Online corpus of spoken Ilokano language

F Apostol 1 and A Malicdem 2

1 Mariano Marcos State University, Ilocos Norte, Philippines
2 Don Mariano Marcos Memorial State University, La Union, Philippines

1franklinrapostol@gmail.com, 2amalicdem@dmmmsu.edu.ph

Abstract. There has been a great effort in the collection of different languages in the past years all over the world, and the development of online corpus outside the country brought new possibilities in the Philippines. However, there is a limited resource for the Ilokano Language. This paper introduces the Corpus of Spoken Ilokano Language, an online repository of spoken Ilokano in the Philippines specifically in region 1. The main component of this study is spoken Ilokano. It has been specifically built for natural language processing. It shows the difference of Ilokano language as spoken by Ilokanos in the region. The database consists of 160 speakers, 40 speakers in each province of the region, each speaking about 74 statements. Spoken Ilokano language was audio recorded and transcribed. A web application has been developed making the dataset available online. The corpus was validated to provide a useful resource of data that can be used for automatic speech recognition models.

1. Introduction
With the various language across countries, different corpora have been persistently developed and have become a fertile ground for investigation. The studies of [1][2][3][4] have outlined the development of a repository of British, English, German, Italian and Hindi language in the past years. [5] also illustrated the evolution of corpus specifically for use in the telephone. The birth of the world wide web made these corpuses available over the internet. [6] says that the world wide web is not only a tool for information retrieval and exchange but also a massive repository of authentic data, “a self-renewing linguistic resource” offering “a freshness and topicality unmatched by fixed corpora.” However, with all the researches that have been done, [7] says that the use of corpus by different language service providers and language professional remains limited due to the existence of computing resources that are likely to be perceived as less demanding regarding time and effort required to obtain them.

Corpus, as defined by [8], is a collection of writings, conversations, and speeches in a particular language or languages that are to be stored, managed and analyzed in digital form. The growth of the corpus outside the country has led the birth of the Philippine language corpus [8]. PALITO is a collection of written texts of religion and literature and Filipino Sign Language focusing on the four most commonly used dialects in the Philippines namely Tagalog, Cebuano, Hiligaynon, and Ilokano. There has been a great effort in the collection of Philippine language and making it available online in the past years, however, due to the time, it takes to collect such data, [8] strongly recommend the building of a corpus of spoken Philippine language first in Cebuano, Tagalog, Hiligaynon or Ilokano.

With this, it is believed that the building of an online corpus of spoken Ilokano language, the data collection and keeping them track will leverage the quality of Philippine corpus.
Ilokano is an Austronesian language of the Philippine type spoken by about nine million people [9]. It is a member of the cordilleran language family which comprises many language of Northern Luzon Island, Philippines. It is ranked third in terms of its number of mother-tongue speakers (probably over 7.7 million native speakers or 10.1 % of the total population). Ilokano is also considered as regional lingua franca of the Filipinos residing in the Northern Luzon. A very small minority speaks it in the Mountain Province and Davao and Bukidnon in Mindanao. It is also spoken as a second language by the Ilokano people who have migrated to other parts of the Philippines and of the world. It is spoken by the Ilokanos residing in Metro Manila, Central Luzon and as far as Mindanao. It is used by Ilokanos residing in Hong Kong, Saudi Arabia, European countries and United States of America; where the largest concentration is found in California, Alaska and Hawaii. Aside from the high number of first and second language speakers of Ilokano, it has an immense volume of literature; next in number to Tagalog literature.

While it is true that Ilokano is considered as lingua franca in the region, people in different municipality speak the same word differently. The original Ilokano speaking areas include the present-day provinces of Ilocos Norte and Ilocos Sur [9]. People from these places speak the purest form of the language, however, due to the migration of the Ilokanos southward and eastward, much of northern Luzon is heavily influenced by Ilokano language and culture. The provinces of La Union and Pangasinan are dominated in most areas by Ilokano speakers, speaking the southern dialect which has minimal lexical differences from the northern one, but a significant phonological difference.

