# The Documentation of Endangered Altaic Languages and the Creation of a Digital Archive to safeguard linguistic diversity

Choi Woonho
You Hyun-Jo
Kim Juwon

# The Documentation of Endangered Altaic Languages and the Creation of a Digital Archive to safeguard linguistic diversity

**Choi Woonho**
Lecturer, Department of Linguistics, Seoul National University, Korea

**You Hyun-Jo**
Lecturer, Department of Linguistics, Seoul National University, Korea

**Kim Juwon**
Professor, Department of Linguistics, Seoul National University, Korea

## ABSTRACT

Language is a vehicle of intangible heritage and transmits many social and cultural concepts from generation to generation. Half of the world's languages, including most of the Altaic languages, are now in danger of extinction. The loss of a language means the loss of cultural and intellectual diversity. This paper describes a linguistic project which aims to preserve the endangered Altaic languages. The ASK REAL (Altaic Society of Korea, Researches on Endangered Altaic Languages) team has gathered linguistic resources from thirty-eight Altaic languages and plans to build an extensive digital archive of all fifty-five of them. Through field research in minority language communities spread over a vast area of Eurasia, we acquired nearly three thousand lexical items of multi-media data, a few hundred grammatical constructions and another few hundred examples of daily conversation for each language. The data is converted into a standard digital format and managed in a database. A small part of the collection is currently available to the public via an easily accessible web interface with multilingual annotations for international users.

## Keywords

Altaic languages, endangered languages, documentation, documentary linguistics, Manchu-Tungusic, Mongolic, Turkic, linguistic diversity

## Introduction

This paper aims to introduce the documentation and digital archiving project of the REAL (Researches on Endangered Altaic Languages) team, which was initiated by ASK (the Altaic Society of Korea) to create a collection of multi-media resources and materials relating to the Altaic languages. The documentation of endangered languages does not directly prevent their becoming extinct, but it is a fundamental step towards preserving endangered languages and their related cultural heritage.

Language is one of the most important factors in preserving culture and institutions and it is an essential element of self-identification for indigenous communities, even though it is not included as one of the five broad intangible heritage domains of UNESCO's 2003 *Convention for the Safeguarding of the Intangible Cultural Heritage*. Language plays an essential and necessary role in transmitting and maintaining intangible heritage from generation to generation, and it can be viewed as a vehicle of intangible cultural heritage (Smeets 2004: 161).

It is estimated that over six thousand languages are spoken in the world today (Krauss 1992: 5). UNESCO reports that half of them will have disappeared by the end of this century if nothing is done to protect them. Loss of a language does not simply mean the replacement of one majority language by another, it is part of *a much larger process of LOSS OF CULTURAL AND INTELLECTUAL DIVERSITY* (Hale 1992a: 1).

Altaic languages are widely spoken throughout Eurasia and many of them are in danger of dying out. The Dongxiang language, for example, is one of the Mongolic languages and is the native language of the Dongxiang people. The Dongxiang people live in the Dongxiang Autonomous County of Linxia Prefecture in Gansu Province, China. In 2000 it was reported that there were 513,805 Dongxiang people in China, and it was estimated that half of them use their native language. In February 2009, field research on this language was conducted in Lanzhou in Gansu Province, China. Two Dongxiang speakers were interviewed by the ASK REAL research team. One was then thirty-nine years old and the other was thirty. Both are female and work as teachers. When the subjects were asked to count, they gave the numbers from one to ten in Mongolic words. For numbers over ten, however, they used Chinese. The researchers found that the Dongxiang language is full of Chinese words. The words for *sun*, *moon* and *star* in Dongxiang are the same as in other Mongolic languages, but the Dongxiang for *sky* is a Persian word. The words *father*, *mother*, *elder brother*, *younger brother* and *grandson* are Chinese, but the words for *son*, *daughter*, *wife*, *father-in-law*, *mother-in-law*, and *daughter-in-law* are the same as in other Mongolic languages (Kim et al. 2011: 176-178).

The ASK team conducted field work for six years between 2003 and 2009 to collect spoken Altaic language resources and to build a digital archive for the purpose of preserving the languages, for carrying out linguistic studies and possibly to help revive endangered languages at some point in the future.

