Konidioms Corpus: A Dataset of Idioms in Konkani Language

Naziya Shaikh, Jyoti Pawar, Mubarak Banu Sayed
Government College of Arts, Science and Commerce, Goa University, Independent Researcher
Quepem Goa India, Panjim Goa India, Goa India
naziya.gcq@gmail.com, jdp@unigoa.ac.in, muskantemp@gmail.com

Abstract

Konkani is a language spoken by a large number of people from the states located in the west coast of India. It is the official language of Goa state from the Indian subcontinent. Currently there is a lack of idioms corpus in the low-resource Konkani language. This paper aims to improve the progress in idiomatic sentence identification in order to enhance linguistic processing by creating the first corpus for idioms in the Konkani language. We select a unique list of 1597 idioms from multiple sources and proceed with a strictly controlled sentence creation procedure through crowdsourcing. This is followed by quality check of the sentences and annotation procedure by the experts in the Konkani language. We were able to build a good quality corpus comprising of 6520 sentences written in the Devanagari script of Konkani language. Analysis of the collected idioms and their usage in the created sentences revealed the dominance of selective domains like ‘human body’ in the creation and occurrences of idiomatic expressions in the Konkani language. This corpus is made publicly available.

Keywords: Idioms, Corpus, Konkani

1. Introduction

Idiomatic expressions are sensed with a different semantic meaning as compared to the meaning of the individual words in the expression. The nature of these expressions has posed a challenge to natural language processing systems. With the onset of deep learning methods, data has played an important part in simplifying the processing related to idioms in the sentences. To counter the problem of idiomatic sense processing, multiple datasets have been developed in high-resource languages like English. Some of the recent idioms related corpus in English language such as the MAGPIE corpus consist of more than 50000 sentences (Haagsma et al 2020). Although a large number of datasets consisting of idiomatic expressions are available in languages like English, a corresponding corpus in low-resource languages like Konkani are missing. Konkani is a language spoken by over 2.5 million speakers in the western coast of the Indian subcontinent (Registrar General and Census Commissioner - census, 2011). This language has been under-resourced, with regard to the availability of digital resources in the language. There is no idioms-related corpus developed for the Konkani language till date, to our knowledge. In this paper we aim to create a new corpus that incorporates a large range of idioms in the Konkani language. Our goal is to provide a ground for training and testing idiomatic sentences in Konkani language for the purpose of idiomatic sense identification and processing in order to enhance the natural language processing research in the Konkani language. This corpus is primarily intended for the task of automatic idiom recognition. This task is further intended to be used to improve the efficiency of Konkani to English language machine translations. This corpus can be used as a fine-tuning dataset for large language models to create applications like idiom-suggestion for Konkani language essay writing. This corpus will also act as a preliminary dataset for fine-tuning the language model for the automatic extraction of idiomatic sentences from a corpus in order to create a larger dataset of idioms in Konkani.

Crowdsourcing as a method for corpus creation has been used by MAGPIE corpus (Haagsma et al., 2020). This study made the workers annotate the data from the British National Corpus (BNC, Consortium et al., 2007) as the base dataset through the use of an interface. In this paper we have used a crowdsourcing approach for the creation of sentences from scratch as there are no related resources in the Konkani language available to our knowledge that can be used as a base dataset to extract remarkable number of instances containing idiomatic expressions. We further use a combined approach, where experts guide our crowdsourced work adding quality to the dataset. The corpus is termed as the “Konidioms Corpus” standing for Konkani language idioms corpus. The main contributions of this paper include the following:

- Describes the construction of the first (to our knowledge) dataset of idioms in the Konkani language.
- Provides analysis of frequency and usage distribution of the idioms in the Konidioms corpus that can reflect the overall pattern of idiomatic expressions in the language.
- Provides analysis of the domain distribution of idiomatic expressions as well as sentences in the Konidioms corpus.
- Provides an insight into the benefits, limitations and difficulties of using a combined approach of crowdsourcing and expert annotations for the creation and annotation of the Konidioms corpus.

This paper is divided into six sections beginning with the introduction that describes our motivation to create the dataset and our major contributions through this research. This is followed by the section on related work that mentions the current research with regard to idioms dataset, including the corresponding existing benchmarks and datasets in other languages. We further describe the process and experience of the corpus creation in the ‘Method’ section. This is followed by the section on ‘Konidioms Corpus’ where we mention the dimensions and the details of the created corpus. In the next section we proceed with the frequency and domain analysis of the collected idiomatic expressions, as well as the sentences in the corpus. We conclude in the last section by summing up the methods, process and analysis that we contributed.

2. Related Work

The availability of rich base corpus branching out into different genres in the high-resource English language has led to major work in creation of corpus containing idiomatic expressions. This work mainly focused on the extraction of sentences from the base corpus to create a new corpus, rich in the usage of the idiomatic expressions. The overall dimensions of the existing idiomatic corpus and their creation and extraction process is discussed briefly in this section.

