Adaptation and Development of Universal Dependencies for Punjabi (Shahmukhi) Script: Challenges and Linguistic Insights

Muhammad Shoaib Tahir
M.Phil Applied Linguistics, Government College University, Faisalabad
Email: shoaibtahir410@gmail.com
Minnaa Ahmad
M.Phil Applied Linguistics, Kinnaird College for Women, Lahore
Email: minnaa.ahmad90@gmail.com
Syeda Malika Zahra
Lecturer, National University of Modern Languages, Islamabad

Abstract

This paper explores the Universal Dependencies (UD) framework applied to Punjabi (Shahmukhi) through the development of a treebank, addressing both theoretical and practical aspects. Universal Dependencies is a standardized annotation scheme designed to enhance the development of multilingual parsers, facilitate cross-linguistic research, and promote consistency in syntactic annotation across different languages. Originating from the Stanford Dependencies and incorporating principles from Google's universal tagset and Interset, UD aims to provide a universal set of grammatical categories applicable across various languages. Punjabi, spoken primarily in Pakistan and written in the Shahmukhi script, presents unique challenges due to its less standardized orthography and script-related ambiguities. This paper discusses the adaptation of UD for Punjabi by addressing these orthographic issues, such as the absence of consistent diacritics, which can lead to significant semantic ambiguity. The project aims to offer a comprehensive linguistic resource for Punjabi in Shahmukhi script, thereby supporting research and applications in Natural Language Processing (NLP) and contributing to the broader UD ecosystem. By detailing the specific adaptations made for Punjabi, including handling of diacritics and grammatical features like word classes, gender rules, and tonal characteristics, this work seeks to enhance the utility and accuracy of the UD framework for low-resource languages. The paper highlights the importance of this adaptation for advancing multilingual parsing and syntactic analysis in less-resourced linguistic contexts.

Introduction

Universal dependencies play an important role in natural dependencies where it is responsible for enabling the development of multi-lingual parsers, cross-linguistic learning and research on parsing. Stanford (universal) dependencies (de Marneffe et al., 2006, 2008, 2014), google universal parts of speech tags (Petrov et al., 2012), and the interest interlinguas for morphosyntactic tag sets (Zeman, 2008) are the root of this annotation scheme. The main idea is to ensure the availability of a universal inventory of categories and a guiding principle to promote similar constructions in various other languages. The system was initially created in 2005 to assist in Recognising Textual Entailment systems by serving as a backend to the Stanford parser. It has now become the widely accepted standard for analysing the dependencies in English sentences. Over time, it has also been modified to work with various other languages (Chang et al., 2009, Bosco et al., 2013, Haverinen et al., 2013, Seraji et al., 2013, Tsarfaty, 2013, Lipenkova and Souček 2014).

The Google universal tag set was first used by Das and Petrov (2011) for unsupervised part-of-speech tagging. It was produced as a result of cross-linguistic error analysis conducted by McDonald and Nivre (2007) using the CoNLL-X shared task data. It has been generally accepted as a standard for converting different tagsets into a single standard. The Interset (Zeman, 2008) was first created to convert between the morphosyntactic tagsets of many languages. The term "cross-lingual delexicalized parser adaptation" was first introduced in 2006 and was then used in the initial research conducted by Zeman and Resnik in 2008. Subsequently, it was used as the morphological layer in Hamle DT (Zeman et al., 2014), a project that unifies many language treebanks using a common annotation system.

This work produces substantive definitions for universal POS categories that are not always precisely equivalent classes of categories in the underlying language-specific treebanks. Typically, the process of translating work into UD POS codes necessitates the use of context-sensitive guidelines or manual corrections. The UD morphological characteristics, which are universally present, are designed to offer a simplified fundamental set of characteristics that are most pertinent for research. The dependency representation of UD is derived from Stanford Dependencies (SD), which is in turn based on the principles of grammatical relations-oriented description that are prevalent in numerous language frameworks. In other words, it is fundamentally organised around concepts such as subject, object, clausal complement, noun determiner, noun modifying, and so forth. The new universal version aimed to improve the grammatical structures of typologically distinct languages by incorporating or enhancing relations, as some of the more eccentric and English-specific components of the previous version were removed. Consequently, the new taxonomy contains fewer relationships than the original SD.

