Some personal reflections on (national) corpora

Kikuo Maekawa
(National Institute for Japanese Language and Linguistics, NINJAL)

1. Introduction

The theme of this forum is national corpora. The notion of language corpora is a matter of linguistics, while the notion of nation is concerned with many mutually entangled branches of humanities and social sciences including, for example, politics, sociology, history, and ethnology.

Because my specialty is limited to phonetics and linguistics, it is beyond my ability to discuss the latter aspect of the issue in any consistent and exhaustive manner. Therefore, in this talk, I will deal almost exclusively with the linguistic aspect of the theme, especially the issues of corpus design and compilation. But I will try to present some notes on the Taiwanese national corpora project from my own perspective at the end of the talk. Until then, I would like to discuss the basic issues of corpus design and compilation based upon my own experience.

But before I go into the details, please let me summarize my career in the development of Japanese language resources. This information is necessary to understand the main issues in my talk.

It was in 1999 that I was involved in a language resource development project for the first time. It was a collaborative project among the Tokyo Institute of Technology (TiTech), the National Institute of Communications Technology (NICT) and the NINJAL that aimed at the development of a prototype of the next generation automatic speech recognition (ASR) system that could recognize spontaneous monologue. This task was considered to be a very difficult goal at that time.

The project was supported by a time-bounded (five years) budget from the former Science and Technology Agency (currently it is a part of the Ministry of Education, Culture, Sports, Science and Technology –MEXT). And the mission of the NINJAL group was to develop a corpus of spontaneous speech that could be used as the learning data for the system.

The project ended in the spring of 2004, and the corpus, which is known today as the Corpus of Spontaneous Japanese (CSJ), was released for public use one month after the project ended [1]. There is a wide consensus that this five-year project brought real breakthrough in the ASR of spontaneous speech. Professor Sadaoki Furui of TiTech who supervised the whole project received various awards including the Medal with Purple Ribbon from the Japanese government. The total amount of budget used for the construction of the

CSJ was about 600 million yen.

After one-year break, I started the second project in 2006. It was the development of the Japan's first balanced corpus known as the Balanced Corpus of Contemporary Written Japanese (BCCWJ) [2]. This time the research fund was supported by the Kakenhi (Grants-in-Aid for Scientific Research) grant, which is a competition-based research fund of the Japanese Government. Unlike the CSJ whose goal was to set a new standard at the world level, the aim of the new project was to catch up the standard of the corpus linguistics in the English-speaking countries. At that time, one of the factors that impede the development of Japanese linguistics was the lack of a reliable large-scale balanced corpus of written Japanese. The goal of the new project was to solve this problem by compiling a balanced corpus whose size is comparable to the renowned British National Corpus (BNC).

A five-year priority-area Kakenhi program entitled "Japanese Corpus" started in the spring of 2006. This was a research program by the collaboration of Japanese linguists and the researchers of natural language processing. The program had two goals, namely the compilation of the BCCWJ, and the promotion thereby of the corpus-linguistic analyses of the Japanese language. It was the NINJAL researchers that played the central role in achieving the former goal. The one hundred million words BCCWJ was released for public use in December 2011. The total amount of the budget used in this project was 780 million yen.

After the completion of the BCCWJ project, I launched the project of NINJAL Web Japanese Corpus (NWJC), which is a 24.8 billion words web corpus compiled to complement the low word-coverage rate of the BCCWJ. This was a five-year project (2011-2015) supported directly by the MEXT, and the NWJC has been publicly available on the Web since 2016. Although I was the leader of the project superficially, the project was conducted almost exclusively by a young colleague [3].

Currently, there are several ongoing corpus compilation projects in NINJAL, but I am not directly responsible for any of them. I have already handed the torch over to the next generation. As a result, I'm in a position where I can talk about the research and development of language resources from a more objective point of view. From here, starts the central part of this talk.

