Konidioms Corpus: A Dataset of Idioms in Konkani Language

Naziya Shaikh, Jyoti Pawar, Mubarak Banu Sayed

The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation

(LREC-COLING 2024) , 20-25 May, 2024

Introduction

- Construction of the first dataset of idioms in the low-resource Konkani language.
- Combined approach of crowdsourcing and expert annotations for the creation and annotation process.
- Contribution to language grammar through division of idiomatic expressions into categories based on their properties.
- Analysis of domain and frequency distribution of the collected idiomatic expressions and sentences in the Konidioms corpus.

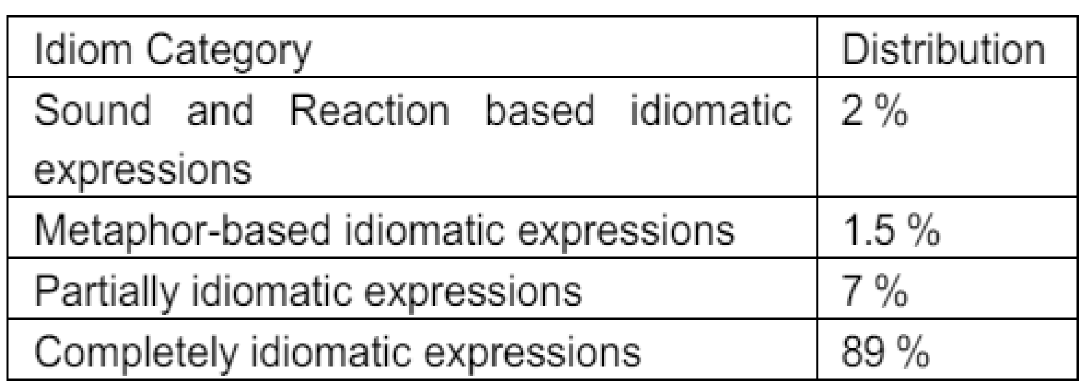
Applications

- A ground for training and testing idiomatic sentences in Konkani.
- Primarily intended for the task of automatic idiom recognition.
- To improve the efficiency of Konkani to English language machine translations.
- Fine-tuning dataset for large language models to create applications like idiom-suggestion for Konkani language essay writing.
- 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.

Dataset Construction Process

- Idiom Collection
  - Dictionaries, educational books, survey of elderly Konkani speaking people
- Appointing Crowdworkers
  - Test on basic Konkani language usage, interface
- Expert Quality Check
  - 438 sentences discarded out of 7060 sentences initially created through crowdsourcing
  - Criteria – Serious grammatical and semantic errors, non-inclusion of correct idiomatic expressions, ethical considerations
- Expert Annotations
  - 102 sentences discarded by experts due to annotation difficulties
  - Annotation issue categories- Ambiguous, Unclear Interpretations, Discontinuous, Others