This building of an online corpus of spoken Ilokano language will be built specifically for natural language processing. It aims to provide an excellent resource of data to be used for automatic speech recognition models and therefore, hope to enrich the quality of Philippine corpus. Likewise, the results may provide possible implications for Ilokano language teaching in the Philippines and outside. For researchers and scholars who are interested in Ilokano language, the study can provide additional input to their studies thereby contributing to the continuing evolution of studies in Ilokano language. For teachers of Ilokano, this can also provide data for activities such as spoken practices and exploration in their use of Ilokano.

With the existing infrastructure as provided by the World Wide Web and the Internet that virtually connects people from various physical locations, contributing to the development of such a collection of spoken language is now a reality.

The primary objective of this study is to build an online corpus of spoken Ilokano language particularly the language used by native speakers in Region 1. Specifically, the study aimed to: (1) Determine the difference of Ilokano language as spoken by Ilokanos in Region 1, (2) Build the corpus of spoken Ilokano language by: (a) recording voices of native citizens of Ilocos Norte, Ilocos Sur, La Union and Pangasinan, (b) transcribing the audio recordings, (c) creating a database of spoken Ilokano language, (3) Develop a web application for the corpus; and (4) Determine the validity of the corpus for speech recognition.

2. Methodology
The researcher used the descriptive and developmental research methods to organize the presentation, description, and interpretation of data gathered. The study was descriptive in distinguishing the differences of Ilokano language as spoken by Ilokanos in region 1, the collection of data and in the selection of respondents. The developmental research was considered the most suitable since the study dealt with the development of Online Corpus of Ilokano Language.

2.1. Voice Recording
With the use of purposive sampling, participants were selected and limited to those: a) who were born and raised in their home provinces; b) whose first language is Ilokano; c) who were educated in their local schools, colleges, and universities; and d) who used Ilokano language in their conversation. This was to ensure that the participants represent the genuine Ilokano speakers of Ilokano language. Furthermore, the number of respondents was determined using the central limit theorem. The number
of participants was limited to 160, with 40 participants from each province, a number which was well within the recommendation of the theorem. The study adapted the age labels used in DARPA Switchboard Corpus for the classification of age groups of Ilokanos in the region. Perceived age was grouped into four main categories: Youth (16-25), Intermediate Adult (26-35), Adult (36-50), Senior (51 and above) with 5 males and 5 females in each group. Table 1 shows the distribution of respondents with equal participants from each province and age group. Equal distribution was employed to ensure objectivity and reliability in the gathered data.

Table 1. Distribution of Participants Per Province.

| Age Group | Ilocos Norte | Ilocos Sur | La Union | Pangasinan |
|-----------|--------------|------------|----------|------------|
| A1 (16-25)| 10           | 10         | 10       | 10         |
| A2 (26-35)| 10           | 10         | 10       | 10         |
| A3 (35-50)| 10           | 10         | 10       | 10         |
| A4 (51 Above) | 10     | 10         | 10       | 10         |
| Total     | 40           | 40         | 40       | 40         |

Legend: N – Population; n - sample

Data were gathered from the Ilokano-speaking provinces of Region 1. The provinces were limited to Ilocos Norte, Ilocos Sur, La Union, and Pangasinan as these are the homelands of genuine Ilokano-speaking people. Furthermore, data were limited from the recorded oral reading of a short story from BANNAWAG magazine, the umbrella organization of Ilokano writers in the Philippines and other countries. This magazine contains the purest Ilokano language needed in the study. The selected article was a folktale story entitled “Ti Ari ken ti Kawitan” written by Juan S.P. Hidalgo, Jr, issue of March 28, 2011. It contains the four types of sentences (Imperative, Declarative, Interrogative, Exclamatory). These types of sentences were used in the determination of the different pitches of Ilokanos. The article is composed of 74 sentences, 970 words, including the title of the story. The instrument that was used in data gathering for this study aimed to elicit the participants’ personal profiles and speech samples which were audio recorded. The participants were requested to read the article written in Ilokano language in an audible manner and normal voice. Each participant’s oral reading was audio-recorded. Efforts were made to take advantage of the diverse population in schools, local government units, local congregation, and some barangays. Speakers were recruited by advertising throughout the said environment. The materials used in the recording were ZOOM H1 and BY-M1. ZOOM H1 is a handy recorder for audio and video that runs in a 24 bit and 96 Khz audio, with Lo-cut filter and auto level controls for reducing noise and a X/Y microphone design that captures perfect stereo image. It also includes microphone function for external microphone. BY-M1 is a clip-on lavaliere microphone micro-cravate used for audio recorder. A 8 gigabyte of micro SDHC was used as storage of the recorded audio. The recordings were distributed in home, office, streets, rooms as recommended in LREC 2004 Workshop: Speech Corpus Production and Validation. Participants were given instruction prior to the recording. The speaker was seated in a mono block chair in one corner of a closed room. The microphone was covered with foam windscreen for filtering noises and was clipped 5 inches below the chin. Air condition, electric fans and other appliances were turned off to lessen the noise of the environment. During the recording, no one was allowed to speak aside from the speaker to avoid distraction.

