Cost Analysis of Linguistic Resources

e-published by
Central Institute of Indian Languages, Mysore

First Published
September, 2018

Author
Narayan Choudhary

ISBN No.: 978-81-7343-283-5

CIIL Publication No.: 1184

This is an open policy document subject to change as more feedback comes in. All feedback on this document may be sent to **n.choudhary [AT] gov.in**

Available only in e-format.

TABLE OF CONTENTS

LIST OT TABLES

ACKNOWLEDGEMENTS

This document has been prepared with feedback from several scholars working in the field. It would not be feasible to write the names of everyone in this document. However, the following must be named:

Dr. Pushpak Bhattacharya, IIT Patna/Bombay
Dr. Dipti Mishra Sharma, IIIT Hyderabada
Dr. Aadil Amin Kak, University of Kashmir
Dr. Karunesh Arora, CDAC, NOIDA
Shri Atul Saha, CDAC, Pune
Dr. Girish Nath Jha, JNU
Dr. Sobha L., AU-KBC, UoH
Smt. Swaran Lata, Meity, TDIL
Shri Vijay Kumar, Meity, TDIL
Dr. Niladri Shekhar Dash
Kalika Bali, MSRI, Bangalore
Dr. L. Ramamoorthy, CIIL
All PAC members of LDCIL
All LDCIL Staff
And many others

1 INTRODUCTION

The present document attempts to estimate a reasonable cost of creating a language resource or data set that would be developed by agencies/resource development personnels and organizations. This document gives only an estimated cost of creating such a resource and by no means intends to state that this would be the cost of a language resource.

The cost of creating a language resource depends on quite a lot of factors including the following:

1. Type of resource being created
2. Sources of the resources
3. Types of additional work done on the resources and ease of availability of such resources
4. Administrative and management cost along with a time factor

Linguistic Data Consortium for Indian Languages (LDCIL), a scheme of Govt. of India implemented by the Central Institute of Indian Languages, Mysore has been working on developing such resources since 2008 for all the scheduled languages as well as also in the Indian Sign Language. The scheme has so far developed a lot of raw corpora and annotated corpora in 21 scheduled languages of India.

Similarly, the Technology Development for Indian Language (TDIL) programme of Ministry of Electronics and Information Technology (MeitY) has been sponsoring several resource development programmes through various agencies in the country for several years. TDIL also has collected a huge amount of linguistic resources of various types that it has been distributing for non-commercial purposes for the past several years. However, due to lack of any policy decision with regard to the cost of these resources, these have not been made available for commercial consumption.

The lack of any clarity on how to decide the price of these resources have been a bottleneck for several years. To bring some clarity on the pricing of such resources, it was decided by the competent authorities in CIIL and TDIL that a document be prepared which analyses the contemporary cost of developing a language resource which would be used as the benchmark for deciding the cost of a language resource to be distributed through the government agencies for commercial consumption as well.

The author of this document in his capacity as Officer-in-Charge of Linguistic Data Consortium for Indian Languages (LDCIL), CIIL, Mysore was entrusted with the task of preparing a formula to arrive at the cost of the most universal types of language resources. The guiding principle for this formula would be the current cost if one wants to develop the same corpus afresh not take into consideration the actual cost that has already been incurred on developing such resources.

This document takes feedback from various stake holders including LDCIL, TDIL, CDAC, IITs, IIITs, central universities, industries and other entities who have worked or are working in the area of developing such resources.

2 TYPES OF LANGUAGE RESOURCES

There are mainly three types of language resources required in development of language resources. These are mainly text, speech and image that contain linguistic content and are required for various purposes of linguistic analysis done through computers.

These corpora can be further divided into sub-categories based on the types of annotations done on them. If no annotation is done on them, these corpora are called raw corpus. A brief on some of major types of language corpora currently being actively developed and in demand are given below.

1. Text Corpora
    a. Raw Text Corpus (Digitized)
    b. Raw Text Corpus (Santized)
2. Text Corpus with Value Addition
    a. PoS Annotated Text Corpus
    b. Chunked Text Corpus
    c. Dependency Text Corpus
3. Speech Corproa (Read Speech Dialogue)
    a. Raw Speech Corpus
    b. Sentence Segmented Speech Corpus
    c. Word Segmented Speech Corpus
4. Image Corpora (Corpus for OCR)
    a. Scanned Images
    b. Handwriting Recognition Corpus
5. Sign Language Corpus
6. Other resources
    a. Lexicon
    b. Dictionaries
    c. Thesaurus
    d. Other Cognitively networked resources (WordNets, etc.)

There might be different types of language resources depending upon the needs.

3 COST OF CREATING RAW TEXT CORPORA

Raw text corpus might seem easy to have been developed for most of the people, specially in recent years when the content on the Internet is growing exponentially. There might be an easy way out to say that such a corpus can be collected by crawling the internet. However, this does not hold good for all the languages. Many of the scheduled languages of India does not have enough content available on the internet. Therefore, it is often not possible to collect enough amount of corpus from the internet. The alternative, however, is to get the text already written (in print format), typed again and included into a representative corpus.

Thus, here we see a few steps in creating such a raw corpus. These steps also have bearing on the overall cost of the resource being created. Let's call these steps as factors deciding the cost of a text corpus.

3.1 FACTORS AFFECTING THE COST OF THE RAW TEXT CORPUS

As there are many factors, these factors have been noted in this section. All types of raw text corpora do need to involve these factors.

3.1.1 Data Source

The source of data/text included in the corpus may be from different sources. If the text is collected through crawling, it would cost less than if the text is collected manually and then typed. For this reason, the data source for now is divided into two categories of crawled and typed data.

3.1.1.1 Typed Text Data

It is a typical case for the Indian languages that digital data is not available while there are books published in print version. To have a representative text corpus from all genres, therefore, it is often required that the text are sourced in print format and then a language editor/typist is hired to type the selected text to be included into the corpus.

Text sourced using this method involve the cost of sourcing the data as well as typing them and proofing them.

For the sorting and indexing purposes, let's call it typed text data.

3.1.1.2 Crawled Text Data

For languges where digital text is availabel (such as English, Hindi and some other languages as well as specific domain text such as twitter feeds, facebook posts, comments on social networks and forums etc.), the text may be crawled. This type of corpus may be called the Crawled Text Data.

A crawled text data may or may not involve proofing costs. For the costing purposes, a corpus may have this information to justify the cost attached to a corpus.

3.1.2 Data Inputting Cost

If a corpus is a typed corpus, data inputting cost should be measured as per the current market rates or approved government rates.

To arrive at a cost for such a task, the current rates proposed have been derived from the rates of National Translation Mission (NTM) and that of the Bharatavani Project of Central Institute of Indian Languages. The cost of typing a page in Indian languages containing a total of 300 words as per NTM approved rates is Rs. 15. This comes to a cost of Rs. 0.05 per word.

However, the unit in the case of raw text corpus should be a character (UTF-8) which is equivalent to one keystroke and not a word because there are several Indian languages where the average character length of a word delimited by a white space is much higher than others. For example, in Malayalam, the average character length of a word is more than 11 while that for Hindi it is around 5. Given that the effort made in typing/inputting the raw corpus cannot be ignored, the reasonable unit for the raw corpus is to be counted as character.

Thus we have to arrive at the cost for per key stroke or character. Taking the cue from the NTM rates above where one word containing an average of 7 characters cost Rs. 0.05, the cost of a key stroke/character would be equivalent to the cost of one word divided by the average character length of a word.

For the purpose of simplicity across all Indian languages, the average word length has to be drawn from across languages. This exercise has been done using the available text corpus with LDCIL for all the languages (Excluding Sanskrit and Sindhi for which we do not have any corpus). Our analysis says that average word length across these languages stand at 6.55 characters per word. For simplicity sake, we have rounded it to the next decimal number of 7 (as used in all the Bharatavani Project tasks). So, we can consider that all the Indian languages contain an average of 7 characters/key strokes per word.

Thus the cost of inputting one UTF-8 character/keystroke equals to the cost of inputting a word divided by seven i.e. 0.05/7 which equals to Rs. 0.007.

3.1.3 Vetting/Proof Reading of the Text Data

Text typed by a typist may contain errors. Thus it may require further vetting or proof reading at the level of a language expert/language editor. Therefore, the cost involved in this step also has be counted while fixing the price of a text corpus.

The standards of proofing cost may be taken from various sources. For this proposal, we will consider the proofing cost as approved for Bharatavani project works because in the NTM's rate card, the number of words are not mentioned. Bharatavani's project's rates were also finalized through a committee at CIIL and is based on the rates as approved by the NTM's sub-committee on costing.

The cost of vetting/proof reading a 200 word document in Indian languages is fixed at Rs. 6. Here also we would consider the character as final unit for cost and we will consider that a word contains an average of 7 characters per word. Thus the cost of proofing one character in Indian languages would be equivalent to Rs. 6/200*7. This roughly equals to Rs. 0.004.

3.1.4 Storage and Management Cost

Roughly a 10% of the overall cost in typing and proofing would be earmarked as the storage and management cost to be levied on top of the other costs.