2. Seven basic requirements of modern language corpora

According to the online Cambridge dictionary of English, the word "corpus" is defined as "a collection of written or spoken material stored on a computer and used to find out how language is used." This is a good dictionary definition, but it needs to be augmented in view of today's corpus design. In one of my writings, I pointed out seven basic requirements of modern corpus design, viz., the representativeness, balance, size, authenticity, machine-readability, public availability, and annotation [4]. Below, I will touch upon each of these.

2.1 Representativeness

Representativeness and balance are mutually related notions. Some linguists treat them as near synonyms, but I think it is important to make the distinction. In my view, representativeness is a quasi-mathematical notion, while balance is subjective. A corpus is regarded to be representative of the target language if users can obtain unbiased information about the whole body of the language just by analyzing the corpus, which is a small, usually a very small, subset of the target.

One well-known mathematical way of assuring corpus representativeness is the random sampling. But in order to apply this technique, we must have information, especially that of the size, about the statistical population from which samples are to be drawn. Once such information becomes available, we can design a corpus to which various techniques of statistics can be applied. Unfortunately, however, it is not always easy to design such corpora. Especially in the case of spoken corpora, it is usually impossible to define the statistical population in a meaningful manner. I will return to this issue later in this talk.

2.2 Balance

As mentioned in the previous section, balance is a subjective notion. In a plain term, a corpus is regarded to be a balanced corpus, if the corpus covers many different registers of the target language. Note, however, that there is no established way of knowing the total number of registers in a given language; this is why I regard this notion subjective. Note also that the word "register" is used here to refer to the variation of language in relation to "the situation of use of the variety" [5].

It is important to recognize that the classification and choice of registers in the design of a "balanced" corpus depends deeply on cultures. For example, in the renowned Brown Corpus (world's first "balanced" corpus, one million words of written American English), 17 texts out of the total of 500 texts (3.5% of the corpus, namely) were devoted for the category of "religion". This choice appears to be very strange for most Japanese people, because contemporary Japan is a highly postreligious society.

A direct consequence of this culture-dependency is that the compilation of so-called comparable corpora is possible only among languages used in the societies that share the same or very similar cultures. Comparable corpus is a collection of similar texts in different languages or different varieties of a language. Oslo Bergen Corpus (British English), Kolhapur Corpus (Indian English), Wellington Corpus (New Zealand English), and Macquaries Corpus (Australian English) are the example of Brown-compatible comparable corpora. These all belong to Christian society of West or its colonies.

The text balance of the Brown Corpus was designed on the basis of the consensus of a group of language specialists. So, were the group organized differently, the resulting design could be different.

One way to overcome such uncertainty would be to make full use of random sampling. This is what I and my colleagues of the NINAJL tried in the design of the BCCWJ. In the design of so-called library sub-corpus (30 million words) of the BCCWJ, we did not pay any attention for the text balance deliberately. Samples of this sub-corpus were chosen by extensive random sampling; all texts in a population (that is all books registered in the public libraries of Tokyo metropolis during the years 1986-2005) had the same probability of being chosen as the samples. Put differently, we designed the balance of the sub-corpus so that the extracted samples approximated maximally the text balance of the population that we cannot observe directly.

There are cases, however, where the application of random sampling is not favored. For example, if we conduct random sampling of books and magazines published in Japan or Taiwan, or any 'free' countries, the drawn samples will include definitely samples of obscenities or pornographies, and the number of such samples will not be very small. In the case of the BCCWJ's publication sub-corpus, we were surprised to know that about 10% or more of the samples from books and magazines were this sort. Our decision was to include all the obscenity samples to keep the exactness of statistical sampling, but the decision would be different depending on the goals of the corpora. If the corpus is to be used for pedagogical purposes, exclusion of the obscenity samples would be the appropriate decision.

There is also a case where samples are chosen randomly from the pre-fixed registers. Sampling of the so-called publication sub-corpus (35 million words) of the BCCWJ was this type. In this case, samples were randomly drawn from the three pre-fixed registers, namely books (28.6 million words), magazines (4.5 million), and newspapers (1.4 million). Note the differences in the sample sizes reflect the sizes of the populations.