## The classification and distribution of the Altaic languages

The Altaic language family is divided into three branches: the Manchu-Tungusic, Mongolic and Turkic branches. However, there are no authoritative guidelines for identifying individual Altaic languages. SIL Ethnologue (http://www.ethnologue.com) provides sixty-five unique language codes for the Altaic language family, but it is unlikely that there really are sixty-five individual Altai languages considering that Ethnologue assigns different codes to the same language when it is spoken in different countries. ASK REAL identifies fifty-five individual languages and classifies them into three branches as follows:

**I. The Manchu-Tungusic branch**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | Ewen | 2 | Ewenki | 3 | Solon | 4 | Negidal |
| 5 | Nanai | 6 | Uilta | 7 | Ulchi | 8 | Udihe |
| 9 | Orochi | 10 | Manchu | 11 | Sibe | | |

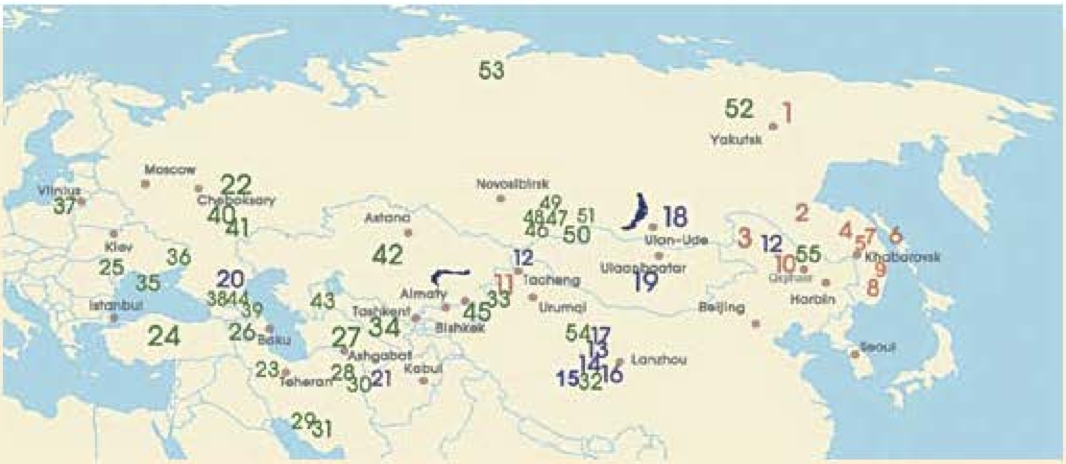


Figure 1
The distribution of Altaic languages.

II. The Mongolic branch

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 12 | Dagur | 13 | Monguor | 14 | Bonan | 15 | Kangjia |
| 16 | Dongxiang | 17 | East Yugur | 18 | Buriat | 19 | Mongolian |
| 20 | Kalmyk-Oyrat | 21 | Moghol | | | | |

III. The Turkic branch

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 22 | Chuvash | 23 | Khalaj | 24 | Turkish | 25 | Gagauz |
| 26 | Azerbaijani | 27 | Turkmen | 28 | Khorasan Turkish | 29 | Qashqa'i |
| 30 | Afshar | 31 | Aynallu | 32 | Salar | 33 | Uyghur |
| 34 | Uzbek | 35 | Crimean Tatar | 36 | Urum | 37 | Karaim |
| 38 | Karachai-Balkar | 39 | Kumyk | 40 | Tatar | 41 | Bashkir |
| 42 | Kazakh | 43 | Karakalpak | 44 | Nogai | 45 | Kirghiz |
| 46 | Altai | 47 | Khakas | 48 | Shor | 49 | Chulym Turkish |
| 50 | Tuvan | 51 | Tofa | 52 | Yakut | 53 | Dolgan |
| 54 | West Yugur | 55 | Fuyu Kirghiz | | | | |

The language numbers match the numbers on the Altaic language distribution map. The Manchu-Tungusic languages are spoken mainly in Eastern Siberia and Manchuria. For historical reasons, the Sibe language has been spoken in Xinjiang in China since 1764. The Mongolic languages are mostly spoken in Mongolia and the surrounding areas. The Kalmyk-Oirat language, however, is mainly known to the west of the Caspian Sea. This language is one of the Mongolic languages, but the name 'Kalmyk' comes from a Turkic word. It means 'remain' or 'people who remained abroad and did not come back to their home towns'. The Turkic languages are spoken widely across the Eurasian continent from Eastern Europe to Eastern Asia.