2.2. Transcription of Audio Recordings

In the transcription of the audio recording, the software audacity was used. Audio clips were selected from the speakers recorded oral reading as samples for transcription and interpretation. The four types of sentences, imperative, declarative, interrogative and exclamatory were applied as criteria in selecting the samples. The researcher selected 4 sentences from the story, one for every type of sentence. The spectrum of these audio clips was plot for analysis using a mathematical algorithm known as Fast Fourier Transform (FFT). Seven points, from 1 through 7, in the spectrum were chosen for each type of sentences in the determination of the pitch pattern. Point 1 represented the beginning of the statement,
4 represented the middle and 7 represented the end of the statement. The frequency of each points were recorded. The mean of these seven points in every three statements in each sentences was computed in terms of frequency and were presented digitally through a chart and a table.

2.3. Creation of Database
With regards to the process of the creation of a databank of spoken Ilokano language, cloud computing was utilized as the online storage of the data. The data was stored in the local server with 1 Gigabyte Random Access Memory with unlimited bandwidth for the server to cater the functions of the web application that is not supported by a third party server. In the development of the Web application used in the Corpus, the Model/View/Controller (See Figure 1) design pattern was used to have a systematic approach in development. PHP, HTML, CSS were used to design the Application Program Interface.

![Model/View/Controller Architecture](image)

Figure 1. Model/View/Controller Architecture

2.4. Data Analysis
Validation was made to ensure the reliability of the data provided in such a corpus. The validation covered the voice transcription, digital representation of pitch pattern, and the analyses and interpretation of the data being presented from the selected samples of speakers. Seven validators that were expert in Ilokano language were requested to validate the corpus. Every validator indicated their grade to the description related to the speaker. They evaluated the description and interpretation of the digital representation of voice to see if it matched to the audio recording. This was done to determine whether the audio came from a genuine Ilokano speaker or not. The validation was interpreted using the 3-point Likert scale.

\[
\text{Weighted Mean} = \frac{(\text{Number of Validator}) \times (\text{Weight of 1st Column})}{\text{Total Number of Validator}} + \frac{(\text{Number of Validator}) \times (\text{Weight of 2nd Column})}{\text{Total Number of Validator}} + \frac{(\text{Number of Validator}) \times (\text{Weight of 3rd Column})}{\text{Total Number of Validator}}
\]

The scale and their equivalent that were used to interpret the weighted mean are as follows:

| Point | Range     | Descriptive Equivalent | Descriptive Interpretation |
|-------|-----------|------------------------|----------------------------|
| 1     | 1.00 – 1.66 | Poor                   | Not Valid                  |
| 2     | 1.67 – 2.33 | Good                   | Valid                      |
| 3     | 2.34 – 3.00 | Excellent              | Valid                      |

Weighted mean with average ranging from 1.67 to 3.00 was interpreted as valid which means spoken data comes from genuine Ilokano speakers, and 1.00 – 1.66 was interpreted as not valid and means
speakers were not genuine Ilokano, which specifically needs to be changed for a better corpus of Ilokano Language.

