# Japanese Effort Toward Sharing Text and Speech Corpora

**Shucihi Itahashi*+**
*National Institute of Informatics
2-1-1 Hitotsubashi, Chiyoda-ku,
Tokyo, Japan 101-8430
itabashi@nii.ac.jp

**Koiti Hasida+**
+National Institute of Advanced
Industrial Science and Technology
1-1-1 Umezono, Tsukuba, Ibaraki,
Japan 305-8568
Hasida.k@aist.go.jp

## Abstract

This report introduces the activities of the two organizations related to collection and distribution of text and speech corpora in Japan. One is the Language Resource Association (GSK) and the other is NII-Speech Resources Consortium (NII-SRC).

## 1 Introduction

Although the need for shared speech and text data has long been acknowledged, its realization has been slow to develop in Japan. The Language Resource Association (GSK) was established in 1999 in order to share and distribute text and speech corpora. It was renovated as an NPO in 2003 with an emphasis on text corpora. The National Institute of Informatics (NII) has decided to initiate the Speech Resources Consortium (SRC) in 2006, with the goal of creating future value in information media, particularly speech media.

## 2 NII-Speech Resources Consortium

The National Institute of Informatics (NII) was founded in Tokyo, Japan in April 2000 as an inter-university research institute organized to conduct comprehensive research on informatics and to develop an advanced infrastructure for disseminating scientific information. As a part of promoting the missions, NII decided in 2006 to initiate the Speech Resources Consortium (SRC) for creating future value in information media, particularly speech media. NII is promoting this consortium together with GSK (Itahashi and Ohsuga, 2006).

### 2.1 SRC objective and activities

The SRC aims at the collection, distribution, investigation, research, and standardization of electronic data and software tools that are necessary for the development of science, education, and industry concerning speech. The consortium will contribute to the development of the information society through these activities.

The SRC investigates the existing speech resources, catalogues them, and shows them on its homepage in order to promote the research and development of speech information processing. It also requests research institutions to offer their speech resources to the SRC. Based on these activities, it urges the distribution, promotion, and publicity of speech resources. The consortium will conduct the additional production and distribution of the speech resources that are frequently requested. It also conducts investigations and research on speech resources. Users will be able to obtain the speech resources or data they need and use them through simple processes offered by the SRC. SRC distributes 23 speech corpora.

### 2.2 Organization

The SRC is made up of a chairperson, a researcher, an adviser, a secretary, and a Speech Corpora Promotion Committee. The committee works to promote the development of the SRC. Around 15 members have been invited to join the committee from the fields of speech processing,

linguistics, acoustics, speech and language corpus creation, and speech and language resource provision. The committee meets a few times a year.

## 3 Language Resource Association (GSK)

The GSK was renovated as an NPO in 2003 and is qualified as a corporate body and can mediate between the producer and users of a language corpora (Hasida and Tanaka, 2006).

### 3.1 GSK objective and activities

The GSK aims at almost the same objectives as those of the NII-SRC mentioned in the preceding section, including both the text and speech corpora, but the text corpora are the main concern of GSK at present.

Prof. Y. Mikami of Nagaoka University of Technology proposed a project on the "Construction of Networks for Asian Linguistic Information Technology Resources" together with GSK. This proposal was adopted as a three-year project starting from fiscal 2005. It is supported by the Science and Technology Promotion & Coordination Fund, and Yen15 million ($125,000) is available for GSK. The mission of the project is to create a network of qualified Asian partners to specify and support the development of high priority language resources for Asian languages. As a part of this project, the GSK organized the "Asian Language Resources Workshop 2007" in March. Twenty one people from 13 countries participated in the workshop.

The GSK activities are almost the same as those of the NII-SRC. The GSK is made up of a president, two vice presidents, 11 board members, 25 steering committee members, a secretary, and two office clerks. GSK supplies two text corpora and a speech corpus; it plans to distribute seven more text corpora soon.

### 3.2 Corpora distribution system

We have three systems of corpora distribution to be conducted by the NII-SRC and GSK. (1) No-fee distribution: As a rule, the cost of handling the corpora is to fall on the user, although the corpus itself is free of charge. (2) Agency: The producers of the corpora entrust the SRC/GSK with receiving requests from the users. The SRC/GSK advertises the corpora to speech researchers through the homepage. It mediates the user's requests to the producer or provider of the corpora. (3) Fee-based distribution: Making speech corpora usually requires some money, including royalties. Some corpora cost users a certain amount of money to obtain, although they are not so expensive.

## 4 Present Issues

So far, we have established the GSK and NII-SRC for the text and speech corpora, respectively, while the LDC and ELRA distributes both the speech and text corpora.

The GSK is supported by a project until the end of this fiscal year, but it will loose its financial support at that point. The GSK has a relatively small amount of corpora to be distributed, so it is still too early to stand on its own feet. The NII-SRC activities will be supported by a project for a few more years, but it can not gain profit from distributing the corpora created by the researchers outside of NII. We will search for better ways by which both the NII-SRC and GSK will be able to act as corpora agents like the LDC or ELRA. There are some more organizations related to language resources such as ATR, NICT, and NIJL. We need much more collaboration and coordination among these.

## 5 Conclusion

This report explained the activities of the GSK and the NII–Speech Resources Consortium (NII-SRC). GSK and NII-SRC will facilitate the distribution, promotion, and publicity of the language resources, and in so doing, will contribute to the information society in Japan and in Asia. For further information, please refer to the following URLs.
http://research.nii.ac.jp/src/eng/index.html
http://www.gsk.or.jp/index_e.html

## References

S. Itahashi, T. Ohsuga. 2006. Introduction of NII-Speech resources Consortium, *Proc. Oriental COCOSDA-2006*, Penang, Malaysia: 38-43.

K. Hasida, H. Tanaka. 2006. Nonprofit Organization "Language Resource Association (GSK)", (in Japanese) *Japanese Linguistics*, 20:107-110.