Some of the recent idiom related datasets in English language include the following:

- IDIX Corpus - For the creation of IDIX corpus (Sporleder et al., 2010) worked on the extraction of 4022 idiomatic phrases from the British National Corpus (BNC, Consortium et al., 2007) using a select list of idioms. The extraction process involved pattern comparison, followed by manual supervision.
- The Potential Idiomatic Expression – PIE dataset (Haagsma et al., 2019) consists of around 20 thousand samples that were divided into 10 multiword classes.
- ‘SemEval 2013 Task 5b’ (Korkontzelos et al., 2013) provided the dataset with 4350 instances from UkWac (Ferraresi et al., 2008) base corpus for 65 different types of idioms. The instances in this corpus were annotated using labels – literal, idiomatic, both, unpredictable.
- PARSEME shared task dataset (Savary et al., 2017) - Carried out verbal multi-word expression annotation (with idiomatic expressions as a subset) for 18 languages that were divided into 4 groups - Balto-Slavic, Germanic, Romance and Others. The created corpora included 4.5 million tokens for training, out of which 52,724 were annotated as ‘verbal multi-word expressions’ and 900 thousand tokens for testing out of which 9,494 tokens were annotated as ‘verbal multi-word’ expressions.
- EPIE Dataset (Saxena and Paul, 2020) – This dataset includes around 25000 sentences extracted based on a list of 717 idioms. These instances contain both literally sensed as well as idiomatically sensed sentences along with their annotations.
- MAGPIE Corpus (Haagsma et al., 2020) is an open-source corpus containing more than 50000 instances extracted from BNC (BNC, Consortium et al., 2007) base corpus and annotated through the process of crowdsourcing using a fixed list containing 2007 idiom types in English.
- IMPLI dataset (Stowe, 2022) includes instances containing idioms and other multi-word expressions like metaphors prepared as paired statements with automated 24K silver pairs and 1800 manual pairs.
- ‘SemEval-2022 Task 2’ (Madabushi et al., 2022) – The idiomatically annotated dataset was provided for the shared task of identification of statements in the Portuguese and English languages as either idiomatic or literal. This corpus provided the previous and the next sentence in the context, along with each sentence containing the idiom.
- IMIL Parallel Dataset (Agrawal et al., 2018) - Idiom dataset in the Hindi language that shares the same Devanagari script with the Konkani language, includes a parallel corpus that collects the idioms in English and translates them into 7 different Indian languages, including the Hindi language.

3. Method

We used the following process to build the Konidioms dataset. First, we collected a fixed list of 1597 idiomatic expressions from multiple sources including collection from dictionaries, educational books as well as collection through conducted surveys of native Konkani language speakers (section 3.1). We further crowdsourced the task of creating Konkani language sentences with the collected idiom expressions to the native Konkani speakers (section 3.2.1). The crowdsourced work was analyzed for correctness by language experts (section 3.2.2). A different set of language experts further annotated the cleaned data and added idiomatic and literal sense labels to each of the sentences, along with other metadata details like domain and related keywords (section 3.2.3). The sentences obtained after completion of this process, along with their annotations, were aggregated to convert into a labelled dataset for idiomatically and literally sensed statements containing idiomatic expressions.

3.1 Collecting Konkani Language Idioms

The process of formation of Konidioms corpus began with the collection of idioms from multiple sources. We attempted to capture idiomatic expressions of multiple types to incorporate the essence of the idioms from general spoken dialogues in the Konkani language. The Konkani dictionary served as a major source for the collection idioms. Additional idioms from school and college level educational books were added to this list. To ensure complete representation of the spectrum of idiom usage in the Konkani spoken language, we conducted a survey of selected elderly Konkani speaking people who were known to use a rich set of idioms in their spoken Konkani. The idioms collected through this method were added to the originally collected list of idioms to obtain a total of 1597 idioms in Konkani. We fixed this list of idioms for the creation of our dataset.

A further type-based division of idioms in the Konkani language is not currently available with respect to lexical, semantic or syntactic categories. Therefore, we decided to create and contribute to a semantic type-based categorization by studying the collected idioms, with the guidance of Konkani language experts. The major criteria for this division, has been to semantically categorize the available idioms, while maintaining the consistency in type-division theories in English language as well as emphasizing the characteristic uniqueness in Konkani language idioms. The collected idioms were studied and classified into the four different categories based on their characteristics. The four categories of idioms in Konkani language were defined as:

3.1.1 Category 1: Sound and Reaction - based idiomatic expressions

The first category of idioms includes those idiomatic expressions whose first word does not have any individual meaning and is mostly made up of some kind of sound description. Usually, the literally sensed counterparts of such idioms are very rare. Consider for example the following idiom:

ठो जावप
.tʰo javəp
(Sound of gunfire) + (to occur)

'ठो' is assumed to be the sound of a gunfire/explosion (similar to 'bang' in English). In the idiomatic sense, this expression is used to refer to the event of 'failing at something'. Such expressions can be used for example to denote a student failing in an exam.