## How to collect language resources in the field

We acquired resources from thirty-eight (out of fifty-five) Altaic languages including ten (out of eleven) Manchu-Tungusic ones, eight (out of ten) Mongolic ones, and twenty (out of thirty-four) Turkic languages during our six years of field work. In the first project, from September 2003 to August 2006, nine professors, seventeen doctoral students and forty-seven graduate and undergraduate student assistants participated. Five professors, six researchers and twenty assistents took part in the second project which ran from July 2006 to June 2009.

Since most of the Altaic languages are used in Russia, China and Mongolia, most of their users can speak an 'official' language as well as their mother tongue. Field research is done through intensive interviews with native speakers who can also speak one of those official languages.

All conversations and interviews with the subjects are recorded by audio and video devices. The audio materials are digitised with high-definition sound quality. The video devices record the shape of the speakers' mouths and lips. This provides complementary information if a researcher cannot determine the speaker's pronunciation from the audio material alone.

The usual way of doing field research into language is illustrated below. The subject sits in position 1 and faces the video camera so that it records his/her face and especially the shape of the mouth. The language specialist sits to the left or right of the speaker in position 2. He monitors the quality of the audio recording while transcribing the speaker's answers in IPA (the International Phonetic Alphabet). An assistant researcher is in position 3. He/she monitors sound and video quality while recording the interview, and makes sure that the subject is looking at the video camera while answering. The interpreter sits behind the video camera in position 4. Because most subjects can speak the official language of their country as well as their native tongue, the official language is used as a meta-language in interviewing. A meta-language is a language which describes the object language or concept. So, when the interpreter asks a question from the questionnaire in the official language, the subject understands the question but answers in his/her mother tongue. In the Sakha Republic of Russia most people speak Russian. Therefore Russian was used as the meta-language when interviewing subjects there.

## The questionnaire

The questionnaire used in interviews by ASK REAL researchers is set in advance because the research is focused on investigating the current status of the languages in use and making inventories of lexical and grammatical forms of those languages. A linguistic survey can be carried out most efficiently if the questions are carefully formulated in advance; anthropological and ethnographic surveys are generally more flexible.

The ASK REAL questionnaire consists of five parts: (1) questions about lexical items (a 'lexical item' is a word or a sequence of words that acts as a unit of meaning), (2) questions about basic conversations, (3) questions about