Punjabi ranks tenth in terms of global language use. Moreover, the most often utilised language in Pakistan is this particular one. According to the 2008 Census Report of Pakistan, over 75% of the country's total population, which amounts to 76,335,300 people, speak Punjabi as either their primary or second language. The Punjabi language is divided into two main dialects: Eastern Punjabi, mostly spoken by Punjabi people in India, and Western Punjabi, primarily spoken by Punjabi people in Pakistan (Kumar et al, 2013; Kaur et al., 2010; Sharma, 2011). Pakistanis use the Perso-Arabic (Shahmukhi) script, whereas Indians utilise the Gurmukhi / Devanagari script (LRWiki, 2014; Saini & Lehal, 2011; Virk et al., 2011; Malik, 2006). The Punjabi language has historical connections to the Indo-Aryan languages (Dua et al., 2012; Gill & Lehal, 2008). Over time, the Punjabi language has included terms from Persian, Arabic, and Turkish. The alphabet also presents another problem. Punjabi lacks a standardised alphabet. Typically, it is transcribed using the alphabets of the Urdu language.

The Punjabi language, especially the one spoken in Pakistan, is considered to have less resources available for its speakers. Punjabi (Kumar et al, 2013; Kaur et al., 2010) is often neglected or receives little attention. The languages written in Gurmukhi and Shahmukhi scripts. The primary objective of this project is to provide an extensive repository of this language in Shahmukhi script. This resource may be used by language learners, users, and linguists to gather specific information for NLP applications. Shahmukhi is a script that is written in a right-to-left direction and is derived from the Nastalique style of Persian and Arabic writing. The morphology of letters inside a word is context-dependent, meaning that a letter will have a distinct form depending on whether it appears at the beginning, middle, or end of a word (Virk et al., 2011). This script consists of a total of thirty-eight letters, which include four long vowels: Alif (ا), Vao (و), Choti-ye (ی), and Badi-ye (ے); three short vowels: Zer (ِ), Pesh (ُ), and Zabar (َ); diacritical marks such as Shad (ّ), Khari-Zabar (ِ), do-Zabar (ً), doZer (ٍ), and the symbol hamza (ء). Ten consonants (بھ، پھ، تھ، ٹھ جھ، چھ، دھ، ڈھ، کھ ،گھ) are often aspirated, in contrast to the remaining six aspirates (لھ ،ڑھ ،رھ). The user's text is (مھ، نھ، وھ) Content written in Shahmukhi script often does not use short vowels and diacritical marks.

The presence or absence of diacritics, or the use of different diacritics, can alter the meaning of a word. This creates ambiguity for both machines and individuals who are not fluent in Punjabi. For example, the word ترنبcan be written in two different ways: رنب ت(meaning "to swim/float") and رنب ت(meaning "to walk/move"). Similarly, the phrase بیل translates to "the vine" whereas the word یل بtranslates to "the bull" in English. Therefore, diacritics are essential for eliminating the uncertainties between the meanings or senses of words (Virk et al., 2011). What was the approach used in the current implementation to address this? Extract the text

from the text box, remove any diacritics, and query the database for the diacritic-free word. If the word is discovered in the database, get the corresponding ID(s) and proceed to search for the same word with diacritics in the subsequent fields of the retrieved IDs. Alternatively, prompt the user to provide an alternative search term with diacritical marks (aerabs) so that the user may specifically choose the desired phrase. Otherwise, display all terms without any diacritical marks.