3.1.5 Formula to Arrive at the Cost of a Typed Corpus

Based on the discussions made above, the formula to arrive at the cost of a text corpus can be delineated as given in the table below. The ensuing table gives an example calculation for a raw text corpus that is typed and cleaned.

As crawled text corpus are not much in vogue for selling purposes, no formula has been drawn yet for this.

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

Table 1: Formula to Calculate the Cost of a Raw Text Corpus

Example Cost Calculation for a Corpus:

SL	Particulars	Values
1	Resource Type	Text
2	Corpus Type	Raw Corpus
3	Language	Hindi
4	Data Source	Typed+Cleaned
5	Unit	Character/Key Stroke
6	DTP Cost per Unit	0.007
7	Proofing Cost per Unit	0.004
8	Corpus Word Count	24765268
9	Corpush Character/Key Stroke Count	124486539

10	DTP Cost of the Corpus (=6*9)	871405.773
11	Proofing Cost of the Corpus (=7*9)	497946.156
12	Actual Cost of the Corpus	1369351.929
13	Total Cost of the Corpus	1506287.122

Table 2: Example Cost Calculation for a Raw Typed and Cleaned Text Corpus

3.2 Cost of Creating PoS Annotated Text Corpora

Parts of Speech annotated corpus are another type of text corpus for which text corpus is first sourced. The text corpus sourced for PoS annotation has an additional cost that is not included in the formula given below.

3.2.1 Factors affecting the cost of Parts of Speech Annotation

Below we discuss the factors affecting the cost of parts of speech annotation.

3.2.1.1 Type of Tagset Used

There are different types of tagsets with different sorts of complexity. Time taken to annotate a token is dependent upon the complexity of the tagset.

However, in the year of 2012, the Industry and academia in India proposed a BIS standard of a PoS tagset to be followed for all Indian languages with required language specific modifications. For the purposes of this analysis, we are at present taking the BIS tagset as the standard tagset and calculations are based on this tagset.

If a different tagset such as the MSRI tagset (Bali et. al. 2010), which also encodes morphological information such as gender number and person, the time taken may be different and on the higher side than that of the BIS tagset.

It is understood that the PoS annotation is done using some aid tools that helps enhance the speed of the annotations.

3.2.1.2 Whether Segmentation Required

Parts of Speech tagging is done on space separated tokens. It is expectd that each token belong to one of the parts of speech categories. However, this is not true of some languages whose morphological structure is of agglutinating nature. This means that the natural space separated tokens in these languages contain more than one parts of speech which need to be separated before parts of speech labels are attached to them.

This processing of separating the tokens is called segmentation. Amongst the Indian languages, segmentation is required in languages like Malayalam, Tamil and a few other languages. The task of segmentation can be done automatically however we are not aware of its accuracy. Therefore, for PoS annotation, the tasks of segmentation is usually done manually either before PoS annotation or during the PoS annotation itself.

The cost of PoS segmentation is another feature that needs to be taken into account while deciding the cost of PoS annotated corpus for languages where segmentation is required.

If segmentation is done during the process of annotation itself, the time and cost taken in this process may be included in the PoS annotation task itself.

3.2.1.3 Annotator Cost

The cost of annotation is usually measured in the amount of money spent on the annotation task. For this, it is essential to arrive at the cost of hiring an annotator who is a person trained in linguistics (having at least an MA or BA in linguistics) and having an expert level knowledge in the language concerned.

The cost of hiring such a resource may vary depending upon several factors. However, to arrive at a cost plan, we have taken our standards as seen in a government setup at Mysore (i.e. LDCIL scheme under CIIL, Mysore, Karnataka).

The output of PoS tagging would also depend quite a bit on the language being annotated. Thus, this factor would be subjective and depend upon the resource creator's discretion to fix.

3.2.2 Annotation and Validation Procedure

Any kind of human annotation usually requires validation before it is finalized. The process of coming at a consensus for such kind of annotation is called inter-annotator agreement. Usually each text is annotated by two annotators and for any kind of inter-annotator disagreement (reached by comparing the files), is handled by a third annotator called as arbitrator who may hold discussions in case of serious issues arise before finalizing it.

A corpus usually validated before it is released for such purposes. Therefore, the cost of validation would also add up as cost of corpus, if the released corpus is validated.

Thus the additional cost of validation would be calculated as follows:

Annotation Cost = Annotation cost by one annotator * 2

Validation cost (for time spent on resolving the inter-annotator differences): Half of the time spent on the Annotation cost

The validation cost is kept at half of the time spent on the actual annotation time cost spent by one trained annotator. This time may also vary depending upon the quality of the annotator and consensus amongst the team, apart from other subjective factors.

3.2.3 Formula to Calculate the PoS Annotated Text Corpus

Thus, a formula for such a PoS annotated corpus could be delineated as given in the table below:

SL	Parameter	Value
1	**Resource Type**	Text
2	**Corpus Type**	PoS Annotated
3	**Language**	
4	**Unit**	Token/ word (Segmented word, applicable)
5	**Total Number of Tokens/Words**	XX
6	**Total number of words/key strokes**	XX
7	**Annotator Cost Per hour**	Rs. 400 (This may vary based on price index changes. Current cost is taken as per the cost of hiring a contractual employee for such a task in a standard setup.)
8	**Annotated Word Per Hour**	300 (This may vary language wise as some languages require segmentation while others do not. The current example is taken for Kannada as reported at CIIL.)
9	**Cost Per Token**	=Annotator Cost per hour/Cost per token
10	**Actual Cost of a PoS Annotated Corpus**	
10.1	**One Time Annotation Cost (One time annotation by one annotator)**	Number of token*Cost per token
10.2	**Annotator Cost (Validated)**	One Time Annotation Cost * 2.5

Table 3: Formula to Calculate the Cost of a PoS Annotated Text Corpus

Example Calculation for a Kannada PoS Annotated Text Corpus

SL	Parameter	Value
1	**Resource Type**	Text
2	**Corpus Type**	PoS Annotated
3	**Language**	Kannada
4	**Unit**	Token
5	**Total Number of Tokens/Words**	719845
6	**Total number of words/key strokes**	6100640
7	**Annotator Cost Per hour**	400
8	**Annotated Word Per Hour**	300
9	**Cost Per Token**	1.333333
10	**Actual Cost of a PoS Annotated Corpus**	
10.1	**One Time Annotation Cost (One time annotation by one annotator)**	957393.9
10.2	**Annotator Cost (Validated)**	2393485

Table 4: Example Cost Calculation for a PoS Annotated Text Corpus

3.3 Cost of Creating Chunk Labelled Text Corpora

Chunking is the task of assigning labels to a group of words identified for particular purpose as a single unit. The purposes for such a chunking or grouping task can vary depending upon the suitability of the task.

This task is commonly understood to be done for linguistic phrase marking following the general phrase structure grammar of a language. However, this can also be done for various other purposes such as named entity recognition, semantic concept labeling and so on.

Each of the chunk labeling task has its own guidelines which the annotators have to diligently follow and mark it accordingly.

Chunk labeling task is usually a second step where the text to be chunked are already identified and most of the times, are already PoS annotated. PoS information may help the annotators to identify the right groups/chunks. However, this task can be done even without the PoS information.

3.3.1 Factors affecting the cost of Chunk Labelling Task

Below we discuss the factors affecting the chunk labelling task

3.3.2 Type of Chunk Labels Used

There are no standard chunk labels specified by any agency. It depends on the annotating agency what types of labels are being used. As discussed above, the number of labels may differ depending upon the types being used. For example a single grammatical Noun Phrase may have more than one label if the chunk labelling is done concept wise. The cost may differ if more than one such labels are asked in the guideline to be annotated.

3.3.3 Annotator Cost

The cost of an annotator would be almost the same or a bit higher depending upon various factors. Chunk labelling task is a bit more complex than that of the PoS tagging. It may also require multi-discipline knowledge if concept labelling is involved. For example if health/clinical domain corpus is annotated, one would like difference between different types of clinical entities such as names of diseases, procedures, medicine names and attached values, person names and so on. Therefore, the cost of annotator may decided by the agency who is doing the annotation. For the purposes of this document, we are considering that the task is that of a grammatical chunk labelling and a linguist having enough knowledge in the syntax of the concerned language would be fine to go with.

3.3.4 Source Data Preparation Cost

Chunking task is usually done after the texts are selected, cleaned and PoS tagged. Therefore, the costs of creating the cleaned raw corpus and the PoS annotation over it may also be optionally included for the customers/buyers who do not already have purchased those text.

However, this cost may be dropped for the customers who already have purchased the corresponding raw text and PoS annotated corpus.

3.3.5 Annotation and Validation Procedure

Just like the PoS annotation procedure, the standard procedure to validate the chunk data is also the same. That is, the chunking is usually done by two annotators which are then compared and arbitrated by a third annotator for any discrepancies.

The effect of this procedure on costing would be the same as is seen with the PoS annotation i.e. 2.5 times of the cost of the total chunk annotation cost by one annotator one time.