3.1.2 Category 2: Metaphor-based idiomatic expressions

Some of the idiomatic expressions in the collected list have evolved from indirect comparisons to certain living objects or their behaviours. Consider for example the following idiom:

कावळ्या आवय आसप
kavɭya avai asəp
crow + mother + (to-be / exist)
To be the mother of the crow

This idiom literally means 'to be the mother of the crow'. In the idiomatic sense it refers to a gossipmonger. In the comparative inception, it uses an indirect reference and considers the 'crowing' of the crow to be an unpleasant sound and compares it to gossiping.

3.1.3 Category 3: Partially idiomatic expressions

Those idiomatic expressions whose literal and idiomatic interpretations overlap to a certain extent are included in this category. Usually, one of the words in these idiomatic expressions, has some semantic relation with the literal meaning of the sentence. Consider the following expression as an example:

नांव बुडप
nãv buɖəp
name + drown
To drown one's name

The idiom word-wise means to 'drown one's name'. It semantically means 'to spoil the name and reputation'. The first word name is already giving a brief idea about the idiomatic sense, but the verb used along with it (to drown) does not semantically relate to the first word.

3.1.4 Category 4: Completely idiomatic expressions

These are expressions whose constituent words do not correlate at all to the idiomatic meaning of the sentence. Consider for example:

खोबरे जावप
kʰɔbrɛ̃ javəp
(crushed dry coconut) + (become/occur)
To become a crushed dry coconut

The idiomatic sense of this expression refers to being ruined or destroyed and has no connection with the constituent words that refer to the dry coconut.

The linguistic model for this classification was inspired from the idiom-type classification in English language by Fernando (Wray, 2000; Liu, 2008; Salih, 2017). This model divides idioms into three categories – pure-idioms, semi-idioms and literal idioms. Based on our study we also felt the need for metaphor-based category based on idiom type classification by Moon (McCarthy, 1998) that classifies the idioms based on their transparencies. But in the case of Konkani, the number of metaphor-based idioms contributed to only 1.5% of the total idioms and hence we did not feel the need to include further transparency classifications as divided by Moon (McCarthy, 1998). Structural classifications of idioms (David, 2008; O'Dell, 2002) like prepositional phrase, binomial and trinomial idioms were not considered for classification in Konkani language as we did not find such structural fixedness in Konkani idioms (for example binomial structure of "something + and + something" is very uncommon in Konkani idioms). We found that the sound and reaction-based idioms occurred very frequent in the Konkani language and we did not find a similar category in the other classifications in linguistics. Based on this observation, we added an additional 'sound and reaction based idiomatic expressions' category to our classification.

The percentage distribution of collected idioms according to the four categories is shown in table 1.

Idiom Category	Distribution
Sound and Reaction based idiomatic expressions	2 %
Metaphor-based idiomatic expressions	1.5 %
Partially idiomatic expressions	7 %
Completely idiomatic expressions	89 %

Table 1: Category-wise Distribution of Collected Idioms

Completely idiomatic expressions form a major part of the collected idioms list. The idioms from category 1 and category 2 – Sound and Reaction -based and Metaphor-based idioms will rarely possess literal counterparts. We can assume less than 96% of the total idiomatic expressions collected are more likely to be ambiguous (potentially idiomatic), meaning they can have either an idiomatic sense or a literal sense in any particular context.

3.2 Procedure for Dataset Creation

Due to the lack of resources for the under-resourced Konkani language, the creation of the idiomatic and literal sense sentences had to be taken up from scratch. This idiom corpus is, to our knowledge, the first to be developed in Konkani language. To create the sentences, we decided to use the crowdsourcing approach. We provided the crowdsourced participants with an interface that mentioned the list of potentially idiomatic expressions along with their meanings. The workers were allowed to create any number of grammatically correct sentences with the freedom to create sentences with either idiomatic meaning or literal meaning as long as any specified idiomatic expression from the list was included in the sentence. The participants were allowed to create the sentences where the component words of the idiomatic expression were not used together, but were scattered throughout the sentences. They had to choose the idiom from the list using the interface and type out formulated sentences in the textbox provided on the interface. The participants were instructed to create any valid sentence from their regular spoken language dialogues that they use at home, office or otherwise in daily life.

3.2.1 Appointing Crowdworkers

The selection of the crowdworkers was initially done through a simple online test as we needed only native Konkani language speakers that confirm to the selected dialect of this language in Devnagari script. Initially 10 participants were selected for this task based on the online test. After review of their initial work, 7 crowdworkers were retained for the creation of the Konidioms dataset, based on the quality check done by the experts (section 3.2.2). The retained workers were asked to work for at least 3 hours/day and create a minimum number of 50 sentences per day. They were paid Rupees 100/hour as wages. The work was assigned with a timeline of 20 days. After expert supervision of the crowdsourced work and management of some of the participants, the system became stable and was able to reliably create a minimum of 350 sentences per day. At the end of this crowdsourcing phase, we collected a total of 7060 sentences.