Lastly, there are situations where sampling from the pre-fixed registers are the only choices; sampling of spoken language is the typical case. In this type of sampling, it is the application goals of the corpus that determine the sampling method. In the case of the CSJ, 95% of the samples was devoted for monologue samples because the principal aim of the corpus was the development of next-generation ASR system for spontaneous monologue; but the remaining 5% was devoted for various speech registers including dialogue and read speech for the sake of the comparison with the monologue. Recently, a new technique of sampling for everydaty conversation was devised by my colleagues. I will touch upon this topic later.

2.3 Authenticity

In corpus linguistics, it is presupposed that all samples in a corpus are authentic ones. By authenticity is meant that the samples are the records of real language behavior that occurred without the intervention of the corpus compilers. In this sense, most example sentences in the scientific articles of syntax, semantics, and pragmatics are not authentic, because they were created by the authors. It is also true with most examples cited in dictionaries. Recently, there

are many dictionaries that are said to be corpus-based, but even in these dictionaries, the examples are frequently edited by the dictionary compilers for the sake of readability.

At this point, readers might have the impression that a given sample is either authentic or non-authentic. Theoretically, it is true, but in practice, it is not always easy to make such distinction. Here again, spoken language poses many problems. For an example, the BNC is a balanced corpus covering both written and spoken registers, but the samples of the spoken part consist of the transcription texts of recorded speech; this implies that prosodic and/or paralinguistic information is completely lacking. Can this type of plain transcriptions be regarded to be authentic? The answer depends on the research purposes. If the user is interested in the lexical aspects of the target language, the transcription texts can be regarded to be authentic, but if the user is interested in the communicative aspect, they can hardly be authentic, because without the reference to prosodic and paralinguistic information, it is often impossible to identify the communicative intentions of speakers.

In this respect, it is to be noted that speech transcription cannot be perfect. Here is an example; in one academic presentation speech of the CSJ, the presenter talked about [se:ʃitsu]. In Japanese, this sequence of speech sounds is either "characteristic" (性質) or "voice-quality" (声質). In our case, we couldn't decide which was the right transcription, since the theme of the talk was about the computer processing of Japanese speech. So we asked the presenter (whom I happened to know well) his intension, but only to receive an answer telling that he was unable to decide. This is by no means an exceptional case.

Also, spontaneous speech is characterized by the presence of speech errors and various hesitation phenomena. Some corpus compilers make it a rule to correct all errors and remove all hesitations, but this can cause serious problems; for one thing, it is not always possible to know the "correct" form (like in the example above) , for another, speech errors and hesitations provide rich information about the things like mental state of the speakers, the management of dialogue, and the psycholinguistic process of speech production.

Similar problem can happen in written language as the result of mistyping or misunderstanding of the authentic phonological form of the word. Here I present an example that I encountered in the BCCWJ and NWJC.

In Japanese, Sino-Japanese word 原因("cause") is pronounced as [genin], but there is also a relaxed variant that sounds something like [ge:in]. Some native speakers wrongly believe that the latter is the authentic form of the word, so they type the word in the word processing system as "geiin" (Note the long vowel [e:] is written in hiragana as "ei"). The output from the kana-kanji conversion system is 鯨飲! I'm not sure if these letters make sense in Chinese, but in Japanese they are used as a part of a four characters idiom 鯨飲馬食 meaning "Drink like a whale and eat like a horse". You can find a lot of 鯨飲 used to refer to 原因 in Japanese blogs. In addition, I found recently some blogs where the bloggers used 鯨

飲 to refer to 原因 on purpose. Is this an authentic usage or mistype? It's a difficult question to answer.

2.4 Size

Traditionally, the size of a corpus was not the focus of serious scientific discussion; it is because the corpus size is not determined by the scientific factors but by research cost and the information technologies available at the time of construction. It does not mean, however, that the corpus size has no relevance to the scientific achievements brought about by the corpus. On the contrary, the size of a corpus is one of the most important factors that determines what can be done with the corpus. As a rough approximation, the larger the corpus size, the more complex and high-quality research is possible, given that the quality of annotations remains the same.