3.3.6 Formula to Calculate the Cost of a Chunk Labelled Text Corpus

Based on the discussions made above, the formula for arriving at the cost of a chunk labeled corpus be delineated as shown in the table below:

SL	Parameters	Value	Remarks
1	**Resource Type**	Text	
2	**Corpus Type**	Chunk Labelled	
3	**Language**	XX	
4	**Unit**	Chunk Label	
5	**Number of Characters (Raw Text)**	XX	
6	**Number of Words (Raw Text)**	XX	
7	**Number of Chunk Labels**	XX	
8	**Annotator Cost Per hour**	400	
9	**Annotated Chunk Labels Per Hour**	125	This may vary language wise.
10	**Cost per Chunk Label**	3.2	
11	**Actual Cost of a Corpus**	-	
11.1	**One Time Annotation Cost (One time annotation by one annotator)**	Number of token*cost per chunk label	
11.2	**Annotator Cost (Validated)**	One Time Annotation Cost * 2.5	
11.3	**With Cost of the raw text**	Annotation Cost + cost of the raw text	
11.4	**With cost of the raw text and PoS Annotation**	Annotation Cost + cost of the raw text + Cost of creating the PoS annotation	

Table 5: Formula to Calculate the Cost of a Chunk Labelled Corpus

Example cost of a chunk labeled corpus:

SL	Parameters	Values
1	**Resource Type**	**Text**
2	**Corpus Type**	Chunk Labelled
3	**Language**	Tamil
4	**Unit**	Chunk Label
5	**Number of Characters (Raw Text)**	1871372
6	**Number of Words/Pos Tagged Tokens (As per Raw Text)**	194606
7	**Number of Chunk Labels**	160386
8	**Annotator Cost Per hour**	400
9	**Annotated Word Per Hour**	125
10	**Cost per Chunk Label**	3.2
11	**Actual Cost of a Corpus**	
11.1	**One Time Annotation Cost (One time annotation by one annotator)**	513235.2
11.2	**Annotator Cost (Validated)**	1283088
13.3	**With Cost of the raw text**	1414084.04
13.4	**With cost of the raw text and PoS Annotation**	1667071.84

Table 6: Example Calculation of the Cost of a Chunk Labelled Corpus

3.4 COST OF CREATING DEPENDENCY/SYNTACTIC PARSED TEXT CORPUS (TREEBANK)

The task of parsing is done in two ways. It may be called either dependency or syntactic parsing. Although, there are mechanisms using which the dependency labels may be converted to syntactic parsed sentences, it may not always work fine or the process of working on both the ways of typing may be different affecting the costing factors. Therefore, both the tasks are considered different and separate costing formula may be devised for both of the types of the corpus.

3.4.1 Syntactic Parsing

Syntactic parsing is one of the most complex task in linguistic analysis. This is done at the level of sentence and involves interlinking of different phrase level tags to show the inter-connection of them to a mother root level which is the initial point of a sentence.

This task usually requires advanced training in linguistics, especially advanced syntax as well as the knowledge of the language concerned.

The discipline of linguistics is still not so popular yet to have attracted such a kind of trained linguist in all the languages (or even the scheduled languages of India). Therefore, such kind of corpus are still not much available.

However, there have been efforts to create such a corpus for English (e.g. the Penn Tree Bank) or Hindi/Urdu (e.g. the IITH corpus).

3.4.1.1 Factors affecting the cost of Treebank Corpus

Below we discuss the factors affecting the cost of a treebank corpus.

3.4.1.2 Type of Guidelines Used

The guidelines for a treebanking project are always a heavy a one, going into several pages. The preparation of the guidelines itself is a huge task as it requires the experts of both linguistics as well as the language concerned.

However, it is proposed that the cost incurred on preparation of the guidelines for a treebanking project would not be included in the cost calculation.

3.4.1.3 Annotation Cost

Given that this is a complex task and requires usually a person having a high level of expertise (i.e. MA, MPhil/PhD with some kind of specialized experience/exposure to syntax in general and syntax of the language concerned in particular), the cost of finding such a qualified person is usually higher than the ones required for the tasks of PoS tagging or chunking.

Therefore, the hourly rate given for this task would be kept a bit higher than the PoS tagging task. The hourly rates for an annotator may vary upon various factors. However, for the purposes of this document, we have estimated that the hourly cost may be fixed at Rs. 500 per hour.

3.4.1.3.1 Estimates of Hourly Output

As syntactic labeling project has been undertaken only for a select few languages and at only a few places on a trial basis, coming up with an estimate for the same may be a bit difficult. Due to this, coming with the estimates on the output generated by an annotator is difficult.

However, the author of this document had undertaken the parsing task for the clinical domain in his previous role as a researcher with a firm. So, the estimates from that project will be taken as a base point here. This was the task of syntactic parsing undertaken for the clinical domain at the ezDI Labs, Ahmedabad. This was an English text and we largely followed the parsing guidelines as used for the Penn Treebank project with some modifications here and there as suitable for the clinical domain.

With our project on this, it was estimated that an annotator would parse a total of 100 sentences in four hours where the average size of the sentence stood at 10 words. The sentences in the clinical domain text is usually a smaller one and may not reflect the actual size in other domains. Therefore, it is proposed that the costing unit should be done at the level of the second level tags attached to the trees.

A normal file of 50 sentences in this domain contains a total of around 500 syntax level tags. Time spent on doing this is around 4 hours. So, we can say that it takes around 4 hours to give a total of 500 syntactic labels or 125 labels per hour.

In any annotation task, usually it is the label that should be the unit to calculate the cost. Therefore, the the cost per unit for doing the syntactic annotation would be calculated as such, which can be formulated as follows:

Cost of assigning one syntactic label= Cost per hour of an Annotator / Total labels annotated

Considering that the per hour rate of an annotator doing syntactic labeling is Rs. 500/-, the cost of assigning one label would be Rs. 4 (500/125=4).

Any syntactic labeled corpus would have to give information as to how many secondary and higher level tags it contains. This would be noted as syntax level tags

3.4.1.4 Source Data Preparation Cost

Akin to the chunking task, the parsing task is done over after the other layers of annotation are done on the raw text. In other words, this task is a higher step towards the language processing and the precursor steps prior to this task are as follows:

Raw Text Preparation -> PoS Annotation -> Chunking -> Parsing

As there are costs involved in the pre-cursor tasks, these may also be included in the costing plan if the purchasers do not already have purchased those pre-cursor files.

3.4.1.5 Annotation and Validation Procedure

The procedure for annotation and validation in parsing is also the same as followed for PoS tagging and chunking. Therefore, the cost would calculated in the same manner if the annotation is validated.

3.4.1.6 Formula to Calculate the Cost of a Syntactically Parsed Text Corpus

With the discussions as noted above, a formula for such a corpus can be given as delineated in the table below:

SL	Parameters	Value	Remarks
1	**Resource Type**	Text	
2	**Corpus Type**	Syntactic Labelled	
3	**Language**	XX	
4	**Unit**	Chunk Label	
5	**Number of Characters (Raw Text)**	XX	
6	**Number of Words/Pos Tagged Tokens (As per Raw Text)**	XX	
7	**Number of Chunk Labels**	XX	
8	**Number of Syntax Level Tags**	XX	
8	**Annotator Cost Per hour**	500	

9	**Annotated Syntax Level Tags Per Hour**	125	This may vary language wise.
10	**Cost per Syntax Level Tags**	4	
11	**Actual Cost of a Corpus**	-	
11.1	**One Time Annotation Cost (One time annotation by one annotator)**	Number of token* Syntax Level Tags	
11.2	**Annotator Cost (Validated)**	One Time Annotation Cost * 2.5	
11.3	**With Cost of the raw text**	Annotation Cost + cost of the raw text	
11.4	**With cost of the raw text and PoS Annotation**	Annotation Cost + cost of the raw text + Cost of creating the PoS annotation	

Table 7: Formula to Calculate the Cost of a Syntax Labelled Corpus

3.4.2 Dependency Labelling

As noted above, dependency labeling should be taken up separately as the method of annotating is different in this process even though the pre-processing tasks are almost similar.

In India, the dependency labeling task has been done mostly at IIIT Hyderabad. The feedback taken from IIIT Hyderabad has been taken as the base point for the calculations given in the formula given below.

The unit taken for has been taken as sentence with an average number of sentences reported to be produced being 20 sentences per day per annotator. This includes data preparation, annotation, error correction and avalidation. This is a lumpsum estimate but has been considered as such for now. If better calculation mechanisms are arrived at later, it will be included and this section may be updated here.