b) Punjabi Grammar

Therefore, we shall now examine several morphological and syntactic features of the Punjabi language. Punjabi, like Urdu, follows the canonical word order of Subject-Object-Verb (Sharma, 2011). The six cases are accusative, nominative, instrumental, ablative, dative, and locative. There are two sorts of adjectives: inflected and uninflected. There are also two types of affixes: prefix and suffix. Additionally, there are masculine and feminine forms, as well as singular and plural forms. The phrase "on the roof" is written as "تے" in Punjabi. Instead of using prepositions, it uses postpositions.

c) Punjabi-word Classes

Punjabi words exhibit both inflected and uninflected characteristics. Suffixes are primarily employed as inflections to convey grammatical information such as number, person, and tense. Nouns undergo inflection to indicate number and case. For instance, the word "منڈاُ,," is used for the singular form of "boy," while "منڈے,," is used for the plural form of "boys." At times, a word is capable of being utilised for both singular and plural forms based on the situation in which it is applied. For example, in the statement "منڈے تے ا چھت چڑھ گئے-" the word "منڈے" signifies the plural form, while in the statement "منڈےنے اپنب گال کٹ لیب-" it denotes the singular form.

d) Gender Rule for Punjabi Nouns

If a noun concludes in alif (ا), it indicates that the word is a masculine noun. Conversely, if its final character is Choti-ye (ی), the noun is feminine. In this context, the term "منڈاُ" signifies the masculine entity, while "کڑیُ" symbolises the feminine person. The vast majority of Punjabi words conform to this rule, with few notable exclusions.

f) Punjabi Adjectives

Additionally, adjectives may be classified as either inflected or uninflected. Punjabi adjectives are also inflected for singular and plural, masculine and feminine, and so forth. In accordance with the aforementioned regulations for the Noun category, Inflected Adjectives are also identified by their endings, gender, number, and the noun cases in which they qualify. Consequently, it is necessary to contact the appropriate adjective forum that is most compatible with the noun. For instance, the word "کبال" is a masculine adjective, and "بکرا" is a masculine noun. Similarly, "کبلی" is a feminine adjective, and "کبلی" is a feminine noun. These terms are consistent with the gender of the noun in each instance. For

instance, the adjectives that are uninflected are entirely constant and rigid, with a final consonant or vowel.

e) Tonal Features of Punjabi

Modern Punjabi is classified as a tonal language (Sharma, 2011) due to its phonetic nature, which means that the same word can be pronounced in a variety of ways, resulting in a distinct meaning or word. For example, the word کوڑا can be pronounced as KoRa (horse), KoRRa (leprosy/a disease), or KoRaa (whip). Inherent properties of word pronunciation are known as tonal/melodic features. Punjabi comprises three tones: high-falling / high tone, low-rising / low tone, and mid tone / level tone (Baart, 2003).

Literature review

The construction of a treebank requires various resources and tools. An annotation guide provides guidelines for annotators to follow during their work. A software tool is necessary to assist with the annotation process. In the case of semi-automated treebank construction, a part-of-speech (POS) tagger, morphological analyser, and/or syntactic parser are also required. Constructing trees manually is a laborious and prone-to-error procedure. A commonly used approach for constructing a treebank involves a blend of automated and human procedures, however the specific technique of implementation might differ significantly. While there are treebanks that have been annotated entirely by hand, the use of taggers and parsers to automate some tasks has made this approach less common in modern treebanking.

Annotated data holds an extremely important place for empirical research in linguistics and natural language processing. The constitutional or functional structural schemes are actually the basis for the kind of annotation the linguistic resource is done according to, in the treebank. The name treebank comes from the tree like shape of the linguistic constructions as they were first developed in the form of phrase-structure framework. For example; the Penn Treebank for English (Marcus et al., 1993).