An example is the word-embedding (aka distributed representation), a technique that captures the meaning of words (or any other linguistic units) and their compositionality by means of a dense vector consisting of some hundreds or thousands of dimensions. This big breakthrough would not have been possible without the development of the Web and the incredible amount of text data available from the Web.

As you know well, recent technological breakthroughs in the fields of natural language processing and speech processing are largely supported by the technique of machine-learning including deep-learning, and these techniques –especially the latter–require extremely large amount of data for learning. It is not a mistake to say that the size of corpus has become more important in recent years.

The research cost required to create the same amount of data differs considerably depending on the register of samples. At the time we constructed the CSJ, the costs required for spoken data are dozens of times higher than those for written data, due mainly to the cost of manual speech transcription and prosodic labeling.

Today, the difference could be diminished because the speech transcription could be replaced partly by the ASR system. The overall performance of today's ASR system is by far better than the time of CSJ (and we are proud of our contribution to the technological breakthrough through the development of the CSJ).

But even today some phases of spoken corpora construction remain very time and cost consuming compared to written corpora. Collection of spontaneous speech sample (with the clearance of copyright issues) and the annotation of prosodic events like intonation are the typical cases.

2.5 Machine-readability

Those were the days when machine-readability was a crucial issue in corpus creation, today, there is not much to write about machine readability. Most of the issues discussed seriously

in the past like encoding technique and character sets and the cost and speed of storage are no longer serious issue. The amazing dissemination and downsizing of electronic computers as well as the establishment of the Unicode have made these issues nearly obsolete.

To my view, important issues that remains to be solved in this domain include the standardization of the format for data exchange and the format for the recording of corpus analysis process, but I will skip these issues in this talk due to the limitation of time.

2.6 Annotation

Corpus annotation means the structural information added to a corpus in view of greater ease of corpus search. It includes word-segmentation and POS information, dependency structure, phrase structure, anaphora/cataphora, segmental labeling, prosodic structure, and so forth.

Unlike the five requirements discussed so far, corpus annotation is language-dependent. As the result of this, it is not easy to make a general consideration on the issue. I will eschew detailed discussions and show some examples from Japanese that reveal the importance of annotation in corpus analysis.

In Japanese, the most important annotation is word-segmentation and POS analysis; a process known collectively as morphological analysis. As you know probably, Japanese is a so-called agglutinative language in which word boundaries are difficult to identify uniquely. It is the reflection, at least partly, of this morphological nature of the language that the Japanese writing system does not have the habit of word-segmentation by blank.

As the result of this, in Japanese, simple string search often fails to provide satisfactory results. For an example, if we search 国語 in a corpus, the result may contain the strings like 中国語, 韓国語, 外国語, 母国語, 自国語 and so forth. In the same vein, if we search リズム [rizumu] a loan word meaning “rhythm”, the result may include strings like ナショナリズム [naʃonarizumu] ”nationalism”,アルゴリズム[arugorizumu] “algorhythm”, フォーマリズム [fo:marizumu] “formalism”, プリズム [purizumu] “prism”, and much more.

Japanese corpus without the morphological annotation is difficult to analyze, and, as you can predict easily, the difficulty increases quickly as a function of the corpus size.

Japanese has one more cumbersome, still intriguing, property. That is the multiplicity of orthographic representation, meaning that in Japanese a word may often have multiple ways of being written. This property stems from the complexity of the Japanese writing system that mixes multiple characters including Kanji (Chinese characters), hiragana and katakana (Japanese syllabaries), and Roman alphabets.

The multiplicity is particularly evident in the Japanese native words (the Yamato vocabulary). For an example, according to a Japanese dictionary, the simplex verb /waku/ “grow hot” can be written either in hiragana as わく or in kanji and hiragana as 沸く. Similarly, another simplex verb /okoru/ “happen” can be written either as おこる and 起こ

る. But in this case, there is one more variant 起る which differs in the part written in hiragana (the hiragana part stands for so-called conjugation ending).

Thus, the compound verb /wakiokoru/ "burst" has the possibility of six variants. In the one hundred million words BCCWJ, all these variants can be found. Actually we find more variants in which /waku/ is written with different kanji as 湧く. Moreover, there are much more possibilities of this compound because verbs conjugate in Japanese. As I checked the BCCWJ, I found more than two hundred possibilities for this word to be written. And this is not an exceptional case.