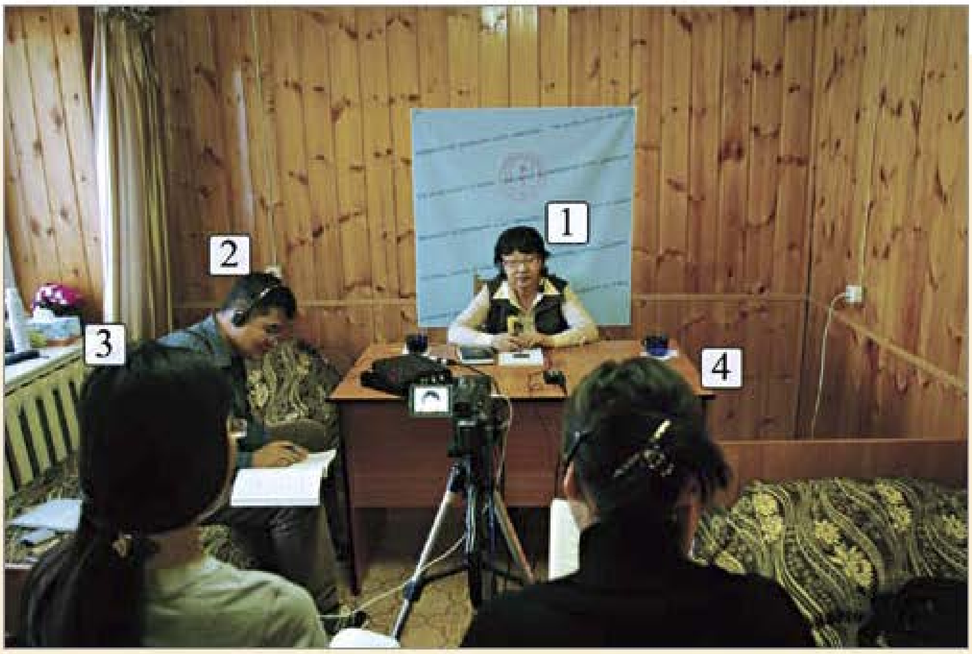


Photo 1
A scene from the field research project on the Dolgan language in Yakutsk, Sakha.

| 어 휘 | | | | | |
|---|---|---|---|---|---|
| 번호 | | 러시아어 | 한국어 | | 전사 |
| 1 | va001 | солнце | 해(태양) | 1 | |
| 2 | va006 | луна | 달 | 1 | |
| 3 | va009 | звезда | 별 | 1 | |
| 4 | va011 | небо | 하늘 | 1 | |
| 5 | va013 | свет | 빛 | 2 | |
| 6 | va014 | земля | 땅 | 1 | |
| 7 | va015 | почва | 흙 | 1 | |
| 8 | va017 | поле | 밭/들 | 2 | |
| 9 | va018 | песок | 모래 | 2 | |
| 10 | va031 | лес | 숲 | 1 | |
| 11 | va033 | пастбище | 목장 | 2 | |
| 12 | va034 | гора | 산 | 1 | |
| 13 | va047 | река | 강 | 1 | |
| 14 | va048 | речка | 시내/내 | 2 | |
| 15 | va051 | море | 바다 | 2 | |
| 16 | va052 | озеро | 호수 | 2 | |
| 17 | va055 | лёд | 얼음 | 2 | |
| 18 | va056 | вода | 물 | 1 | |
| 19 | va057 | колодец | 우물 | 2 | |
| 20 | va058 | родник, ключ | 샘 | 1 | |

**Figure 2**
shows part of a questionnaire which has Russian as the meta-language

grammar, (4) detailed information about the speaker and (5) lists of the audio and video materials which are produced during the research. The lexical part consists of single words, or groups of words, which describe concepts. The grammatical part is focused on investigating the declensions of nouns and the conjugation of verbs.

All the questions have unique identification codes for managing and processing the data. For example, the unique identification code number 'va001' is assigned to the concept 'sun' in the part of the questionnaire dealing with lexical items. Since most of the Altaic languages are used in China, Mongolia and the former Soviet Union, the unique identification code number 'va001(sun)' is annotated in Chinese, Mongolian and Russian, in other words, the questionnaire is edited and published in versions of these three languages. In addition, every unique identification code number in the questionnaire database is also annotated and has links in Korean, English and Pinyin. This multilingual (Chinese, Russian and Mongolian) annotation enables questionnaires to be generated automatically when the meta-language is specified in advance.

Figure 2 shows part of a questionnaire which has Russian as the meta-language. The first column contains serial numbers for convenience when a printed version of the questionnaire is used. The second column contains the unique identification code numbers for the concepts. In the third column Russian words are annotated for an interpreter who asks the questions of the native speaker. The fourth column contains the Korean annotation for use by the language specialist and the assistant researcher(s). The fifth column gives the weighting of the information on a scale from one to four. The subject's actual words are transcribed in the IPA at the time of the interview in the blank space in the last column.

The lexical questions consist of 2,856 concepts which are classified into twenty-four semantic fields. These semantic fields are as follows:

| | | | | | |
|---|---|---|---|---|---|
| 1 | astronomy/ geography | 2 | weather | 3 | time/period/ season |
| 4 | relationships/ occupation | 5 | politics/ economy/ culture | 6 | military/ transportation |
| 7 | human body | 8 | disease | 9 | residence/ instruments |
| 10 | clothing | 11 | food/tableware | 12 | animals/hunting |
| 13 | livestock/ breeding | 14 | birds | 15 | fish |
| 16 | insects | 17 | plants | 18 | metal/jewellery |
| 19 | direction | 20 | number/ measurement | 21 | action |
| 22 | pronouns | 23 | property/state | 24 | etc. |

The lexical items are classified into four levels according to how they are weighted (ie. their importance). Level one includes 250 or more concepts which come from basic vocabulary in everyday use. There are about 560 concepts in level two and 1,323 in level three. This level contains generic terms like *animal* and *plant* or abstract nouns like *condition/state, substance/essence, happiness, secret* and *promise*. Level four includes those lexical items which express modern concepts like *newspaper, broadcasting, politics, economy* and *culture*.