3. Results and Discussion

3.1. Difference of Ilokano Language as Spoken by Ilokanos in Region 1
Based on the use of e/ə, Ilokano may be divided into two dialects: the front medial dialect (e) and the high back unrounded (ə). [9] pointed the “Southern Dialect” to La Union and Pangasinan which use both the high back unrounded (ə) and front medial dialect (e) and the “Northern Dialect” to Ilocos Sur and Ilocos Norte which use the front medial dialect (e) only. Ilokano has six contrastive vowels (five in northern dialect) represented in orthography by five letters /a,e,i,o,u/. The orthographic symbol ‘e’ in Ilokano constitutes two separate vowel sounds in southern dialect and one sound in northern dialect. Table 2 shows the summary on the difference of Ilokano language as spoken by Ilokanos in Region 1.

| Province       | Dialectal Background | Use of e/ə in Pronunciation | Ilokano Vowels |
|----------------|----------------------|------------------------------|----------------|
| Ilocos Norte   | Northern Dialect     | front medial dialect (e)     | a,e,i,o,u      |
| Ilocos Sur     | Northern Dialect     | front medial dialect (e)     | a,e,i,o,u      |
| La Union       | Southern Dialect     | high back unrounded (ə) and front medial dialect (e) | a,e,i,o,u,ə |
| Pangasinan     | Southern Dialect     | high back unrounded (ə) and front medial dialect (e) | a,e,i,o,u,ə |

3.2. Building the corpus of spoken Ilokano language

3.2.1. Recording of Voices Respondents were requested to read a folktale story aloud while it was being recorded. The place of recording was distributed in home, schools, office, streets, and rooms as recommended in LREC 2004 Workshop: Speech Corpus Production and Validation. The dimension of the room was 4 x 4 meter, the ceiling was 4 m high. During the recording in the street and school community, the microphone was placed 2 inches below the chin. This was to ensure that the voice was captured and at the same time lessen the noise of the environment captured by the microphone. Teachers, employees of Local Government Units, Barangay Officials and students were also requested to become part of the data gathering.

3.2.2. Transcription of audio recordings Audio clips were selected from the speakers recorded oral reading as the samples for transcription and interpretation. The four types of sentences, imperative, declarative, interrogative and exclamatory were used as criteria in selecting the audio clips. These were believed to give the necessary data to be used in determining the pitch pattern of speakers in the region. The spectrum of the selected audio clip was plot for analysis. See Figure 2. Seven points, from 1 through 7, in the spectrum were chosen for each type of sentences in the determination of the pitch pattern. Point 1 represented the beginning, point 4 represented the middle and point 7 represented the end of the statement. The points were where the pitch in the spectrum fades in and fades out. The frequency of each points were recorded. The mean of these seven points in every three statements in each sentences was computed in terms of frequency and were presented digitally through a chart and a table. Three statements for each type of sentences were analyzed in the determination of the pitch pattern of speakers. Seven points in the spectrum were obtained for each type of sentences to see the pattern of pitch. The mean of the seven points in every three statements in each sentences was computed in terms of frequency.
and was presented digitally. Audio clips from all of the 40 speakers in each province were plot and analyzed to come up with a single pitch pattern of speakers in each province. The average was computed and combined into single pitch patterns. Based on the computed means of the four types of sentences, interpretations were drawn.

![Figure 2. Screenshot of Audacity Spectrum](image)

### 3.2.3. Creation of Database of Spoken Ilokano Language

This study used the cloud computing environment for the online banking of spoken Ilokano language. Databases were made for the audio, respondents, province, users, and for those who will request to become contributors. Entities were constructed for each database. Attributes were also defined. Primary keys and foreign keys were also assigned. Relationships were established between these databases. The data was stored in the local server with 1 Gigabyte Random Access Memory with unlimited bandwidth for the server. The data was stored in the local server to cater the functions of the web application that is not supported by a third party storage.

### 3.2.4. Development of Web Application

In the development process, sublime was used as the text editor and Node.js was the platform. In designing the Graphical User Interface and other functions of the web, both for the client and the admin, Angular JS was used as a framework. Yeoman was used as scaffolding of the web app. Bootstrap was used in the styling of the website. Bower and NPM was utilized to manage the frameworks, libraries, utilities and node.js packages. GruntJS was used as a task runner. The domain name used for the corpus was ‘ilokanocorpus.net’. Figure 3 shows the front page of the developed web application for the corpus.