The examples shown above suggest strongly that, generally speaking, meaningful search of corpus texts can only be done only after morphological analysis. And the analysis needs to be automated, since the manual analysis will pose intorelable limitation on the size of corpus. Today, Japanese texts can be automatically analyzed with the F-value of 0.95 or higher, as long as written texts are concerned. In the construction of the BCCWJ, we manually checked the results of automatic analysis and raised the F-value to about 0.98.

The automatic morphological analysis of the Japanese texts is a technology that reached the level of practical use in the 1990s. It was one of the greatest contributions that the natural language processing studies did to the Japanese linguistics. On the other hand, my colleagues of the NINJAL made two important contributions to the morphological analysis of the Japanese.

First, we proposed general criteria of word-segmentation in Japanese. It was necessary because at the time we entered this field, each morphological parser followed the criterion of word-segmentation of its own, and quite often the criterion was inconsistent from a linguistic point of view.

For example, a compound 国立国会図書館 "National Diet Library" can be analyzed in several different way including 国立+国会+図書+館, 国立+国会+図書館, and 国立国会図書館; the last one is the case where the whole string is recognized as a single word. When I compared the performance of several morphological parsers in 2006, I was surprised to know that all these examples appeared in the output of the parsers. Here, note that it is not the "god's truth" correctness that matters. Result of analysis can be different depending on purposes. What really matters is the consistency of analysis.

We proposed criteria for two-level morphological analysis based upon Short Unit Word (SUW) and Long Unit Word (LUW). Roughly speaking, SUW corresponds to the analysis like 国立+国会+図書+館 and LUW corresponds to 国立国会図書館 (i.e., a single compound). The criteria were not simple; they are described in detail with examples in two volume manuals of about 360 pages [6].

Second, we compiled a machine-readable dictionary for morphological analysis named UniDic. UniDic is a SUW-based dictionary in which morphological information are

hierarchically organized. Figure 1 shows the representation of an adjective 大きい "big". At the level of dictionary index, there is just one lexeme to which its reading and lexeme class (i.e. rough POS class, one of Nominal, Verbal, and Adjectival) are specified. At the second level (shown in the middle of the figure), two word-forms of the lexeme are registered, i.e., {ookii} and {okkii} (where 'kk' stands for a so-called geminate). These two forms share the same linguistic meaning but differ in their styles; the former being formal and the latter being casual. At the third level (shown in the right), one or more orthographic forms (called 書字形) and phonetic forms (called 発音形) are specified for each word-form. In the case of {waku} discussed above, at least three orthographic forms, i.e., 沸く, わく, and 湧く are to be registered.

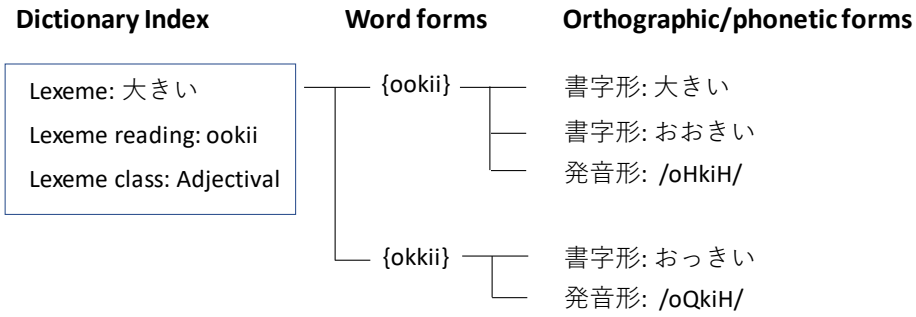


Figure 1: Hierarchical representation of lexical information in the UniDic

Currently, UniDic is available for public use free of charge in three different models: modern written Japanese model, modern spoken Japanese model, and classical Japanese model. And the last model consists of sub-models each developed for different historical period of the Japanese language like Edo and Meiji. Users can use these models (and sub-models) for morphological analyses with the MeCab parser [7].