3.2.2 Managing the Reliability of Crowdworkers

The validity of the sentences created by the crowdsourced workers was evaluated by three experts for correctness. The experts were highly qualified post-graduates in the Konkani language with a prior working experience of minimum 8 years associated with the language. The experts were paid Rupees 300/hour for the evaluation task. They evaluated the sentences for grammatical correctness, correct inclusion of the idiomatic expressions in the created sentences and for ethical correctness of the sentences. Based on the feedback from the experts, certain crowdworkers who were consistently giving poor quality statements were discontinued from giving inputs in the dataset creation work. Other workers were made to incorporate the changes as advised by the experts, to maintain the correctness of the sentences in the dataset.

3.2.3 Annotation by Experts

The annotation of the sentences was carried out by two experts. Each expert assigned a label to every listing in the dataset that identified the meaning of the idiomatic expression used in the sentence as idiomatic or literal. The annotators were also asked to add the overall domain that the sentence with the idiomatic expression covered along with the related keywords. The two experts were qualified with a post-graduation degree in Konkani Literature and more than 10 years of teaching experience in the Language apart from being native Konkani language speakers. Each of the experts was paid Rupees 300 per hour for this task. An effective inter-annotator agreement was obtained with Cohen's kappa constant value of 0.92 for the annotation task of labelling sentences as idiom or literal. The different keywords and domains specified by the annotators were merged to create the final genre metadata.

We also carried out a separate evaluation of the sentences in the dataset, wherever the experts had difficulty annotating them, or where each of the annotators had assigned different labels to the sentences. Through the discussion between the evaluation committee and the annotators, the contentious sentences were divided into the following four categories.

- Ambiguous – This category included the sentences whose sense can be inferred as both idiomatic as well as literal. Such statements would need access to the previous and following statements in the context for interpreting the meaning correctly. Such sentences contributed to 9% of the inter annotator disagreements. As a solution to this category of statements, some of the sentences were retained after making minor modifications to the sentences to reduce their ambiguity. Some of the highly ambiguous instances were discarded.
- Unclear Interpretation – Around 6% of the idiomatic expressions in the Konkani language have more than one semantic meaning even when used in only idiomatic sense. The correctness of such sentences depends on what meaning the idiomatic expression is interpreted with. Due to this unclear usage some of the valid sentences with idiomatic sense were deemed as literal. This was solved by discussion with the annotators about the multiple meanings and the varying viewpoints were considered while annotating the sentences. Accordingly based on the context and usage of such expressions, the annotators decided whether to retain the label or update it.
- Discontinuous – There were instances in the dataset where idiomatic expressions were not used in contiguous space. The component words of the idiomatic expression were not used together as an expression, but were scattered throughout the sentences. For this

category of sentences, an additional flag called ‘continuous’ was added to every sentence in the Konidioms corpus. This decision was taken based on the fact that although some of the words appeared in between the idiomatic expression in the sentence, it still held on to the correct idiomatic or literal sense of the sentence. This may provide a good ground for testing of ‘token level’ word-wise tagging of each word in the sentence as idiom or literal. The label ‘continuous’ in each listing is assigned either a value ‘yes’ indicating the expression is used as a whole together in the sentence or it is assigned a value ‘no’ for discontinuous expressions in the sentence.

- Others – Some sentences were questioned based on the usage as being neither idiomatic, nor literal, but instead belonging completely to other multiword classes. Such sentences were discarded from the current dataset.

3.2.4 Crowdsourcing and Annotation Outcome

As a result of the complete creation and annotation process, a total of 6520 sentences were successfully included in the dataset. Out of the originally created 7060 instances by the crowdworkers, 438 sentences were discarded by the experts on account of incorrect grammar or non-inclusion of the correct idioms or due to ethical considerations. Around 102 sentences were discarded by the expert annotators owing to annotation difficulties. 92% of the originally created sentences are now a part of the final Konidioms Corpus after a thorough quality check.