3.4.2.1 Formula to arrive at the cost of a Dependency Labelled Corpus

SL	Parameters	Value
1	**Resource Type**	Text
2	**Corpus Type**	Dependency Labelled
3	**Language**	XX
4	**Unit**	Sentence
5	**Total Number of Dependency Labelled Sentences**	XX
6	**Annotator Cost Per hour**	500
7	**Annotated Sentences Per Hour**	2.5
8	**Cost Per Dependency Labelled Sentence**	Annotator Cost Per hour / Annotated Sentences Per Hour
8	**Actual Cost of a Corpus**	Total Number of Dependency Labelled Sentences * Cost Per Dependency Labelled Sentence

9	**Maintenance Cost**	10% of Actual Cost
10	**Overall Cost**	Actual Cost + Maintenance Cost

Table 8: Formula to arrive at the cost of a Dependency Labelled Corpus

Example Cost of a dependency labeled corpus

SL	**Parameters**	**Value**
1	**Resource Type**	Text
2	**Corpus Type**	Dependency Labelled
3	**Language**	Hindi
4	**Unit**	Sentence
5	**Total Number of Dependency Labelled Sentences**	50000
6	**Annotator Cost Per hour**	500
7	**Annotated Sentences Per Hour**	2.5
8	**Cost Per Dependency Labelled Sentence**	200
8	**Actual Cost of a Corpus**	10000000
9	**Maintenance Cost**	1000000
10	**Overall Cost**	11000000

Table 9: Example Cost of a dependency labeled corpus

4 SPEECH CORPROA

Speech processing has made long strides in the past decade so much so that fairly good quality speech command based tools are made available to common people in several languages on hand held devices. However, majority of such tools for other languages, including that of major Indian languages, are not working well.

Speech based tools usually work with a machine learning approach which a data intensive mechanism and depends on good quality data to be provided to these models to ensure good output is generated.

However, speech recognition or speech generation are not the only two types of speech resources that a researcher might be interested in. There might be other types of speech resources which may act as additional or special resources of interest to people working on the speech of various languages.

For example, the Linguistic Data Consortium for Indian Languages (LDCIL) has been working on developing different types of speech resource set which may be of interest to different set of people. A brief of these sets are given below:

1. Command and Control Words:

   This is a list of words and phrases used as commands in the language concerned. Commands like “come here”, “Bring me a glass” etc. are part of such list. This list is prepared and the speakers are asked to render it.

2. Created Text

   This is a fictional story created and the speakers are asked to render it in their speech. A set of 5 to 10 such text are created and recorded for different sets of people.

3. Form and Function Word

   A list of function class words are selected and recorded from different speakers.

4. News

   News items in the language concerned are selected randomly from different sources and recorded.

5. Phonetically Balanced

   This is a list of words carefully selected such that all types of phones and phonemes of the language are covered in it at all the positions as envisaged in the language concerned. The list is usually prepared in consultation with a phonologist and an expert in the language.

6. Place Name

   A list of common global and national place names are selected and users asked to record it. The list also includes some place names local/native to the language concerned.

7. Person Names

A list of common global and national person names as well as that of common native/local person names are included in this list.

8. Sentence

A list of representative sentences is prepared such that it covers both the phonology and different types of syntactic structures of the language concerned.

9. Most Frequent Words

The most frequent word list is generated out of a sampled raw corpus and recorded.

10. Date Format

There are various ways of speaking a date and time expression. This small set requires the speakers to speak the dates and time in the common style they would speak it.

4.1 FACTORS AFFECTING THE COST OF CREATING A SPEECH CORPORA

There might be different factors affecting the cost incurred on creating a speech corpus. The major factors towards this are noted below.

4.1.1 Text Preparation Time (for unique data set)

For the data that are guided, a text is first prepared before being recorded. This usually called a pre-recording process where the data collectors do the preparation before the speech is recorded from the informant.

This process would not be applied if the data is recorded impromptu from natural conversations (such as telephonic conversations or other discussions, speeches done for the purpose of showing/telecasting/webcasting etc.

As done for the text data, this is part of the guideline preparations. And therefore, the cost incurred in this procedure is not included here.

However, it is advisable that the cost of it may be included at the discretion of the competent person handling the resource, if so required. Text preparation may of different kind and some of those texts e.g. phonetically balanced set requires an additional effort and careful designing that can be done only by an expert or by undergoing to other corpus analysis methods. As this may have additional cost, a cost for this may be included into the formulae for such a language resource.

4.1.2 Cost of Field Visits

In the current scenario when the web/internet has reached almost every nook and corner of the world, it may be feasible that some kind of data may be collected without going into the actual field and meeting the actual people for such tasks.

But the traditional way to collect the speech data is by visiting the field where the speakers reside and collect the data or invite the informants to a place where the speech varieties could be recorded.

The table below gives an overview of the estimated costs of a field visit for Kannada.

SL	Particulars	Cost Per day	Remarks
1	Rate for Field Linguists Collecting the Data (INR 400 per hour for 8 hours a day)	3200	
2	Local travel allowance for one linguist (as per current Govt. of India norms)	225	As per GoI Norms
3	Daily Allowance for the linguist	300	As per GoI Norms
4	Hourly Honorarium for an informant from whom the data is sought (500 per informant * 5 informant for a day)	2500	As per GoI Norms
5	Daily output (in minutes) in field by a linguist recording: 15 minutes *5 speakers	75 minutes	
6	Cost per day:	6225	
7	Overhead Charges (Rs. 25 per speaker)	125	As per GoI Norms
8	Overall Cost per day	6350	
9	Cost per minute Collecting Data (Overall cost per day / overall speech recorded per day) (i.e. 6350/75	85	

Table 10: Cost of Data Collection (Field Work Method)

The above table does not give the cost of travelling expenses from the head quarters to the field or the place of the informant or the cost of an informant coming from different places to the place where the recording takes place.

This has purposely been left out so that as this cost may vary depending upon the geographical location of either the linguist or the informant. However, this cost may be included as a separate costing factor.

We are also not including the cost of equipments required for recording.

4.1.3 Segmentation and Warehousing

After the data has been recorded, the process of transferring the data on to computers, segmenting and storing them takes place. This is the first process which is done in-house after the data have been collected in the field work. Each linguist is required to segment the speech as per the respective text and create the corresponding metadata before it is stored at a designated place. This process usually takes around a month or 21 working days for a 5 day field visit or a total of 75*5=225 minutes of data.

This calls for an additional cost of 21 days for a linguist at the rate of INR 400 per hour which equals to 400*8*21=67200.

Thus the cost of segmentation and ware housing would be calculated as follows:

Cost of Segmentation/Warehousing (per minute = time spent on segmentation/warehousing divided by cost of the annotator for the corresponding time

The same may be stated in the tabular form as follows:

SL	Particular	Values	Example Value (for Kannada)
1	Total Minutes of Recorded Speech Data collected in the field	Xx minutes	225 minutes
2	Total time spent on segmentation and warehousing	Xx hours	168 hours
3	Total cost of the annotator for the time spent on segmentation and warehousing	INR 400/hour	67200
4	Per Minute cost of segmentation and warehousing of the raw data	Total Minutes of Recorded Speech Data/Total cost the linguist (i.e. 1/3)	67200/225 = 299

Table 11: Cost of Segmentation and Warehousing (for Field Data)

4.1.4 Infrastructural Costs

Infrastructural costs such as provision of equipments such as recorder, computers and other peripherals may also be included. However, this is not being considered for the time being on this and may be calculated as per the policy.

4.1.5 Administrative Expenses

There might also be some administrative costs with regard to management of such a task. This again is not being included for this document. May be included later as per the policy.

4.2 Cost of Creating Raw Speech Corpus

Based on the discussions as above where we have specified the unit cost for collecting the speech data as well as that of segmenting and warehousing it, the formula to draw the cost of the raw speech corpus could be drawn as shown in the table below:

SL	Parameters	Value	Remarks
1	**Resource Type**	Speech	
2	**Corpus Type**	Raw Corpus	
3	**Language**	Kannada	
4	**Unit**	Minute	
5	**Number of given words**	xx	
6	**Linguist Cost per hour**	400	May vary depending upon other factors.
7	**Total Minutes of Recorded Speech Data**	4800	
8	**Cost per minute Collecting Data**	85	
9	**Per Minute cost of segmentation and warehousing of the raw data**	299	
10	**Total cost of Speech Data Collection**	Cost per minute Collecting Data * Total Minutes of Recorded Speech Data	
11	**Total cost of Segmentation and Warehousing**	Per Minute cost of segmentation and warehousing of the raw data * Total Minutes of Recorded Speech Data	
12	**Total Cost of Raw Speech Data for Release**	Total cost of Speech Data Collection + Total cost of Segmentation and Warehousing	

Table 12: Formula to Calculate the Cost of a Raw Speech Corpus

As an example, the cost of the Kannada speech corpus can be drawn as shown in the table below:

SL	Parameters	Value
1	**Resource Type**	Speech
2	**Corpus Type**	Raw Corpus
3	**Language**	Kannada
4	**Unit**	Minute
5	**Number of given words**	xx
6	**Linguist Cost per hour**	400
7	**Total Minutes of Recorded Speech Data**	4800
8	**Cost per minute Collecting Data**	85

9	**Per Minute cost of segmentation and warehousing of the raw data**	299
10	**Total cost of Speech Data Collection**	408000
11	**Total cost of Segmentation and Warehousing**	1435200
12	**Total Cost of Raw Speech Data for Release**	1843200

Table 13: Example Calculation of the Cost of a Kannada Raw Speech Corpus

4.3 COST OF CREATING SENTENCE SEGMENTED SPEECH CORPUS

The raw corpus recorded are further segmented to match with the sentence. So the sentences are cut into pieces or sentence boundaries are marked using an annotation tool for each of the audio tracks. This task is called sentence level annotation.