Currently, dependency treebanks and functional annotation are the talk of the town. Other than this, the existing constitution-type treebanks have been improved by adding grammatical function type annotation in them. Tesnieére (1959) is to be credited for the starting research which further lead to dependency grammar formalisms. Dependency grammars focus on the lexical nodes only while ignoring the phrasal ones. The directed binary relations are then linked with these lexical nodes. The languages that have a free word order most commonly utilize the dependency format for building treebanks. Such languages include Basque, Czech, German, Turkish. While the languages like English deploy constituent formalism while building a treebank. Dependency annotation offers a suitable interface between syntactic and semantic representation, thus its reasons differ from

the fact that the type of structure is the one needed by many, if not most, applications. Dependency structures can also be automatically converted into phrase structures should required, even though they are not always perfectly accurate (Lin, 1995; Xia and Palmer, 2000). A treebank with both phrase structure and dependency annotations is the 50,000 sentence TIGER Treebank of German, a free word order language (Brants et al., 2002).

Related treebanks-Indo Aryan languages

Work that is relevant NLP applications depend extensively on linguistically annotated materials, which fulfil a variety of objectives, ranging from the examination of linguistic theories, the development and assessment of parsing technologies, and the provision of understanding of specific linguistic phenomena of a language (Nivre and Zeman, 2020). Nevertheless, the Indo-Aryan (IA) languages are devoid of high-quality digital instruments due to the scarcity of available corpora. The main languages are represented in these treebanks: Urdu (Bhat and Sharma, 2012), Hindi (Tandon et al., 2016) and Punjabi (Arora, 2022). Furthermore, there are automated conversions of Urdu (Ehsan and Butt, 2020) and Hindi (Bhat et al., 2018) treebanks from constituent annotations. Saraiki has conducted minimal research in the field of NLP. Alam et al. (2023) have developed a morphological analyser for Saraiki, while Asghar et al. (2021) have developed a part of speech (POS) tagger. Additionally, Sarghoda University is currently conducting research on a Saraiki wordnet as part of the Higher Education of Pakistan's funding (Gul et al., 2021); however, the system has not yet been released. It is equally crucial to comprehend the linguistic phenomenon of a language in order to develop NLP-related tools. Bashir and Conners (2019) have published a descriptive grammar for Saraiki, which we utilised as the foundation for our treebank annotations.

In 2013, a collaboration was established, headed by IIIT Hyderabad, to initiate a project financed by TDIL, Government of India, titled The Development of Dependency Treebank for Indian Languages. The primary goal of this project was to revive the annotation effort for monolingual and parallel treebanks in languages such as Hindi, Marathi, Bengali, Kannada, and Malayalam. Besides Universal Dependencies (UD), there exists a significant body of research on syntactic annotation in many South Asian languages that utilises Paninian karaka formalisms (Bhatt et al., 2009; Bhat and Sharma, 2012; Nallani et al., 2020). The Pāṇinian Kāraka Dependency annotation technique, as described by Bharati et al. (2006), was used to implement the treebank model. The annotation approach was previously used to annotate data in Telugu, Urdu, and Kashmiri languages (Begum et al., 2008; Husain et al., 2010; Bhat, 2017).

In the Universal Dependencies framework, Sanskrit, Hindi, Urdu, Marathi, Tamil, and Telugu have treebanks and parsers accessible as of UD version 2.5. These resources are mentioned in the works of Zeman et al. (2019), Straka & Straková (2019). Nevertheless, there are a few smaller

manual treebanks available for less commonly studied languages such as Marathi (Ravishankar, 2017), Bhojpuri (Ojha and Zeman, 2020), Sanskrit (Hellwig et al., 2020; Dwivedi and Zeman, 2017), Ashokan Prakrit (Farris and Arora, 2021), and code-mixed Hindi–English (Bhat et al., 2018). Research in Natural Language Processing (NLP) has resulted in the creation of various tools and systems for the Bhojpuri language. These include a statistical POS tagger (Ojha et al., 2015; Singh and Jha, 2015), a dictionary that can be read by machines (Ojha, 2016), a tool for identifying the language (Kumar et al., 2018), a system for translating from Sanskrit to Bhojpuri (Sinha and Jha, 2018), and most recently, a system for translating from English to Bhojpuri (Ojha, 2019).