4. The Konidioms Corpus

The 6520 datapoint instances finally constituted the Konidioms corpus in the form of training and testing sets. A single datapoint instance in the Konidioms corpus is shown in table 2. Every instance consists of the following –

- A unique identifier ‘Id’ – an integer number that identifies each instance in the dataset.
- The field ‘Idiom’ states the idiomatic expression used in the sentence in Konkani language.
- The field ‘Contextual_meaning’ provides the context-wise interpreted meaning(s) of the idiomatic expression in English.
- The ‘Sentence’ field is the sentence created using the specified idiomatic expression.
- ‘Usage_sense_label’ shows the final annotation label given by the annotators for the sentence. It can take only two values: ‘literal’ or ‘idiom’.
- The ‘Continuous’ field is used to indicate whether the entire idiomatic expression is used together in the sentence or it is split into separate tokens throughout the sentence. ‘Continuous’ can take two values: ‘yes’ or ‘no’. The value ‘yes’ indicates that the idiomatic expression is used as a whole expression and value ‘no’ indicates use of expression in a discontinuous manner.
- ‘Domain’ specifies the overall genre in the English language, of the instances given in the dataset.
- The field ‘split’ is used to denote the division of the sentences in the form of training and testing splits. A ‘split’ can take two values: ‘train’, ‘test’. The ‘train’ value indicates that the instance belongs to the training set. The testing sets are divided into two parts. The ‘test’ split instances consist of a combination of instances with some idioms included from the training set in a different sentence and some unseen idioms. Each of the two sets includes examples from all the four categories of idiomatic expressions.

Id:	1
Idiom:	जीब वाऱ्यार घालप d͡ʒib varjar g^{h}aləp (tongue) + (on-air) + (to-put) to put tongue on air
Contextual_ meaning:	speak with no apparent reason/ divulge secrets
Sentence:	आमची शेजान्न आपली जीब वाऱ्यार घालता amt͡ʃi ʃɛd͡zann apli d͡ʒib varjar g^{h}alta. Our neighbour-Female her tongue on-air puts. Our neighbour (Female) puts her tongue on air.
Usage_sense_ label:	Idiom
Continuous:	Yes
Domain:	Human Body
Split:	Train

Table 2: A single Datapoint instance of the Konidioms Dataset

More examples from the Konidioms corpus with varying complexities is shown in Appendix A. The dimensions of the created ‘Konidioms’ dataset are shown in Table 3.

Total Number of Idioms:	1597
Total Number of Sentences:	6520
Total Number of Unique Words:	11945
Total Number of Idiomatically Sensed Sentences:	4404
Total Number of Literally Sensed Sentences:	2116
Total Sentences with Discontinuous idioms:	148
Total Number of Domains:	17
Total Sentences in Train Split:	4991
Total Sentences in Test Splits:	1529

Table 3: Dimensions of the Konidioms Dataset

It was seen that within these 6520 instances in the dataset, idiomatic interpretations of expressions in the sentence were 35% more compared to the literal interpretations of the expression. 68% of the sentences created in the dataset were idiomatically sensed. This scenario is comprehensible as in the Konkani language, certain idiomatic expressions, especially from categories 1 and 2 rarely possess literal interpretations. This phenomenon is observed in the nature of the regularly spoken natural language Konkani. In the case of English idioms corpus as well, the dominance of idiomatically sensed sentences has been observed in the existing corpus (Haagsma et al., 2020). The Konidioms dataset covers a wide range of genres through large variations in the sentence domains of the instances.

5. Analysis of The Konidioms Corpus

5.1 Frequency Distribution

The crowdworkers were given freedom to write any number of sentences for the idiomatic expressions of their choice without any specific instruction or constraint on writing idiomatic interpretations or literal interpretations of the expressions. This method was followed to ensure the natural valid frequency distribution of the idioms, as the native speakers are more likely to remember and create sentences for the idioms that they hear more often in instances, while speaking the language in general. Therefore, the more the frequency of an idiom in the dataset, the more likely it is to be used in the spoken Konkani language. The overall frequency distribution of the idioms in the Konidioms dataset is shown in table 4.

Frequency Range	Number of Idioms
1 - 5	1262
6 - 10	204
11-15	85
16 – 20	24
21 – 25	18
26 – 30	2
>30	2

Table 4: Dimensions of the Konidioms Dataset

The distribution depicts that 80% of the idioms are used less frequently in the spoken dialogue-based Konkani language statements. Only 1% of the idioms have high occurrence in the language and these idiomatic expressions have contributed to the maximum number of sentences in the Konidioms dataset. This distribution confirms the general belief that although idioms are believed to be a rare occurrence, some small groups of idioms occur very frequently and can be usually found in most of the spoken language conversation dialogues.

The frequency distribution of the idioms category wise is shown in table 5. The maximum number of collected idioms belong to the Category 4 – completely idiomatic expressions. Idioms from this category have occurred most frequently and a total of 5801 instances of the Konidioms corpus belong to this category. Sound-based and Metaphor-based idiomatic expressions form only 2.5% of the Konidioms dataset. The frequency of occurrence for a single idiom in the Konidioms dataset is higher in partially idiomatic expressions. The overall instances in the Konidioms dataset that contain partially idiomatic expressions are comparatively less. This implies that although small number of partially idiomatic expressions exist, their corresponding usage in the Konkani language is very high.