At LDCIL, this task is done using the actual text given to the speakers in the script of the language. Thus this annotation is also done using the same script.

The cost of sentence level annotation is calculated based on the feedback received from various resource persons across languages. This is again calculated as per their output per hour. Based on their feedback, it is seen that a linguist can cover on an average annotation of 2.5 minutes of speech data for sentence level annotation.

Thus, the formula to arrive at the cost of sentence level annotation could be drawn as shown in the table below:

SL	Parameters	Value	Remarks
1	**Resource Type**	Speech	
2	**Corpus Type**	Sentence Delimited	
3	**Language**	Kannada	
4	**Unit**	Minute	
5	**Number of given words**	xx	
6	**Linguist Cost per hour**	400	May vary depending upon other factors.
7	**Total Minutes of Speech Data**	4800	
8	**Average Hourly Output for Sentence Annotation (in decimal Minutes)**	2.5	
9	**Per Minute cost of sentence annotation**	Linguist Cost Per hour / Average Hourly Output for Sentence Annotation	
10	**Total Cost of Sentence Annotation**	Total Minutes of Speech Data * Per Minute cost of sentence annotation	

Table 14: Formula to Calculate the Cost of a Sentence Annotated Speech Corpus

To given an example of the Kannada speech as per the above formula, the cost would come out as follows:

SL	Parameters	Value
1	**Resource Type**	Speech
2	**Corpus Type**	Sentence Delimited
3	**Language**	Kannada
4	**Unit**	Minute
5	**Number of given words**	Xx
6	**Linguist Cost per hour**	400
7	**Total Minutes of Speech Data**	4800
8	**Average Hourly Output for Sentence Annotation (in decimal Minutes)**	2.5
9	**Per Minute cost of sentence annotation**	160
10	**Total Cost of Sentence Annotation**	768000

Table 15: Example Calculation of the Cost of a Kannada Sentence Annotated Speech Corpus

4.4 Cost of Creating Word Segmented Speech Corpus

While raw speech corpus and sentence annotated speech have their own goals, the word level annotation and mapping of the speech corpus has its own high end uses both in ASR and TTS as well in several other kinds of research for both academic and commercial uses.

In this task the speech files are mapped with the corresponding words with boundaries given using speech annotation tools such as Praat.

Cost of creating the word level annotation again may vary depending upon the language being considered. However, for the purpose of this task, an average drawn from the word level annotation tasks undertaken for the languages of India at the LDCIL project. The average productivity for this is 35 seconds per hour or 0.6 seconds in decimal points (required for calculations).

Based on the above calculations, the formula to calculate the speech corpus annotated at the word level would be as shown in the table below:

SL	Parameters	Value	Remarks
1	**Resource Type**	Speech	
2	**Corpus Type**	Sentence Delimited	
3	**Language**	Kannada	
4	**Unit**	Minute	
5	**Number of given words**	Xx	
6	**Linguist Cost per hour**	400	May vary depending upon other factors.
7	**Total Minutes of Speech Data**	4800	

8	**Average Hourly Output for Sentence Annotation**	0.6	
9	**Per Minute cost of sentence annotation**	Linguist Cost Per hour / Average Hourly Output for Sentence Annotation	
10	**Total Cost of Sentence Annotation**	Total Minutes of Speech Data * Per Minute cost of sentence annotation	

Table 16: Formula to Calculate the Cost of a Word Annotated Speech Corpus

To show an example of the Kannada corpus, the following may be seen:

SL	Parameters	Value
1	**Resource Type**	Speech
2	**Corpus Type**	Sentence Delimited
3	**Language**	Kannada
4	**Unit**	Minute
5	**Number of given words**	xx
6	**Linguist Cost per hour**	400
7	**Total Minutes of Speech Data**	4800
8	**Average Hourly Output for Sentence Annotation**	0.6
9	**Per Minute cost of sentence annotation**	666.6666667
10	**Total Cost of Sentence Annotation**	3200000

Table 17: Example Calculation of the Cost of a Kannada Word Annotated Speech Corpus

5 Parallel Corpora

Parallel Corpora is also a type fo text corpora where the same content is available in at least two languages. The parallel corpus is often translated and aligned sentence wise. The standard goal for a parallel corpora is work as the training and development data for machine translation works between language pairs.

Parallel corpora either can be generated using existing translation memories or created afresh for a particular purpose. The method shown here describes the process of creating such a corpus afresh and not through a translation memory or some other means such as crawling.

Though LDCIL has not directly worked on this type of language resource, we do plan to develop such corpora. The author of the document has also worked on the ILCI corpora earlier in his capacity as Senior Linguist with the ILCI project funded by the DIT. Therefore, his experiences of developing such a parallel corpora is also utilized here.

5.1 FACTORS AFFECTING THE PARALLEL CORPORA

The following are the factors affecting the creation of a parallel corpora.

5.1.1 Raw Corpus Creation

The raw corpus, if it is a typed corpus, is collected in the same way as noted above. It is usually seen that the source text remains the same across languages (as is the case with ILCI) while the same text is translated across languages. In such a situation, the raw corpus charges will apply only to the language wherein the source text has been prepared.

The method of calculating the cost of the raw source text would be the same as noted above.

5.1.2 Content Copyright Costs

The source text of such languages are usually extracted and typed from published books and other content where the necessary copyright rules apply. In such a situation, it is the task of the sourcing agency to ensure that the necessary copyrights have been achieved for the purposes as intended for such resources.

Getting copyrights may also involve some costs, both in terms of manpower and royalties being paid to the copyright holders. While efforts should be made such that end users of such resources do not have to be worry about copyrights irrespective of the type of use intended by them (research or commercial), there may be some text where the copyright holders may demand royalties on further distribution. The text demanding further share of royalties should be avoided as it would be difficult to calculate and track such royalty generation and too cumbersome to manage.

5.1.3 Metadata Preparation

The source text often comes with metadata containing information such as domain and sub-domain type, author, publisher, source etc. Prof. Girish Nath Jha, the consortium leader for the ILCI project suggested that the cost of metadata preparation may also be included in such a corpus.

The matter has been considered and included here. This is to further note that metadata preparation is done only in the source and not the target languages for which the metadata remains same with a few additional column denoting the information about the translators and validators working on the translating the source text.

This is also to note that metadata preparation task is part of the source text selection. Source text selection is usually done by one or two linguists working for the source language. It is suggested that the metadata preparation cost be fixed at 30% of the total cost of source data preparation. However, this proportion may vary depending upon the various factors such as the total number of words in the corpus and the proportional number of words for each of the domains and sub-domains. Therefore, it is preferable that this cost may not be fixed by any any external agency/personnel and rather to be decided by the agency developing the corpus as the agency is finally responsible for the success of the language resource being distributed. In the case of the corpora of ILCI, the size of the corpus and domains/sub-domains are too many and as per suggestion of the consortia leader, this may be fixed for ILCI only at 30%.

This is to note that the metadata preparation is done for the text resources of LDCIL as well. The text resources of ILCI is rather much larger than the ILCI and it too has information on domain and sub-domains. However, the cost of metadata preparation is not included separately for LDCIL and is rather included in the maintenance cost of 10% over the overall cost. This is a valid justification because the size of each of sub-domains are much larger.

5.1.4 Translation Cost

The translation cost will apply only to the languages/text that have been translated from a source. Normal, standard translation rates would apply for the same as per the source text.

The existing translation cost as per the standards of National Translation Mission, a scheme of the Government of India functioning within CIIL, is Re. 1 per word for bulk orders. Howevver, this rate is quite old (coming up in 2012). Inflation has risen a lot since then and a revision is required for this.

Therefore, it is suggested that the current rate for translation across languages may be fixed at Re. 1.5 per source word.

5.1.5 Maintenance Cost

The maintenance cost includes any updates and upkeep of the language resource. As per the suggestions from various stakeholders (DIT, IIT-BHU etc.), it has been suggested that the maintenance cost be kept as nil for the resources being maintained by Government funded organizations.