Universal Dependencies for other Low-resource languages

The development of UD treebanks for a number of other low-resource languages has been observed, including Yorùbá (Ishola and Zeman, 2020), Latin Treebank for UD (Cecchini et al., 2020), Hittite (Andersen and Rozonoyer, 2020), Manx Gaelic (Scannell, 2020), Laz (Tụrk et al., 2020), Albanian (Toska et al., 2020), and others. Dash et al. (2021) recently reported the establishment of a treebank in Santhali, an additional low-resource language spoken in India. Nevertheless, the absence of grammatical descriptions and, hence, a reference point for determining the analysis required to provide the dependency relationships is one of the most significant obstacles to the construction of treebanks for a significant number of low-resource languages. There are numerous treebanks that could be essentially categorised into two categories. Treebanks of well-known and well-described languages, such as Hindi, English, and French, as well as treebanks of lesser-known and sparingly described languages. Magahi and Braj are included in the second category of languages that are sparingly described. There is a dearth of linguistic research on these languages, as neither of them has an exhaustive grammatical description or a dictionary. The following are a few linguistics studies that have contributed to the description of Magahi: a basic (albeit not entirely accurate) description of Magahi is provided by Shila Verma (Verma and Verma, 1983; Verma, 1985), a description of the Magahi case system is provided by Lahiri (2021, 2014; Kumar et al., 2014), a discussion of the Magahi honorific system within the minimalist framework is conducted by Alok (2021), the morphosyntactic properties of the nominal particle -wa are examined by Alok (2014), and the spatial postpositions of Magahi are examined by Alok (2012). As far as we are aware, the sole contemporary linguistic research on Braj pertains to its ergativity within the minimalist framework (Chandra and Kaur, 2020 a,b).

Data Analysis

1. Text Preparation:

Text preparation is a fundamental stage in linguistic analysis that involves several meticulous steps to ensure that the corpus is ready for thorough examination. Initially, the process begins with the extraction of the

corpus, where text data is gathered from a variety of sources. This extraction can involve collecting text from books, articles, web pages, or transcriptions of spoken language. The purpose is to compile a dataset that accurately represents the linguistic features or phenomena under investigation. For instance, if the research focuses on a particular genre of literature, texts from that genre are collected to form a representative sample.

Once the text is extracted, it must be converted into a format that is suitable for analysis. This conversion involves transforming the raw text into a structured format that can be easily processed by linguistic tools and software. Typically, the text is converted into plain text (.txt) or more structured formats like XML or CSV. This step also includes cleaning the text by removing any extraneous elements such as headers, footers, or embedded formatting codes that are not relevant to the analysis. This ensures that only the core content of the text is retained, making it easier to work with.

After conversion, proper sentence segmentation is crucial. This involves dividing the text into individual sentences to facilitate detailed analysis. Sentence segmentation is achieved by identifying sentence boundaries, which are generally marked by punctuation such as periods, exclamation points, and question marks. However, this process is not always straightforward, as punctuation marks may also occur in contexts that do not indicate the end of a sentence, such as within abbreviations or in certain types of formatting. Therefore, advanced tools and algorithms, like those provided by sentence tokenizers in natural language processing libraries, are used to accurately segment sentences. In cases where automated tools might fall short, manual review may be necessary to ensure that sentence boundaries are correctly identified, especially in texts with complex structures.

The process of text preparation ensures that the corpus is meticulously organized and formatted, making it suitable for subsequent linguistic analysis. By carefully extracting, converting, and segmenting the text, researchers set the stage for accurate and reliable analysis of linguistic features. This preparation is crucial for producing valid results and insights from the linguistic data.