### 3.2.5. Validity of the Corpus

To ensure the reliability, validation was made for the corpus. The validation was made by experts in Ilokano language in the region. Seven validators were requested to evaluate the corpus against specified requirements. The validation covered the voice transcription, digital representation of pitch pattern, and the analyses and interpretation of the data being presented from the selected samples. Pitch pattern was described based on the digital representation of the four types of sentences and compared as to who has the highest and lowest pitch among the speakers in the region. Every validator indicated their grade to the description related to the speaker. They evaluated the description and interpretation of the digital representation of voice to see if it matched to the audio recording. This was done to determine whether the audio came from a genuine Ilokano speaker or not.
Using the formula for weighted mean, the weighted mean in each province was computed and presented in Table 3.

![Corpus of Spoken Ilokano Front Page](image)

**Figure 3.** Screen Shot of Corpus of Spoken Ilokano Front Page

**Table 3.** Validation Report

| Speaker       | Mean | Descriptive | Equivalent | Interpretation |
|---------------|------|-------------|------------|----------------|
| Ilocos Norte  | 2.29 | Good        |            | Valid          |
| Ilocos Sur    | 2.14 | Good        |            | Valid          |
| La Union      | 2.14 | Good        |            | Valid          |
| Pangasinan    | 2.14 | Good        |            | Valid          |
| Grand Mean    | 2.17 | Good        |            | Valid          |

Referencing to the scale and their equivalent presented in the data analysis, the grand mean, 2.17, was interpreted as valid. The voice recording perfectly matches to the digital representation and interpretation of the pitch pattern of Ilokano in Region 1. Thus, the voice recordings came from a genuine Ilokano speaker. With these results, the overall quality of the data in the corpus was good and the corpus was valid and there should be no problem in using it for its intended purposes and other applications.

4. Summary, Conclusions and Recommendations

In this paper we have discussed the difference of Ilokano language as spoken by Ilokanos specifically in region 1. The use of high back unrounded (ə) and front medial dialect (e) makes a difference in the pronunciation of Ilokano language. The challenges of building the corpus has been elaborated. Some of these challenges were in terms of the selection of native Ilokano speakers, methods of data collection, and the transcription of collected data. We have also shown the process of the development of the web application. The web application was developed in a way that could help interested parties access the data. Finally, the validity of the corpus has been presented. Validators deem that the corpus was valid for its intended purposes for it showed the grand mean of 2.17.

In the future we hope more and more affiliations could participate in this work. Only in this way, can the corpus be constructed more efficiently and be used or shared more easily.

References

[1] Grefenstette G and Kilgarriff A 2003 Introduction to the special issue on the web as corpus *Assoc. for Computational Linguistics* vol 29 no 3 pp 333-47

[2] Baroni M, Bernardini S, Ferraresi A and Zanchetta E 2009 The WaCky wide web: a collection of very large linguistically processed web-crawled corpora *Language Resources and Evaluation* vol 43 no 3 pp 209-26
[3] Belhe S, Dubey M, Kulkarni M, Magdum D, Patil T and Shah R 2015 Methodology for designing and creating Hindi speech corpus 2015 Int. Conf. on Signal Processing and Communication Engineering System pp 336-9

[4] Yin Z and Li A 2017 A standardization program of speech corpus collection 20th Conf. of the Oriental Chapter of the Int. Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA) pp 1-5

[5] Zhigang Y 2016 How to make a telephone speech corpus Proc. of the 2nd Int. Conf. on Communication and Information Processing pp 174-8

[6] Mautner G 2005 Time to get wired: Using web-based corpora in critical discourse analysis Discourse & Society vol 16 no 6 pp 809-28

[7] Bernardini S and Ferraresi A 2013 Old needs, new solutions: comparable corpora for language professionals Building and Using Comparable Corpora pp 303-19

[8] Boutaba R, Cheng L and Zhang Q 2010 Cloud computing: state-of-the-art and research challenges J. of Internet Services and Applications vol 1 no 1 pp 7-18

[9] Burnard L 1995 British National Corpus: Users Reference Guide British National Corpus Version 1.0 (Computing Service, Oxford University)

[10] Dita SN, Roxas RE, Inventado P 2009 Building online corpora of Philippine languages Proc. of the 23rd Pacific Asia Conf. on Language, Information and Computation vol 2

[11] Rubino CR 2000 Ilocano Dictionary and Grammar: Ilocano-English, English-Ilocano (University of Hawaii Press)