5.2 FORMULA TO CALCULATE THE COST OF A PARALLEL CORPUS

Based on the factors above, a formula to arrive at the cost of a parallel corpus may be drawn as shown in the table below:

SL	Particulars	Value
1	Resource Type	Text
2	Corpus Type	Parallel Corpus
3	Language Pair	Hindi-Marathi
4	Word Count (Source Text)	
5	Character Count (Source Text)	
6	Word Count (Target Text)	
7	Character Count (Target Text)	
8	Data Source (for Source Text)	Typed
9	Data Source (for Target Text)	Translated
10	Unit (for Source Text)	Character/Key Stroke
11	Unit (for Target Text)	Source Word
12	Cost per Charcter/Key Stroke (for source Text)	0.007

13	Proofing Cost (for source Text)	0.004
14	Translation Cost / word (for target Text)	1.5
15	Actual Cost of a corpus:	
15.1	Typed Corpus: (number of characters*Data Entry cost per character)	
15.2	Typed+Cleaned Corpus: (number of characters*Data Entry cost per character)+number of characters*proofing cost per character	
15.3	Translation Cost (Source Text Word Count * Translation Cost/word)	
15.4	Copyright Costs (if any)	
15.5	Metadata Preparation Cost (30% typed+Cleaned Corpus)	
15.6	Maintenance Cost	
15.7	Total Cost: (Cost of Typed + Cleaned Corpus)+(Translation Cost) + Copyright Costs + Metadata Preparation Cost)	
15.8	Overall Cost of a corpus: Actual Cost + 10% of actual Actual Cost	

Table 18: Formula to Calculate the Cost of a Parallel Corpus

To show an example of the cost of a parallel corpus, the table below can be seen.

SL	Particulars	Value
1	Resource Type	Text
2	Corpus Type	Parallel Corpus
3	Language Pair	Hindi-Marathi
4	Word Count (Source Text)	419420
5	Character Count (Source Text)	2516520
6	Word Count (Target Text)	421000
7	Character Count (Target Text)	2526000
8	Data Source (for Source Text)	Typed
9	Data Source (for Target Text)	Translated
10	Unit (for Source Text)	Character/Key Stroke
11	Unit (for Target Text)	Source Word
12	Cost per Charcter/Key Stroke (for source Text)	0.007
13	Proofing Cost (for source Text)	0.004
14	Translation Cost / word (for target Text)	1.5
15	Actual Cost of a corpus:	
15.1	Typed Corpus: (number of characters*Data Entry cost per character)	176156.4
15.2	Typed+Cleaned Corpus: (number of characters*Data Entry cost per character)+number of characters*proofing cost per character	186222.48
15.3	Translation Cost (Source Text Word Count * Translation Cost/word)	629130
15.4	Copyright Costs (if any)	0
15.5	Metadata Preparation Cost (30% typed+Cleaned Corpus)	55866.744
15.6	Maintenance Cost	
15.7	Total Cost: (Cost of Typed + Cleaned Corpus)+(Translation Cost) + Copyright Costs + Metadata Preparation Cost)	871219.224
15.8	Overall Cost of a corpus: Actual Cost + 10% of actual Actual Cost	958341.1464

Table 19: Example Calculation of the Cost of a Parallel Corpus

6 IMAGE CORPORA

By image corpora we mean the text written on paper or otherwise that which can be converted into text in the respective scripts. Image corpora may of two types: scanned images of printed materials and the scanned images of handwritten text. While the first one is important in digitization process of old text/printed text, the other one is used mainly to recognized untyped handwritten manuscripts or capture the text while writing on an touch enabled electronic device.

6.1 SCANNED IMAGES

Scanned image of books or printed materials are already available in good quantity. However, to make the same thing available in electronic format, it needs to be digitized. This has been happening either by typing manually in the respective scripts or running an Optical Character Recognition (OCR) software over the images that can extract the text and digitize it automatically.

While OCRs with pretty fine accuracies are commercially available in English and some other major languages globally, the same is still lacking in Indian languages and demands focused attention.

6.1.1 Factors Affecting the Cost of the Scanned Image Corpus

Developing a scanned image corpus requires various steps that are listed below.

6.1.1.1 Data Source

The data is usually taken from printed materials. This may be a copyrighted text in which case the cost of copyrighting will also be included. The cost of copyrights is an amount that may be free or come at a price. This should be left to the agency developing the corpus to see whether a cost can be included for this process or not.

6.1.1.2 Scanning Costs

Scanning of the printed materials may be done using either high end machines or normal desktop scanners or even AIO printing-scanning systems. The cost and output will depend on these factors.

Scanning is a task that may be done by normal Data Entry Operators who may require initial training on how to do the scanning. The rates per hour for a Data Entry Operator is decided to be at Rs. 150 per hour (this rounded off as per the current Ministry of Labour rates which stands at Rs. 617 for a day).

For the estimates provided here, it is understood that the scanning is done on normal scanning machines with a page feed of 20 pages per batch. A page stands for a normal A4 or A5 page size which may have 300 to 600 words printed on it. It is also understood that the page fed to the scanning machines require some pre-feeding exercise such as arranging and sorting the pages and some post-feeding exercise such as saving, renaming and labelling the scanned images.

As per the feedback received from some agencies who have developed OCR corpus, it understood that scanning is done usually at a rate of 20 pages per hour which includes all of the pre-scanning and post-scanning tasks. At the end of the task, the scanned images are archived with proper labels for further processing.

Thus, the cost of scanning one page is same dividing the the total cost for one hour divided by total number of pages done in one hour. As per the estimates given above, this comes to 150/20 which is Rs 7.5 per scanned image.

The above prices are indicators only and may change if the agency doing the scan have different parameters.

6.1.1.3 Data Inputting Cost

The next process in developing a scanned image corpus is to get the scanned images typed using a proper style as required by the OCR training engines. The typing is again done by a Data Entry Operator.

The cost of typing will not be discussed here as the same has been done in the previous section and the same thing will apply here.

6.1.1.4 Vetting/Proof Reading of the Text Data

All typing tasks require proofing. This again will not be discussed here and the costs are taken from the discussions made above.

6.1.1.5 Storage and Management Cost

Roughly a 10% of the overall cost in typing and proofing would be earmarked as the storage and management cost to be levied on top of the other costs.

6.1.2 Formula to Arrive at the Cost of a Scanned Image Corpus

Based on the discussions above, the formula to arrive at the cost of a scanned image corpus would be as follows:

SL	Particulars	Value
1	Resource Type	Image
2	Corpus Type	Image Corpus
3	Language	Malayalam
4	Number of Page Images	XXX
5	Scanning Costs (per page)	XXX
6	Word Count	XXX
7	Character Count	XXX
8	Cost per Charcter/Key Stroke (for source Text)	0.007
9	Proofing Cost (for source Text)	0.004
10	Unit	Character/Key Stroke
11	Actual Cost of a corpus:	
11.1	Scanning Cost	(Scanning Cost per page * Number of Page Images)
11.2	Data Input Cost	(number of characters*Data Entry cost per character)+number of characters*proofing cost per character
11.3	Copyright Costs (if any)	As per actuals.
11.4	Maintenance Cost	10% of Actual Cost
11.5	Total Cost	Scanning Cost + Data Input Cost + Copyright Costs
12	Overall Cost	Overall Cost

Table 20: Formula to Arrive at the Cost of a Scanned Image Corpus

Example Cost of Scanned Image Corpus

SL	Particulars	Value
1	Resource Type	Image
2	Corpus Type	Image Corpus
3	Language	Malayalam
4	Number of Page Images	500
5	Scanning Costs (per page)	7.5
6	Word Count	200000
7	Character Count	1800000
8	Cost per Charcter/Key Stroke (for source Text)	0.007
9	Proofing Cost (for source Text)	0.004
10	Unit	Character/Key Stroke
11	Actual Cost of a corpus:	
11.1	Scanning Cost	3750
11.2	Data Input Cost	19800
11.3	Copyright Costs (if any)	0
11.4	Maintenance Cost	10% of Actual Cost
11.5	Total Cost	23550

12	Overall Cost	25905

Table 21: Example Cost of a Scanned Image Corpus

7 HANDWRITING RECOGNITION CORPUS

Development of Handwriting Recognition Corpus requires handwritten inputs on a touch device. This may be done by giving a predefined set of text to a number of informants who consent to write on the electronic device which are captured electronically in the desired format (a sequence of points with their x, y coordinates, and optionally, the pressure exerted by the writer).

7.1.1 Factors Affecting the Cost of the Handwriting Corpus

This corpus also as a set of processes and its own set of costs for it. These are noted below.

7.1.1.1 Data selection and Preparation

The data for handwriting is usually collected from different types of informants based on different demographic criteria such as age, gender, region, education level etc. The selection may either be guided or impromptu.

The data that needs to be written by each of the informants are prepared before the same is presented before the informants while writing. This task is done by the project implementing agency. We estimate that this cost would be a total of 10% of the other costs involved in this.

7.1.1.2 Ink Collection Cost (Capturing the Handwriting samples)

As noted above, people are invited to come to a particular place or the data collector may go to the informant’s convenient place to give the handwriting samples. This is usually a half and hour to one hour task performed by the informant on an electronic device. For this, the informants are given the incentives in cash or kind.

It is usually understood that an informant is given a sum of Rs. 500/- for this task (either in cash or kind). In turn the informants usually write a total of around 500 words. Considering that a sample collector gets handwriting samples from a total 10 informants everyday, the total collection be 5000 words per day for each data collector working full time on such a project.

Considering that data collector is a linguist with a minimum qualification of Masters, it would require a payment of Rs. 400 / hour and give an output of about 5000 words per day. Given that the linguist works 8 hours a day (excluding the time spent on other activities), the total daily cost would be Rs. 3200. Thus the cost of collecting the sample per word would be 3200/5000 i.e. Rs. 0.64 per word.