2. Part-of-Speech (POS) Tagging:

Each word in the corpus was tagged with its respective part of speech. This includes nouns, verbs, adjectives, adverbs, pronouns, conjunctions, prepositions, and particles.

3. Morphological Analysis:

Morphological features such as tense, aspect, mood for verbs, and case, number, gender for nouns and pronouns were annotated.

4. Dependency Parsing:

Relationships between words in each sentence were annotated, identifying the head of each word and the type of syntactic relation (e.g., subject, object, and modifier).

Part-of-Speech Distribution

The analysis of the extracted corpus reveals insightful patterns regarding the distribution and frequency of various parts of speech, offering a comprehensive understanding of its linguistic structure. Nouns emerge with high frequency throughout the text, underscoring a focus on detailed descriptions and a multitude of references to entities such as people, places, and objects. This prevalence suggests that the text is rich in concrete and abstract references, providing a robust framework for discussing and elaborating on specific subjects and their characteristics. The prominence of nouns highlights a narrative or expository style where naming and identifying elements are central to the content.

Verbs, on the other hand, exhibit a diverse range of forms, illustrating the dynamic nature of the text. The variety of verb forms indicates a narrative that encompasses a broad spectrum of actions, events, and states of being. This diversity reflects the text's engagement with different temporal and aspectual dimensions, thereby allowing it to convey a wide array of activities and processes. The richness in verb forms suggests an emphasis on the progression and development of actions, which is crucial for creating a vivid and engaging depiction of events or arguments within the text.

Adjectives are employed to enhance the descriptive quality of nouns, contributing to a more nuanced and detailed portrayal of the entities discussed. The frequent use of adjectives indicates a deliberate effort to provide in-depth descriptions, which helps in painting a more comprehensive picture of the subjects. This detailed descriptive approach allows readers to gain a better understanding of the attributes and characteristics of the nouns, thus enriching the overall texture and depth of the narrative or exposition.

Pronouns are utilized to indicate the subjects and objects within sentences, facilitating fluidity and coherence in the text. The variety of pronouns found in the corpus demonstrates their role in referring back to previously mentioned entities, avoiding repetition, and maintaining clarity in the discourse. Pronouns help in managing the relationships between different parts of the text, ensuring that the narrative or argument remains coherent and interconnected.

Conjunctions play a crucial role in linking clauses and phrases, which contributes to the complexity of sentence structures within the text. The use of conjunctions shows a sophisticated approach to constructing sentences, allowing for the combination of ideas and the establishment of logical connections between different segments. This connective function enables the text to present more elaborate and intricate arguments or narratives, reflecting a structured and cohesive approach to writing.

Dependency Relations

The dependency parsing analysis of the text provides a detailed view of its syntactic structure, highlighting several key observations about how different parts of the sentences are organized and related. Firstly, the identification of subjects (nsubj) and objects (obj) is a prominent feature throughout the sentences. This clear delineation underscores the relationship between actions and their participants. Subjects typically denote who or what is performing an action, while objects indicate who or what is receiving the action. The consistent identification of these elements demonstrates a well-structured narrative or argument, where the roles of different entities in relation to actions are explicitly defined, facilitating a clear understanding of the sentence's meaning.

Modifiers, including adjectives (amod) and adverbs (advmod), are instrumental in adding descriptive detail to the text. Adjectives modify nouns, providing additional information about their attributes or qualities, while adverbs modify verbs, offering more context about the manner or intensity of the actions. This modification enriches the text by creating more vivid and precise descriptions, which enhances the reader's ability to visualize and comprehend the nuanced aspects of the narrative.

Conjunctions, denoted by cc and conj, play a critical role in indicating coordination between clauses, which contributes to the complexity of the sentence structures. Coordinating conjunctions (cc) connect words, phrases, or clauses that are of equal syntactic importance, while conjunctive words (conj) link clauses or phrases in a way that shows their interdependence. The presence of these conjunctions reflects a sophisticated approach to sentence construction, allowing for the integration of multiple ideas and maintaining a coherent flow throughout the text.