Category	No of Idioms	Frequency Range	Total No of sentences
Sound and Reaction - based idiomatic expressions	31	1 – 13	97
Metaphor-based idiomatic expressions	24	1 – 7	67
Partially idiomatic expressions	112	1 – 21	555
Completely idiomatic expressions	1430	1 – 36	5801

Table 5: Category-wise Frequency Distribution

5.2 Domain Distribution

Domain categorization of the Konidioms dataset is carried out in two parts. The first part calculates the percentage distribution of the collected idioms over the specified list of domains. The second part measures the distribution of the domains across the sentence instances in the Konidioms dataset. We considered the list of 18 different domains based on our understanding of division in the corpus. This list of domains was provided to the experts. The experts were requested to assign a fitting domain from the list to each of the instances. Experts were allowed to assign more than one domain as per the variations in the sentence instance. Table 6 shows the distribution of domains across the idiomatic expressions into the selected 18 domain categories.

It can be inferred based on the distribution that the maximum number of idiomatic expressions in the Konkani language use the human body elements to denote some other concept in the idiomatic sense. The concrete concept of the existing human body is the most referred domain.

The next highest reference for the idioms belongs to the abstract category tilted 'Abstract Life/Soul Concept'. This domain incorporates all the abstract notions of human life – identity, name, soul, humanity, life and death, abstract emotions.

The next major set of idioms referred to the five elements of nature in multiple forms. The five elements of nature are traditionally defined as combination of air, water, fire, earth and sky elements. Use of these concrete reference elements usually corresponded to some form of human behaviour in the idiomatic senses.

8% of idiomatic expressions refer to the local traditions and beliefs that are practiced by the Konkani language speaking people and provide a cultural aspect to the usage of idioms in the sentence. It was also observed that most of the trees and animals referred to in the Konkani language idiomatic expressions contributing to 8% of the total idioms were also the local productions culturally linked to the land of the Konkani language use. For example, idioms referred to various parts of the coconut tree which grows in abundance across the west coast of India in these Konkani speaking regions.

Another major domain with regards to idiomatic expressions that contributed to 8% of the total expressions was human behaviour that denoted typical human characteristics during any interaction.

Domain	Distribution Pattern of Collected Idioms	Domain Distribution of Konidioms Corpus Instances
Human Body	32%	34%
Abstract Life/Soul Concept	14%	6%
Five Elements of Nature	9%	9%
Local Traditions and Beliefs	8%	7%
Human Behaviour	8%	5%
Food and Cooking	5%	5%
Sound and Reaction	5%	3%
Animals and Insects	4%	3%
Trees and Greenery	4%	4%
Place and Position	4%	4%
Time and Numbers	3%	2%
Habitation	3%	3%
Money	3%	2%
Clothing and Accessories	2%	2%
Metal Objects	2%	1%
Games and Hobbies	1%	1%
Colour	0.65%	1%
Temperature	0.5%	1%

Table 6: Distribution of Domains Across the Collected Idioms

The domain distribution pattern of idioms signifies the category division of 1597 collected idioms which have been formed over years over different cultural aspects. Whereas, the domain distribution of Konidioms corpus instances is the categorization of the sentences created in the corpus not including the idiom domain. The distribution difference tests the hypothesis that most of the sentences with idiomatic expressions belong to the domain that is different from the idioms in the sentence itself. The expected major variation in the domain distribution percentages did not occur. The domain categorization of Konidioms corpus instances took a very similar distribution as compared to the idiomatic expression distribution. The notable difference lied only in the remarkable reduction of more than 50% in the domain of 'Abstract Life/ Soul Concepts'.

6. Conclusion

We created the first idioms corpus in the low-resource Konkani language to enhance the research progress in idioms processing concepts in Konkani language. This paper describes the detailed process of creation of idioms corpus through a combined approach that used crowdsourcing in combination with expert supervision and annotation. We collected a rich set of 1597 idioms and studied and divided them into categories. We carried out further analysis of the idiomatic expressions in Konkani language with respect to these categories. Through the method of crowdsourcing, we were able to create a good quality corpus of 6520 instances thus establishing crowdsourcing as a viable method for Konkani language corpus creation. We also presented the analysis of the created corpus in terms of frequency and domain distribution. Through analysis we inferred the strong inclusions of domains like human body throughout the use of idiomatic expressions as well as sentence creations in the Konidioms corpus. We also observed that the relevance of certain percentage of idioms was limited to local traditions and beliefs of the Konkani speaking regions. We hope to pave a way to improvements in the machine translation applications and other natural language processing tasks through this dataset contribution to facilitate further research in low-resource languages like Konkani.

7. Acknowledgments

We are grateful to Dr. Rama D. Murkunde, Shenoi Goembab School of Languages and Literature, Goa University and Assistant Professor Edna Vaz, Government College of Arts, Science and Commerce, Quepem, Goa, India for assisting in translation of Konkani sentences in this paper into IPA style glossing.