7.1.1.3 Annotation of the Handwritten Corpus

The collected handwriting data are further subjected to annotation for further processing. Annotation needs to be done at the stroke, symbol (stroke group), akshara and the word level, and the data is preferably stored as an xml file. Annotation is done at a rate of 250 words per day (as per the feedback taken from CDAC, Pune and vetted roughly the same by Prof. AG Ramakrishnan, IISC, Bangalore and Kalika Bali, Microsoft Research India, Bangalore). The annotation is understood to be done by a linguist (or someone who is trained to do this) being paid at a rate of Rs. 400 per hour or 3200 per day. Thus the annotation costs comes to be Rs. 3200/250 i.e. Rs. 12.8 per word.

7.1.1.4 Validation of the Annotation

Given that the annotation is done by a human annotator, it is further subjected to validation. The estimated productivity of validation task for handwriting recognition is 1250 words per day. This task again is done by another linguist. Thus the total cost per word for validation would come at Rs. 3200/1250 i.e. Rs. 2.56.

7.1.2 Formula to Arrive at the Cost of a Handwriting Recognition Corpus

SL	Particulars	Value
1	Resource Type	Handwriting
2	Corpus Type	Handwriting Corpus
3	Language	Malayalam
4	Word Count	XXX
5	Ink Collection Unit Cost (Per word)	0.064
6	Annotation Unit Cost (Per word)	12.8
7	Validation Unit Cost (Per word)	2.56
8	Ink Collection Cost	Ink Collection Unit Cost * Word Count
9	Annotation Cost	Annotation Unit Cost * Word Count
10	Validation Cost	Validation Unit Cost * Word Count
11	Total Cost	(Ink Collection Cost + Annotation Cost + Validation Cost

11.1	Data Selection and Preparation Cost	10% of Total Cost
11.2	Maintenance Cost	10% of Total Cost
12	Overall Cost	Total Cost + Data Selection and Preparation Cost + Maintenance Cost

Table 22: Formula to Arrive at the Cost of a Handwriting Data Corpus

Example calculation of raw corpus:

1	Resource Type	Handwriting
2	Corpus Type	Handwriting Corpus
3	Language	Malayalam
4	Word Count	457823
5	Ink Collection Unit Cost (Per word)	0.064
6	Annotation Unit Cost (Per word)	12.8
7	Validation Unit Cost (Per word)	2.56
8	Ink Collection Cost	29300.67
9	Annotation Cost	5860134
10	Validation Cost	1172027
11	Total Cost	7061462
11.1	Data Selection and Preparation Cost	706146.2
11.2	Maintenance Cost	70614.62
12	Overall Cost	7838223

Table 23: Example Cost Calculation of a Handwriting Data Corpus

8 ONTOLOGIES

Ontologies are a compilation of conceptual words which are interlinked with each other. In India, the ontology works in Indian languages have been developed with the help of TDIL and other participating institutions, led by IIT Bombay. The project was named as Wordnet. Wordnet is a term derived from the project WordNet at Princeton University whereby a thesauri like structure is made for all the content class words of a language and semantic relations are established amongst them. These semantic relations are like synonymy, antonymy, meronymy, hypernomy, hyponymy and so on. Each word is grouped into a synset correspond to a distinct synset. These synsets are then interlinked to other synsets based on the semantic relations.

Though the term WordNet is a proprietary term developed first for the language of English, the use of term wordnet has become common to the development of similar resource developed for Indian languages.

The wordnet project of TDIL, headed by IIT Bombay, has developed several wordnets for Indian languages which are at different stages. For the price calculations for developing such a resource, the actual expenditure on such a project have been taken into account.

The following table, as prepared by Dr. Aadil Kak, the PI looking after the Kashmiri wordnet project, has been taken as a template. Other wordnets may have a different pricing depending upon the efforts placed in.

8.1 Factors Affecting the Cost of a Ontology Development

Ontology/Wordnet development is a complex task whrein language experts are required from the beginning of the task. It requires identification of the concepts and use of some tools to encode the concepts and their semantic relations with other words in the wordnet.

8.1.1 Compilation Task

Wordnet development task is like a thesauri development task and requires constant updates from various experts drawn from different fields. Though basic wordnet can be developed through some experts focused attention over a period of time, it may require additional inputs from different other experts from different fields. Additionally, the compilation also require additional resources such as corpora, other dictionaries and so on. Depending upon the availability of such resources in other languages, the cost may vary. Therefore, it is difficult to gauge the total expense incurred on developing such a resource.

It is proposed that the expense for it may be drawn based on the total expense done in developing such a task.

We will take the example of TDIL funded wordnet projects and calculate the expenses based on it. The unit taken for this would be the synset.

8.2 Formula to Arrive at the Cost of an Ontology

Based on the discussions above, the following may be the formula to arrive at the cost of an ontology.

SL	Particulars	Values	Remarks
1	Resource Name	XXX	
2	Resource Type	Ontology	
3	Language	Kashmiri	
4	Data Source	Manually Compiled	
5	Total Number of Synset Units	36,685	
6	Budget Allocated for the language	36,29,000	
7	Cost per Synset as on December 2010	Budget Allocated / Total Number of Synsets Developed	
8	Actual Cost	Cost per synset * Total Number of Synsets	
9	Any Inflationary Impact	56% of the Actual Cost	
10	Current Costs	Actual Cost + Inflationary Impact	
11	Maintenance Costs	15% of Current Cost	As it is a software, the maintenance cost is kept a bit higher than the other resources.
12	Overall Cost	Current Costs + Maintenance Cost	
13	Taxes, if any	e.g. GST, if applicable (on overall Cost	

Table 24: Formula to Arrive at the Cost of an Ontology/Wordnet Resource

Example cost of an ontological resource

1	Resource Name	Kashmiri WordNet
2	Resource Type	Ontology
3	Language	Kashmiri
4	Data Source	Manually Compiled
5	Total Number of Synset Units	36,685
6	Budget Allocated for the language	36,29,000
7	Cost per Synset as on December 2010	98.92326564
8	Actual Cost	3629000
9	Any Inflationary Impact	2032240
10	Current Costs	5661240
11	Maintenance Costs	849186
12	Overall Cost	6510426
13	Taxes, if any	e.g. GST, if applicable (on overall Cost

Table 25: Example Cost of an Ontology/Wordnet Resource

9 Cost of Creating Anaphora and Antecedent Annotated Text Corpora for Anaphora Resolution

Anaphora resolution is the task of assigning the antecedent for an anaphor. Anaphors can be anything that requires back reference which includes pronouns, reflexives, one anaphor etc.

This task is done for identifying the reference of a pronoun or any other grammatical constituent which has a reference in the previous sentence or clause. This can be utilized by various natural language processing systems which require knowledge of coherence between sentences in a text.

Anaphor-Antecedent labeling has guidelines of its own which the annotators can utilize for tagging the anaphor-antecedent pair. This annotation is done on raw texts and does not require any other grammatical information to identify the correct pair.

9.1 Factors affecting the cost of Anaphora–Antecedent annotation Task

Below we discuss the factors affecting the Anaphora- antecedent annotation task

9.2 Type of Anaphor-Antecedent Labels Used

There is no standard anaphora – antecedent annotation guidelines specified by any agency. It depends on the annotating agency what types of labels are being used.

9.3 Annotator Cost

The cost of an annotator would be higher as it involves knowledge about syntax and semantic theories. It is a complex task and requires knowledge from outside world for interpreting and identifying the antecedent of an anaphor. It may also require multi-discipline knowledge if the domain of the corpus to be annotated is varying, for example if the domain of the corpus is medical document or biology related document. For the purpose of this document we consider the general domain corpora which can be carried out by a trained linguist with deep understanding of syntax and semantics.

9.4 Source Data Preparation Cost

The data/text needs to be cleaned for annotating anaphor-antecedent pair. Therefore, the costs of creating the cleaned raw corpus may also be optionally included for the customers/buyers who do not already have purchased those text.

However, this cost may be dropped for the customers who already have purchased the corresponding raw text.

9.5 ANNOTATION AND VALIDATION PROCEDURE

Validation of Anaphor –antecedent pair is done in the standard form of using inter annotator agreement. That is, the annotation is done by two annotators which are then compared and automatically verified using Kappa score. If the Kappa score is more than 0.75, the corpus is considered to have valid annotation.