Morphological Features

In languages described as morphologically rich, verbs undergo various modifications to indicate tense (when an action occurs), aspect (the nature of the action's progress or completion), and mood (the speaker's attitude towards the action). Nouns and pronouns in these languages are similarly inflected, meaning they change form to reflect grammatical features such as case (their role in a sentence, like subject or object), number (whether they are singular or plural), and gender (classifying them as masculine, feminine, or neuter). This extensive system of modifications allows for precise and nuanced expression within sentences.

Sentence Structure:

The example sentence in Urdu is: "میں سٹینخورڈ ہسپتال دے اک کمرے وچ پئی نے اَکھاں کھولیاں ہن۔"

Translation: "I am lying in a room in Stanford Hospital with my eyes open."

1. Subject and Main Verb:

The core of the sentence is the subject "میں" (I) and the main verb phrase "پئی نے" (am lying). The auxiliary verb "پئی" (am lying) indicates the present continuous tense, while "نے" here is an auxiliary marker used for present continuous tense.

2. Location and Modifiers:

The phrase "سٹینخورڈ ہسپتال دے اک کمرے وچ" (in a room in Stanford Hospital) functions as a locative phrase that provides information about the location. Here, "سٹینخورڈ ہسپتال" (Stanford Hospital) is the proper noun indicating the specific place, while "دے" (in) and "وچ" (in) are prepositions linking the location to the action.

3. Additional Information:

"اَکھاں کھولیاں" (eyes open) is an additional detail describing the state of the subject while lying. "اَکھاں" (eyes) is the noun, and "کھولیاں" (open) is the verb in this context.

Dependency Parsing:

Root of the Sentence:

The main verb phrase "پئی نے" (am lying) serves as the root of the sentence. The subject "میں" (I) is linked to this verb phrase, indicating who is performing the action.

Subject: "میں" (I) is the subject of the sentence, connected directly to the main verb phrase "پئی نے" (am lying).

Locative Phrase:

"سٹینخورڈ ہسپتال دے اک کمرے وچ" (in a room in Stanford Hospital) is a locative phrase describing where the action is taking place. It is connected to the verb phrase "پئی نے" (am lying) through prepositions "دے" (in) and "وچ" (in), which specify the exact location.

State Description:

The noun "اَکھاں" (eyes) and the verb "کھولیاں" (open) describe the state of the subject while lying. This descriptive phrase is linked to the main verb phrase, providing additional information about the subject's condition.

Dependency Relations:

"سٹینخورڈ" (PROPN) is the subject (nsubj) of "کھولیاں" (VERB).

"ہسپتال دے اک کمرے وچ" (NOUN phrase) acts as a locative modifier (obl) for "کھولیاں" (VERB).

"اَکھاں" (NOUN) is the object (obj) of "کھولیاں" (VERB).

Morphological Analysis:

In the sentence, the verb "کھولیاں" is in the past tense, indicating that the action of opening eyes has been completed. This verb reflects an action that has already occurred. The noun "اکھاں" is in its plural form, suggesting that the reference is to more than one eye. Pronouns play a crucial role in this sentence: "میں" is the first-person singular pronoun, denoting the speaker as the subject, while "نے" is used as a postpositional marker, adding emphasis to the action described.

Complex Sentence Structures:

Consider the example sentence: "کوشش دی پھڑن ہتھ دا اس نال ہتھ سجے اپنے میں ہاں۔ کردی"

In this sentence, "میں" is a pronoun functioning as the subject of the verb "کردی". The phrase "کوشش دی پھڑن" serves as the object of the verb "کردی". Within this phrase, "ہتھ" is the object of the verb "پھڑن". The noun phrase "اس ہتھ دا" acts as a possessive modifier for "ہتھ", specifying whose hand is being referred to. The sentence's structure is complex, with multiple components working together to convey the action of attempting to hold a hand with the speaker's right hand.