8. Bibliographical References

Agrawal, R., Kumar, V., Muralidharan, V. and Sharma, D. (2018). No more beating about the bush: A Step towards Idiom Handling for Indian Language NLP. *In Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC 2018)*, Mayazaki, Japan, May. European Language Resource Association (ELRA).

Cacciari, C. and Tabossi, P. (2014). Idioms: Processing, structure and interpretation. Psychology Press. 19. *BNC Consortium et al. 2007. British national corpus*. Oxford Text Archive Core Collection.

David, C. (2008) A Dictionary of Linguistics and Phonetics. *Oxford: Blackswell.*

O'Dell, F. (2002). English Idioms in Use. Cambridge: *Cambridge University Press*.

Ferraresi, A., Zanchetta, E., Bernardini, S. and Baroni, M. (2008). Introducing and evaluating ukWaC, a very large Web-derived corpus of English. *In Proceedings of the 4th Conference on Web as Corpus Workshop (WAC-4) Can we beat Google?* Marrakech, Morocco.1 June 2008.

Haagsma, H., Bos, J. and Nissim, M. (2020). MAGPIE: A Large Corpus of Potentially Idiomatic Expressions. *In the Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 279–287 Marseille, 11–16 May 2020 copyrighted European Language Resources Association (ELRA), licensed under CC-

BY-NC

Haagsma, H., Nissim, M., & Bos, J. (2019). Casting a Wide Net: Robust Extraction of Potentially Idiomatic Expressions. *ArXiv, abs/1911.08829*.

Korkontzelos, I., Zesch, T., Zanzotto, F. and Biemann, C. (2013). SemEval-2013 Task 5: Evaluating Phrasal Semantics. In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47, Atlanta, Georgia, USA. Association for Computational Linguistics.

Liu, D., (2008). I*diom Definition and Classification, Book-Idioms*, 1st edition, ImprintRoutledge, pages-22, eBook ISBN9781315092843

Madabushi, H., Gow-Smith, E., Garcia, M., Scarton, C., Idiart, M. and Villavicencio, A. (2022). SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. *In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval2022)*. Association for Computational Linguistics.

McCarthy, M. (1998). *Spoken Language and Applied Linguistics*, Cambridge: CUP.

Salih, S. (2017). A NEW TAXONOMY OF ENGLISH IDIOMATICITY, *IMPACT: Journal of Modern Developments in Social Sciences Research (IMPACT: JMDSSR)* Vol. 1, Issue 1, Jun 2017,19-30 © Impact Journals.

Saxena, P., Paul, S. (2020). EPIE Dataset: A Corpus for Possible Idiomatic Expressions. *In: Sojka, P., Kopeček, I., Pala, K., Horák, A. (eds) Text, Speech, and Dialogue.* TSD 2020. *Lecture Notes in Computer Science*, vol 12284. Springer, Cham. https://doi.org/10.1007/978-3-030-58323-1_9

Sporleder, C., Li, L., Gorinski, P., and Koch, X. (2010). Idioms in context: The IDIX corpus. *In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Stowe, K., Utama, P. and Gurevych, I. (2022). "IMPLI: Investigating NLI Model's Performance on Figurative Language". *In the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. (vol 1 – long papers)*. May 2022, ACL, Dublin, Ireland, pages 5375- 5388.

Registrar General, Census Commissioner. (2011). "Statement 1: Abstract of speakers' strength of languages and mother tongues-2011". www.censusindia.gov.in Office of the Registrar General and Census Commissioner, India. Retrieved 7 June 2023.

Savary, A., Ramisch, C., Ricardo, S., Sangati, F., Vincze, V., Zadeh, Q., Candito M., Cap, F., Giouli, V., Stoyanova, I., Doucet, A. (2017). The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. *In the Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain, April 4 ©2017 Association for Computational Linguistics.

Tedeschi, S., Martelli, F. and Navigli, R. (2022). ID10M: Idiom identification in 10 languages. *Findings of the Association for Computational linguistics: NAACL 2022*, pages 2715-2726. July 10-15, 2022 ©2022 Association for Computational Linguistics.

Wray, A. (2000). Formulaic Sequences in Second Language Teaching: Principle and Practice. *Applied Linguistics*, vol.21, no.4, December. Pages 463-489.

Wright, J. (2002). *Idioms Organiser*. *Boston: Heinle*.

Appendix A: Example sentences from 'Konidioms' corpus

Some of the example instances included in the Konidioms Corpus are shown below.