The questions about basic conversations consist of about 340 sentences which are used in seventeen everyday situations, for example, 'meeting people', 'visiting', 'hunting', 'apologising' and 'talking about the weather'.

The queries about grammar are divided into seven sections: (1) the nominal and case markers (2) verb endings (3) the derivation of words (4) the *copula* (This is a word used to link the subject and the predicate) (5) the construction of auxiliary verbs, (6) negation, interrogation and quotation, and (7) other special constructions.

To cover this number of lexical, grammatical and conversational items takes three or four days worth of interviews with a subject, interviewing for six hours a day.

## Building a digital archive

The ASK REAL digital archive was created for two purposes. Firstly, to publicise linguistic diversity and show how it relates to cultural diversity. Secondly, it aims to introduce the endangered Altaic languages and to present some of them in audio/video form.

It is often said that language reflects culture. Mongolian, for example, describes a nomadic lifestyle. We can see this from the examples of Mongolian words for livestock in the table below.

| classification | sheep | goat | cattle | horse | camel |
|---|---|---|---|---|---|
| **generic** | honi | yamaa | üher | aduu | temee |
| **stud one** | huc | uhn | buh | ajarga | buur |
| **emascu-lated male** | ireg | er yamaa | šar | mori | at |
| **male** | em honi | em yamaa | ünee | guu | ingge |
| **one year old** | hurga | išig | tugal | unaga | botgo |
| **two yearsold** | - | - | byaruu | daaga | torom |
| **cry** | mailah | mailah | mööröh | yancgah | builah |
| **dung** | horgol | horgol | argal | homool | horgol |

If the language of the nomads disappears there may be problems transmitting aspects of their particular form of cultural heritage. It is therefore obvious that languages need to be preserved so that cultural heritage can be passed down from generation to generation. For this purpose the ASK REAL team has documented many of the endangered Altaic languages and described their linguistic features (Kim et al.: 2008, Kim: 2011, Ko et al.: 2011, Li et al.: 2008, Li: 2011, Yu et al.: 2008, Yu: 2011). So that the language resources of the ASK REAL may be shared with other researchers, audio/video materials will be made available to the public in stages via the World Wide Web.

All the audio and video materials gathered in the field are digitised and archived for future research. As proposed in Bird & Simons (2003), the file formats of multi-media materials should conform to international standards. The audio materials of ASK REAL are recorded and converted into files as follows:

- file format: PCM WAV (16 bit 48 kHz)
- Mic Input Level: under -20 dB
- audio type: stereo (the voice of the consultant and other researchers are separated and recorded on separate channels)

All the video tapes are converted into DV-AVI files with a resolution of 720*480 and audio sampling rates of 48 kHz.

The ASK REAL multi-media collection is already partially available to the public on the website of the Center for Language Diversity of the Altaic Society of Korea (http://www.cld-korea.org).

The current web-based open digital archive provides access to audio-visual recordings of lexical items from four languages: East Yugur of the Mongolic branch, Chulym Turkic of the Turkic branch, Evenki of the Manchu-Tungusic and Nivkh of the Paleo-Siberian family of languages. The ASK REAL open digital archive can be accessed and retrieved as follows:

- Browse by language
- Multilingual search
- Browse by lexical category
- Browse by vocabulary level

Firstly, the languages in the archive can be browsed by their language name and all the lexical items in the chosen language can be displayed and accessed as shown in Plate 2. Secondly, the site supports multilingual searching. If you find lexical items in Altaic languages for the concept *hunting*, for example, you can find them in this archive along with the Korean, English, Russian, Chinese or Mongolian words for this concept. Plate 3 shows the screen for a multi-lingual search and what can be retrieved from the archive. Thirdly, lexical items in the archive are divided into twenty-four semantic fields and the archive can therefore be browsed by lexical category within the appropriate semantic fields as shown in Plate 4. If you follow the link this will show the results of a search on all the words in the archive in the 'astronomy/geography' category. Finally, you can browse all the words by their vocabulary level (weighting).