9.6 FORMULA TO CALCULATE THE COST OF A ANAPHOR-ANTECEDENT ANNOTATED TEXT CORPUS

Based on the discussions made above, the formula for arriving at the cost of a Anaphor-antecedent labeled corpus be delineated as shown in the table below:

SL	Parameters	Value	Remarks
1	**Resource Type**	Text	
2	**Corpus Type**	Anaphor –antecedent annotated	
3	**Language**	XX	
4	**Unit**	Anaphor-antecedent pair	
5	**Number of Characters (Raw Text)**	XX	
6	**Number of Words (Raw Text)**	XX	
7	**Number of Anaphor-antecedent pairs**	XX	
8	**Annotator Cost Per hour**	500	
9	**Annotated Anaphor-Antecedent pairs Per Hour**	10	This may vary language wise.
10	**Cost per Anaphor-Antecedent pair**	50	
11	**Actual Cost of a Corpus**	-	
11.1	**One Time Annotation Cost (One time annotation by one annotator)**	Number of token*cost per chunk label	
11.2	**Annotator Cost (Validated)**	One Time Annotation Cost * 1.25	

Table 26: Formula to Calculate the Cost of a Anaphor-Antecedent Annotated Corpus

Example Calculation for a Tamil Anaphor-Antecedent Annotated Text Corpus

SL	Parameter	Value
1	**Resource Type**	Text
2	**Corpus Type**	Anaphor-Antecedent Annotated
3	**Language**	Tamil
4	**Unit**	Anaphor-Antecedent pair
5	**Total Number of Tokens/Words**	1,00,000
6	**Total number of Anaphor-Antecedent pairs**	4000
7	**Annotator Cost Per hour**	500
8	**Annotated Anaphor-Antecedent pair per Hour**	10

9	**Cost Per Pair**	50
10	**Actual Cost of a Anaphor-Antecedent Pair Annotated Corpus**	4,50, 000
10.1	**One Time Annotation Cost (One time annotation by one annotator)**	2,00,000
10.2	**Annotator Cost (Validated)**	2,50,000

Table 27: Example Cost of a Anaphor-Antecedent Annotated Corpus

10 Cost of Creating Named Entity Annotated Text Corpora for Named Entity Resolution

Named Entity resolution is the task of automatically assigning labels/tags to words which is a real world object which hold a proper name such as location, personal name, date, time and currency and so on.

This task is done for identifying the named entities in a sentence. This can be utilized by various natural language processing systems such as Information Extraction and other systems which require the knowledge of what the word stands for in real world.

Named Entity labeling has guidelines of its own which the annotators can utilize for tagging the entities. This annotation is done on raw texts and does not require any other grammatical information to identify the correct pair.

10.1Factors affecting the cost of Named Entity annotation Task

Below we discuss the factors affecting the Named Entity annotation task

10.2Type of Named Entity Labels Used

There are no standard Named Entity annotation guidelines specified by any agency. It depends on the annotating agency what types of labels are being used.

10.3Annotator Cost

The cost of an annotator would be higher as it involves knowledge about syntax and semantic theories. It is a complex task and requires knowledge from outside world for interpreting and identifying the named entity. It may also require multi-discipline knowledge if the domain of the corpus to be annotated is varying, for example if the domain of the corpus is medical document or biology related document. For the purpose of this document we consider the general domain corpora which can be carried out by a trained linguist with deep understanding of syntax and semantics.

10.4SOURCE DATA PREPARATION COST

The data/text needs to be cleaned for annotating Named Entity. Therefore, the costs of creating the cleaned raw corpus may also be optionally included for the customers/buyers who do not already have purchased those texts.

However, this cost may be dropped for the customers who already have purchased the corresponding raw text.

10.5ANNOTATION AND VALIDATION PROCEDURE

Validation of Named Entity is done in the standard form of using inter annotator agreement. That is, the annotation is done by two annotators which are then compared and automatically verified using Kappa score. If the Kappa score is more than 0.75, the corpus is considered to have valid annotation.

10.6FORMULA TO CALCULATE THE COST OF A NAMED ENTITY ANNOTATED TEXT CORPUS

Based on the discussions made above, the formula for arriving at the cost of a Named Entities labeled corpus be delineated as shown in the table below:

SL	Parameters	Value	Remarks
1	**Resource Type**	Text	
2	**Corpus Type**	Named Entity	
3	**Language**	XX	
4	**Unit**	Named Entities	
5	**Number of Characters (Raw Text)**	XX	
6	**Number of Words (Raw Text)**	XX	
7	**Number of Named Entities**	XX	
8	**Annotator Cost Per hour**	500	
9	**Annotated Named Entities Per Hour**	25	This may vary language wise.
10	**Cost per Named Entities**	20	
11	**Actual Cost of a Corpus**	-	
11.1	**One Time Annotation Cost (One time annotation by one annotator)**	Number of token*cost per chunk label	
11.2	**Annotator Cost (Validated)**	One Time Annotation Cost * 1.25	

Table 28: Formula to Calculate the Cost of a Named Entity Annotated Corpus

Example Calculation for a Tamil Named Entity Annotated Text Corpus

SL	Parameter	Value
1	**Resource Type**	Text

2	**Corpus Type**	Named Entities
3	**Language**	Tamil
4	**Unit**	Named Entity
5	**Total Number of Tokens/Words**	1,00,000
6	**Total number of Named Entities**	30,000
7	**Annotator Cost Per hour**	500
8	**Annotated Named Entities per Hour**	25
9	**Cost Per entities**	20
10	**Actual Cost of a Named Entities Annotated Corpus**	13,50,000
10.1	**One Time Annotation Cost (One time annotation by one annotator)**	6,00,000
10.2	**Annotator Cost (Validated)**	7,50,000

Table 29: Cost of a Named Entity Annotated Corpus

11 PRONUNCIATION LEXICON DICTIONARIES

Pronunciation lexicon dictionaries are dictionaries with a valid pronunciation of the word and follows the pronunciation lexicon specifications of the W3C. The pronunciation dictionaries may have multiple fields apart from the four basic fields of lexeme, parts of speech, phonemic transcription and sound file of lexeme. Additional fields may include phonetic transcription, alternative pronunciations and so on.

Therefore, depending upon the number of fields included in a pronunciation dictionary, the cost may go up. For a basic pronunciation dictionary, the above four fields are required. Based on the assumption, the following table has been provided by Arup Saha, CDAC, Pune.

SL	Parameters	Value	Remarks
1.	**Resource Type**	Pronunciation Lexicon	
2.	**Corpus Type**	Annotated Lexicon for both ASR and TTS	
3.	**Language**	Bangla	
4.	**Dialect**	Standard Colliqual Bangla / Rarhi	
5.	**No of Fields in Annotation**	4	For eg: Parts of Speech, Phonetic transcription, Phonemic transcription, Alternative pronunciation, etc.
6.	**Unit**	No of words	
7.	**Number of Lexeme**	X	

8.	**Linguist Cost per hour**	Y(Rs. 300)	May vary depending upon other factors.
9.	**Average Hourly Output for Annotated Lexicon(if the annotation field is four)**	Z (30 no of lexeme)	May vary depending upon other factors such as spelling correction of the word, abbreviation normalization, increase in annotation field.
10.	**Cost of linguist Per Minute (M)**	Y/60	
11.	**Number of lexeme annotation per minute(N)**	Z/60	
12.	**Total Cost Lexicon(P)**	X* (M/N)	

Table 30: Formula to Calculate the Cost of Pronunciation Lexicon Dictionary

Please note that the above is derived from Bangla pronunciation lexicon works. The factors may differ for other languages. This section would be updated with greater details later when other fields are also factored in.

12 Multi Word Expressions

Multi word experessions are a group of words (having two or more words) that have a non-compositional meaning not decipherable using the literal senses of the individual constitutent words and the grammatical order that they come in. These group of words need to be marked and labeled differently for various types of tasks to get their actual meaning. For example, to have the right translation, the MWEs need to have a separate sense attached to it.

Identifying the MWE in a running text is a tedius task and has to be done mostly manually. An MWE labeled corpus may be prepared in the first place for some use cases or it may also be repurposed for a different use.

There has not been much work on the MWE identification in Indian languages. However, based on the feedback received from Prof. Pushpak Bhattacharya and Prof. Dipti Mishra Sharma who have worked on some research tasks for a few languages, it is estimated that the annotation of Multi Word Expressions per unit can cost a lumpsum of Rs. 20 per unit.

So, for now, this would considered as a benchmark but may change lateron as we get feedback from more concentrated efforts on such tasks in future.

13 Word Sense Disambiguation

Words having more than one sense require a disambiguation using the context. This is done usually with the help of an ontology (or wordnet like resource). The disambiguation is necessary to identify the right sense as identified with the help of context. WSD is needed for tasks such as machine translation and similar other tasks where a semantic component is involved.

WSD annotation again has not been taken up for most of the Indian languages so far. However, with the few efforts on small scale research works undertaken at IIT Bombay and the feedback received from them (Prof. Pushpak Bhattacharya, some of his students/associates and Prof. Dipti Mishra Sharma), it is estimated that the cost for annotating each ambiguous word element would be Rs. 25 per unit. The unit here stands for each of the sense disambiguated in the running text.

14 Other Resources

There are various other types of language resources that are included in the document given here. These include lexical lists, dictionaries, specialized or domain specific resources etc. These have not been included because there are various factors that are dependent on the languages concerned and other spatio-temporal and technical and methodical concerns. Therefore, it would not be possible to give a generic costing formula for all the types works and it is suggested that the agency developing such a resource should come up with the actual cost incurred on it while developing such resource. While calculating the costs, the basic principles of costing as propounded in this document for the resources mentioned above may be noted such that a consistency is maintained.