9: 161445–68. https://doi.org/10.1109/ACCESS.2021.3132488.

Zarkar, Raghvendra. 2022. "Natural Language Processing." Medium (blog). January 17, 2022. https://medium.com/@raghvendra.zarkar18/natural-language-processing-65f82c8dd7e0.

Zeman, Daniel, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. n.d. “HamleDT: To Parse or Not to Parse?”

Zhang, Wenwen. n.d. “Neural Dependency Parsing of Low-Resource Languages: A Case Study on Marathi.”

Punjabi (India and Pakistan) - Language Snapshot

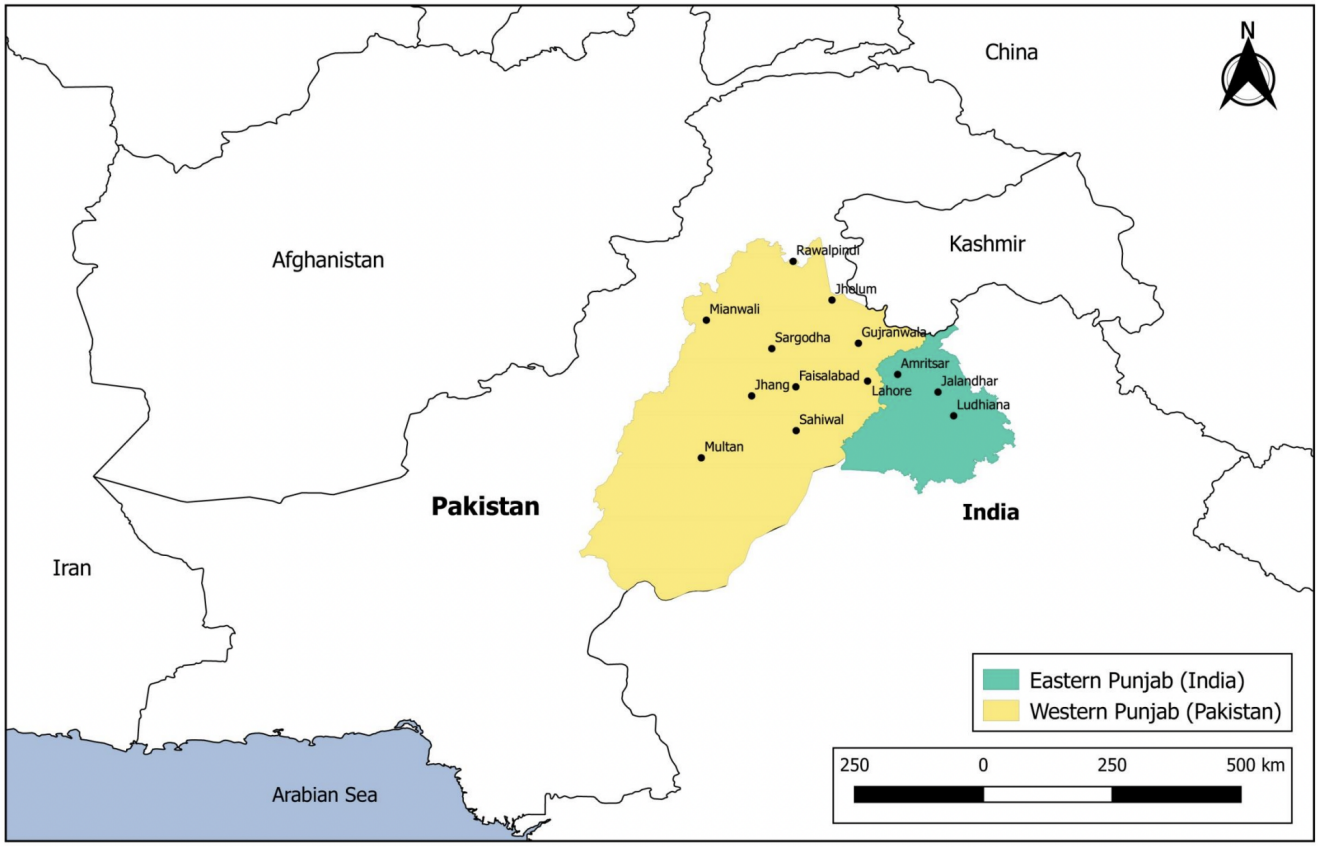


Figure 8: Map of Eastern (India) and Western (Pakistan) Punjab. This is a full scale version of a smaller map on page 10.