## Conclusion

Language itself is a cultural heritage in that it is used as a vehicle for transmitting intangible cultural heritage. Linguists are responsible for documenting and describing endangered languages (Hale 1992a: 1). Nowadays, many Altaic languages are not the official languages of their societies and this means that those languages will probably disappear within two or three generations. This paper has described some of the efforts ASK REAL has made to collect and document endangered Altaic languages. Up until now only two thirds of the Altaic languages have been studied in Korea. Changes in languages, and even their disappearance, can be viewed

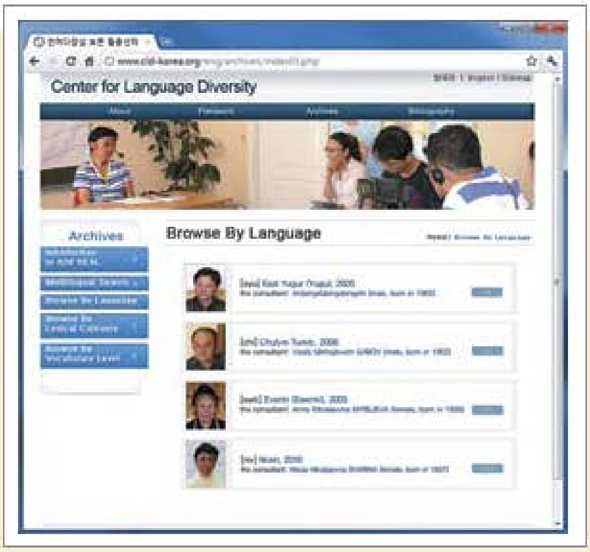


Plate 2
Center for Language Diversity – the ASK REAL digital archive.

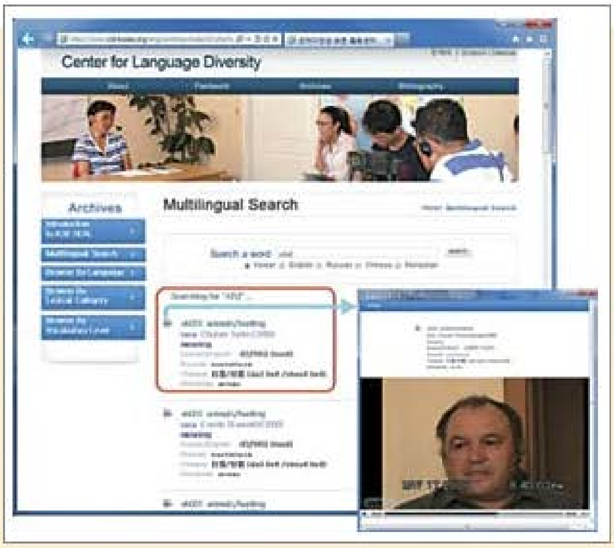


Plate 3
Multi-lingual search.

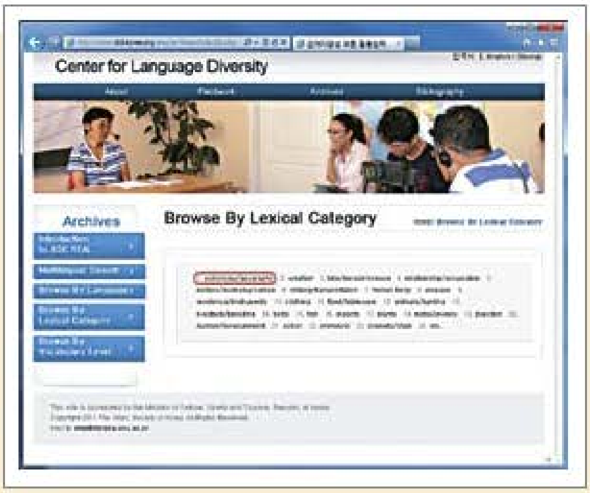


Photo 4
Browse by lexical category within semantic field.