Idiom 1:	बारा वाजप [bara vad͡zəp] (12 o'clock) + (to ring)
Idiom meaning:	Be in trouble / lose in business
Literal meaning:	Clock shows 12 O'clock
Sentence:	निशाक घरा येवपाक वेळ जालो म्हणून बापायन जाप विचरली तेन्ना ताचीं बारा वाजलीं niʃak gʰəra jeʊpak veɭ d͡zalɔ mʰəɳun bapajn d͡zap vit͡sɑrli tɛnna tat͡ʃĩ bara vad͡zlĩ. **Word-to-word**: (Nisha - proper name) home (to-come) time happen (that's why) father answer asked (that time) her (12 o'clock) rang. **Translation**: Nisha returned home late. When Nisha's father enquired about the delay, she realized that she is in trouble and panicked.
Label:	Idiom

Idiom 2:	कोंब फुटप [kɔm fuʈəp] (Bud/sprout) + (break out)
Idiom meaning:	Become mischievous / feel oversmart
Literal meaning:	Germination (of seeds)
Sentence:	आपलो पूत बरो शिकता म्हणून बापायन ताका मोबायल दिलो जाल्यार ताका कोंब फुटले. aplɔ put bərɔ ʃikta mʰəɳun bapajn taka mɔbajl gʰeun dilɔ d͡zaljar taka kɔm fuʈlɛ. **Word-to-word:** My son well/good studying (that's why) father (to him) mobile gave done (to him) sprout (broke out). **Translation:** Father thought his son is studying well and gifted him a mobile. After getting the mobile the son started getting mischievous and acted oversmart.
Label:	Idiom

Idiom 3:	हाताबोटार उडोवप [hatabɔʈar uɖovəp] (hand and fingers) + (to throw)
Idiom meaning:	Donate something (money / eatables)
Literal meaning:	Throw something on the hands and fingers
Sentence:	कोरोना काळांत सिनेमा जगताच्या फामाद कलाकारांनी उपाशी पडिल्ल्या लोकांच्या हाताबोटार उडयलें. korona kaɭãt sinema d͡ʒəgətat͡ʃa famad kəlakarãni upaʃi pəɖillja lɔkãɲt͡ʃa hatabɔʈar uɖəjlɛ̃. **Word-to-word**: Corona (time/during) cinema of-world famous artists hungry fallen of-people (hand and fingers) threw. **Translation**: During corona period famous artists from film industry donated generously to support poor hungry people.
Label:	Idiom

Idiom 4:	हौंटयेंत घालप [hõʈjẽt gʰaləp] (In the lap of a woman) + (to put)
Idiom meaning:	Entrust
Literal meaning:	Put something in the lap of a woman wearing a sari
Sentence:	गरीब आवयन आपल्या माणकुल्या भुरग्याचो फुडार बरो जावचो म्हण ताका त्या भुरगीं नाशिल्ले गिरेस्त बायलेच्या हौंटयेंत घालो. gərib avain aplya maɳkulya bhurgyak bhurgĩ naʃille girest bailet͡ʃa hõʈyet ghalɔ. **Word-to-word**: Poor mother (her own) small child’s future good happen (that is why) (to him/her) that rich child (not having) woman’s (sari garment over the lap) put. **Translation**: In order to make her child’s future bright, that poor mother entrusted her child to the custody of that rich childless woman.
Label:	Idiom

Idiom 5:	तोपी घालप [topi gʰaləp] cap + (to-put)
Idiom meaning:	Defraud / cheat
Literal meaning:	Put a head cap
Sentence:	विदेशांत सर्वीस दितां म्हणून तरणाट्याक तोपी घालून दलाल पळून गेलो. videʃãt sərvis ditã mʰəɳun tərɳaʈjak topi gʰalun dəlal pəɭun gɛlɔ. **Word-to-word**: (In foreign country) job giving saying/pretext (to youngsters) cap put agent ran went. **Translation**: On pretext of giving a job in foreign country, agent cheated youngsters and ran away.
Usage Sense label:	Idiom

Idiom 6:	उदक मारप [udək marəp] water + (to sprinkle)
Idiom meaning:	Put on weight
Literal meaning:	Throw water
Sentence:	बाप्पा पोरसांत वचून झाडांक उदक मारून आयलो. bappa porsãt vət͡sun d͡zʰaɖãk udək marun ajlɔ **Word-to-word**: Father (in the field) gone (to the plants) water sprinkled (has come back). **Translation**: Father went to the field and came back after watering the plants.
Label:	Literal

Idiom 7:	कात उडोवप [kat uɖovəp] skin + (to shed)
Idiom meaning:	Start life a new
Literal meaning:	Snake shedding skin
Sentence:	बंदखणींत आशिल्ल्या तरणाट्याक पुलीस अधिकाऱ्यान सुधारलो आनी आतां तो कात उडोवन नोकरी करपाक लागला. bəndkʰəɳit aʃillja tərɳaʈjak pulis ədʰikarjan sudʰarlɔ ani atã tɔ kat uɖoʊn nokri kərpak lagla. **Word-to-word**: (In Prison) existing youngster police officer improved and now he skin threw job (to do) joined. **Translation**: Police officer changed the mindset of the young prisoner and now he improved and started a new job.
Label:	Idiom