Idioms Categories



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

Sound and Reaction - based idiomatic expressions

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

Completely idiomatic expressions

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

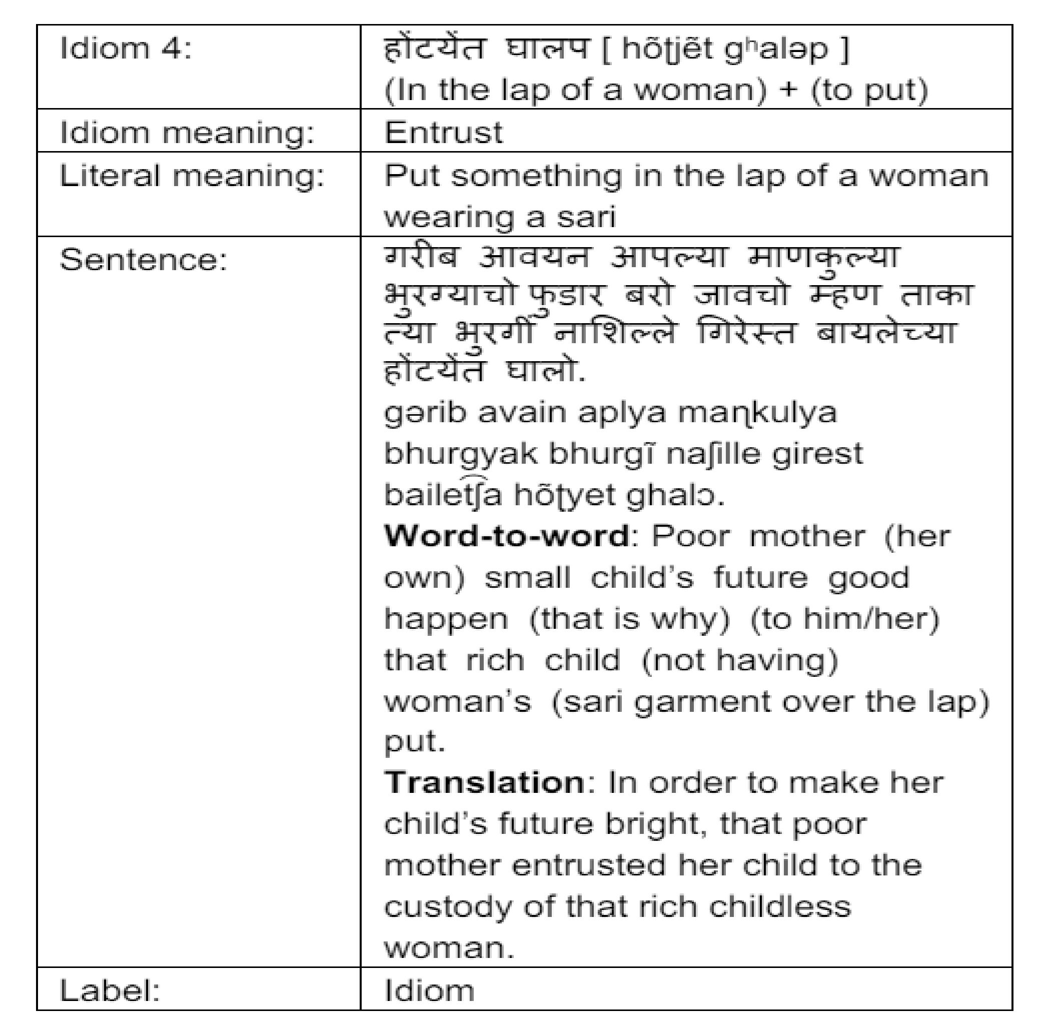
Metaphor-based idiomatic expressions

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

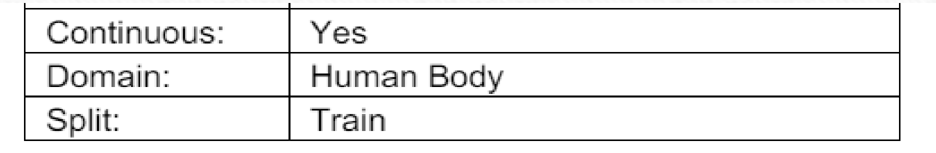
Partially idiomatic expressions

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

Konidioms Corpus



Idiom 4:	होंटयेंत 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 होंटयेंत 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



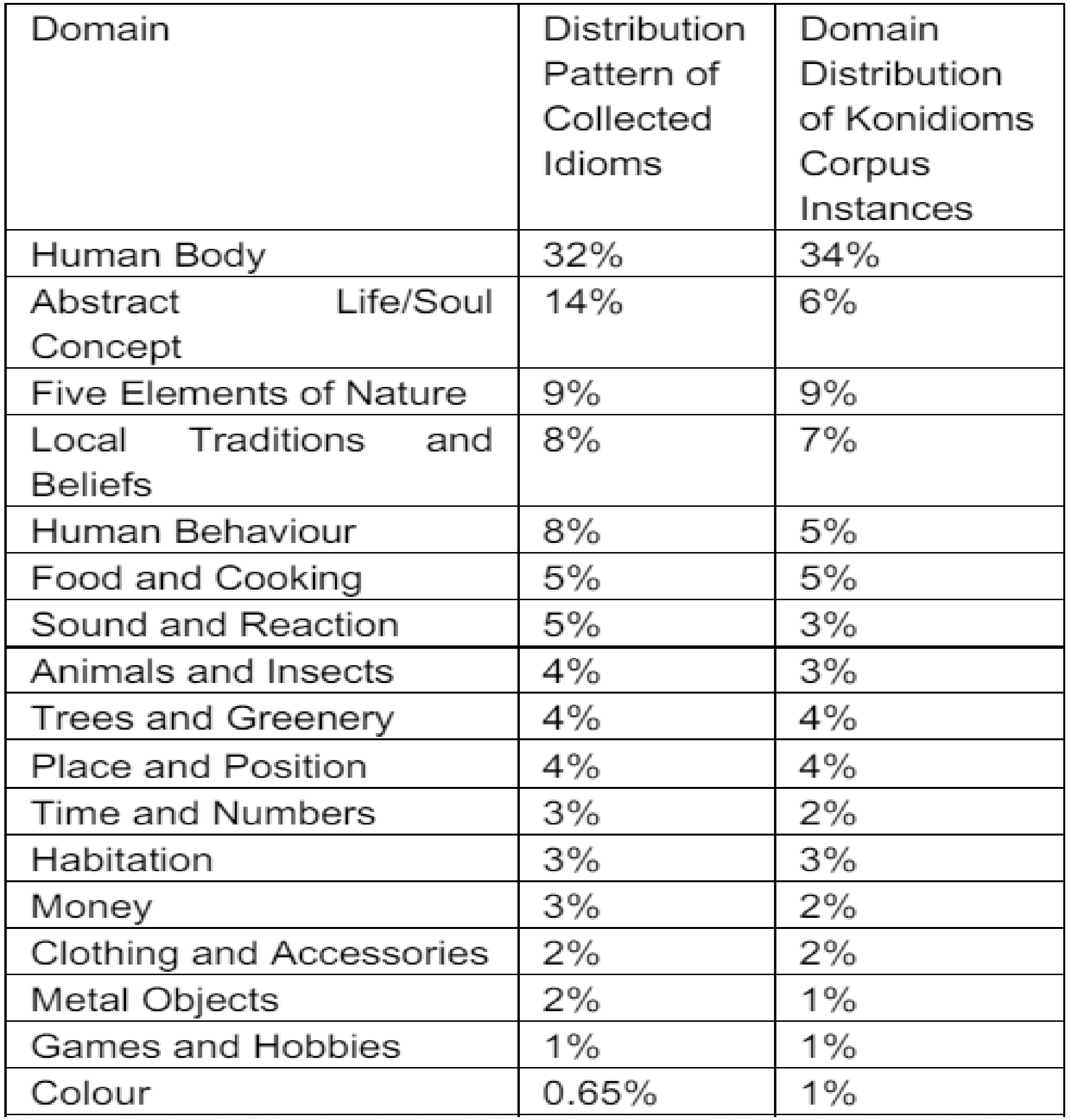
Continuous:	Yes
Domain:	Human Body
Split:	Train

Dimensions:
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

Category-wise Frequency Distribution

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

Domain Distribution



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%

Conclusion

- 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 to create a good quality corpus of 6520 instances.
- Rich set of 1597 idioms in Konkani language studied and divided into categories.
- Frequency and domain distribution analysis of the idiomatic expressions and the created corpus.
- 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.

THANK YOU