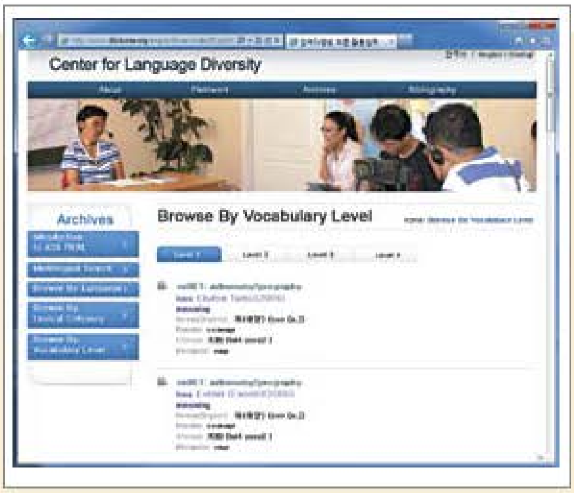


Photo 5
Browse by vocabulary level (weighting).

as part of a process of natural selection, but language diversity is important because it correlates with biological diversity. We conclude this paper with the hope that endangered languages, especially the endangered Altaic ones, may attract more attention, and stand a better chance of being studied and preserved, than they did in the past.

## ACKNOWLEDGEMENTS

- The field research and the digitisation of endangered Altaic languages was supported by KRF (currently NRF) between 2003 and 2009, and the Center for Language Diversity was supported and sponsored by the Ministry of Culture, Sports and Tourism of Korea between 2010 and 2011. We appreciate this support and sponsorship.

## REFERENCES

- Bird, Steven and Simons, Gary, 2003. 'Seven Dimensions of Portability for Language Documentation and Description' in *Language*, 79(3), pp. 557-582.
- Choi, W.H., 2011. *Analyzing Text Materials and Building Digital Archives of Altaic Languages* (in Korean: *alta'i'eon'eo tekseuteujaryo'ui bunseokgwa dijiteol aka'ibeu guchuk'ui silje*), Taehaksa.
- Hale, Ken, 1992a. 'Endangered Languages: On endangered languages and the safeguarding of diversity' in *Language*, 68(1), pp. 1-3.
- Hale, Ken, 1992b. 'Endangered Languages: Language endangerment and the human value of linguistic diversity' in *Language*, 68(1), pp. 35-41.
- Himmelmann, Nikolaus P., 1998. 'Documentary and descriptive linguistics' in *Linguistics*, 36(1), pp. 161-195.
- Kim, J.W., Ko, D.H., Boldyrev, B.V., Chaoke, D.O., Han, Y. and Piao, L., 2008. *Materials of Spoken Manchu*, Seoul National University Press.
- Kim, Juwon, 2011. *A Grammar of Ewen*, Seoul National University Press.
- Kim, J.W., Yu, W., Li, Y.S., Choi, M.J., Choi, W.H., Lee, H.Y., Cheon, S.H., Kwon, J.I., 2011. *Documentation of Altaic languages for the Maintenance of Language Diversity* (in Korean, *eon'eoda'yangseong bojon'eul wihan alta'i'eon'eo munseohwa*), Taehaksa.
- Ko, D.H. and Yurn, G.D., 2011. *A Description of Najkhin Nanai*, Seoul National University Press.
- Krauss, Michael E., 1992. 'Endangered Languages: The World's languages in crisis', in *Language*, 68(1), pp. 4-10.
- Li, Y.S., Kim, K.S., Lee, H.Y., Choi, H.Y. and Ölmez, M., 2008. *A Study of the Middle Chulym Dialect of the Chulym Language*, Seoul National University Press.
- Li, Yongsong, 2011. *A Study of Dolgan*, Seoul National University Press.
- Smeets, Rieks, 2004 'Language as a Vehicle of the Intangible Cultural Heritage' in *Museum International*, Volume 56, Issue 1-2, pp. 156-165.
- Yu, W., Kwon, J.I, Shin, Y.K., Choi, M.J., Borjigin, B., and Bold, L., 2008. *A Study of the Tacheng Dialect of the Dagur Language*, Seoul National University Press.
- Yu, Wonsoo, 2011. *A Study of the Mongol Khamnigan Spoken in Northeastern Mongolia*, Seoul National University Press.