Conclusion

The analysis of the Punjabi text corpus using the Universal Dependencies framework reveals a rich syntactic and morphological structure. The detailed POS tagging, morphological analysis, and dependency parsing provide insights into the linguistic features of Punjabi. This framework can be used to further develop NLP tools and linguistic resources for Punjabi, enhancing the understanding and processing of the language.

References

- Aryaman Arora. 2022. Universal Dependencies for Punjabi. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC), pages 5705–5711.
- Atul Kr. Ojha and Daniel Zeman. 2020. Universal Dependency treebanks for low-resource Indian languages: The case of Bhojpuri. In Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation, pages 33– 38, Marseille, France. European Language Resources Association (ELRA).
- Baart, Joan LG. "Tonal features in languages of northern Pakistan." Pakistani languages and society: problems and prospects (2003): 132-144.
- Begum, R., Husain, S., Dhwaj, A., Sharma, D. M., Bai, L., and Sangal, R. (2008). Dependency annotation scheme for Indian languages. In Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II.
- Bharati, A., Chaitanya, V., and Sangal, R. (2006). Natural Language Processing: A Paninian Perspective. Prentice-Hall of India, New Delhi, India.
- Bhat, Irshad, Bhat, Riyaz A., Shrivastava, Manish, and Sharma, Dipti (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), pages 987– 998. Association for Computational Linguistics, New Orleans, Louisiana. doi:10.18653/v1/N18-1090.
- Bhat, Riyaz Ahmad and Sharma, Dipti Misra (2012). Dependency treebank of Urdu and its evaluation. In Proceedings of the Sixth Linguistic Annotation Workshop, pages 157–165. Association for Computational Linguistics, Jeju, Republic of Korea.

- Bhatt, Rajesh, Narasimhan, Bhuvana, Palmer, Martha, Rambow, Owen, Sharma, Dipti, and Xia, Fei (2009). A multi-representational and multi-layered treebank for Hindi/Urdu. In Proceedings of the Third Linguistic Annotation Workshop (LAW III), pages 186–189. Association for Computational Linguistics, Suntec, Singapore.
- Bornini Lahiri. 2014. A typological study of cases in eastern indo-aryan language. Ph.D. thesis.
- Bornini Lahiri. 2021. The Case System of Eastern Indo-Aryan Languages: A Typological Overview, 1 edition. Routledge India.
- Cristina Bosco, Simonetta Montemagni, Maria Simi. 2013. Converting Italian treebanks: Towards an Italian Stanford dependency treebank, In 7th Linguistic Annotation Workshop and Interoperability with Discourse.
- Daniel Zeman and et al. 2021. Universal dependencies 2.8.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University
- Daniel Zeman, and Philip Resnik. 2008. Cross-Language Parser Adaptation between Related Languages. In *Proceedings of IJCNLP 2008 Workshop on NLP for Less Privileged Languages*
- Reut Tsarfaty. 2013. A unified morpho-syntactic scheme of Stanford dependencies. In *Proceedings of ACL*.
- Ritesh Kumar, Bornini Lahiri, and Deepak Alok. 2014. Developing lrs for non-scheduled indian languages. pages 491–501.
- Riyaz Ahmad Bhat and Dipti Misra Sharma. 2012. Dependency treebank of Urdu and its evaluation. In Proceedings of the Sixth Linguistic Annotation Workshop (LAW), pages 157–165.
- Ryan McDonald, and Joakim Nivre. 2007. Characterizing the errors of data-driven dependency parsing models. In *Proceedings of EMNLP-CoNLL*.
- Singh, S. and Jha, G. N. (2015). Statistical tagger for Bhojpuri (employing support vector machine). In 2015 International Conference on Ad- 38 vances in Computing, Communications and Informatics (ICACCI), pages 1524–1529. IEEE.
- Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC*. (home page)
- Straka, M. and Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 88–99.