The latest version of UniDic contains 226,319 lexemes, 367,463 word-forms, 719,298 orthographic-forms, and 367,705 phonetic forms. All models of UnDic mentioned above share these lexical items, but the models are different with respect to the contextual weightings automatically machine-learned for the Mecab parser.

In our laboratory, UniDic is managed by means of a relational database (RDB). Figure 2 is a snapshot of an RDB application that we call UniDicExplorer, with which we can search the information in the UniDic [8]. This figure shows the case of lexeme No 24199 一寸("a little bit"). This item has eleven word-forms, thirty-six orthographic forms, and eleven phonetic forms.

Corpora analyzed using UniDic (hence the segmentation unit being the SUW) are mutually comparable as long as the morphological information is concerned. Put differently,

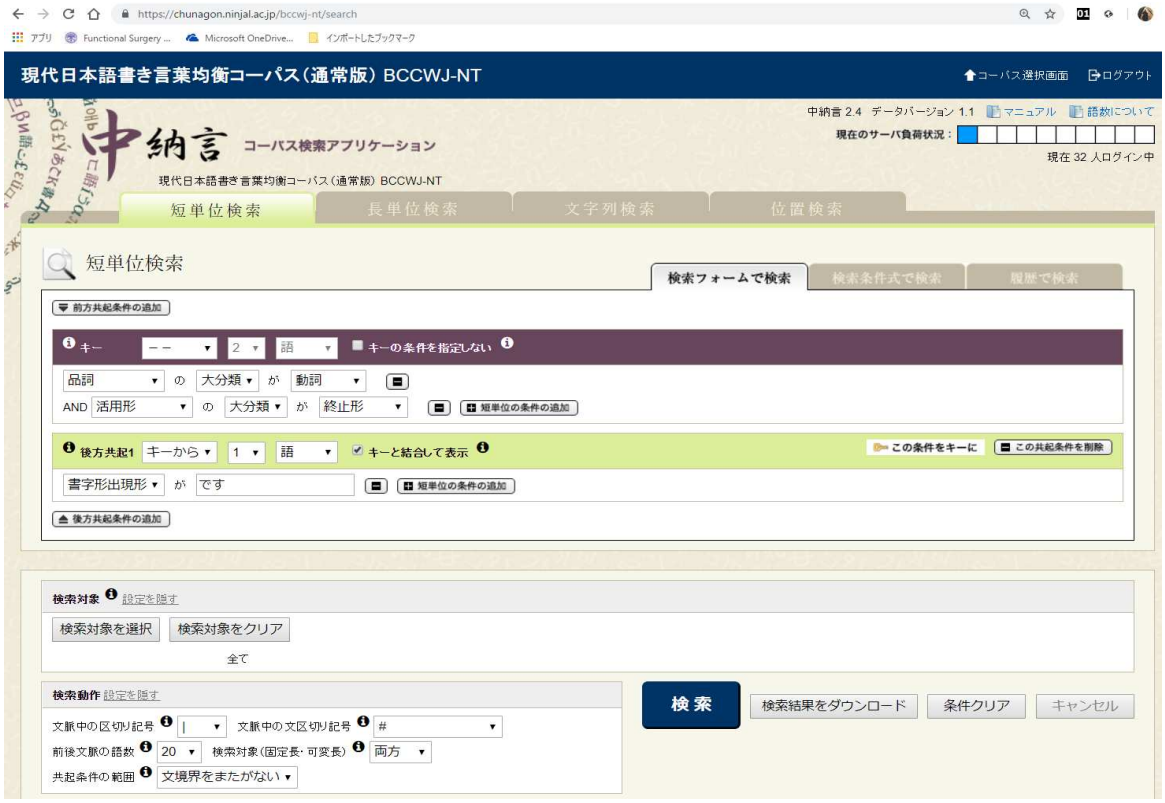
morphological comparison of corpora each analyzed with different segmentation criteria are very much prone to nonsensical results. This problem is known as corpus interoperability.

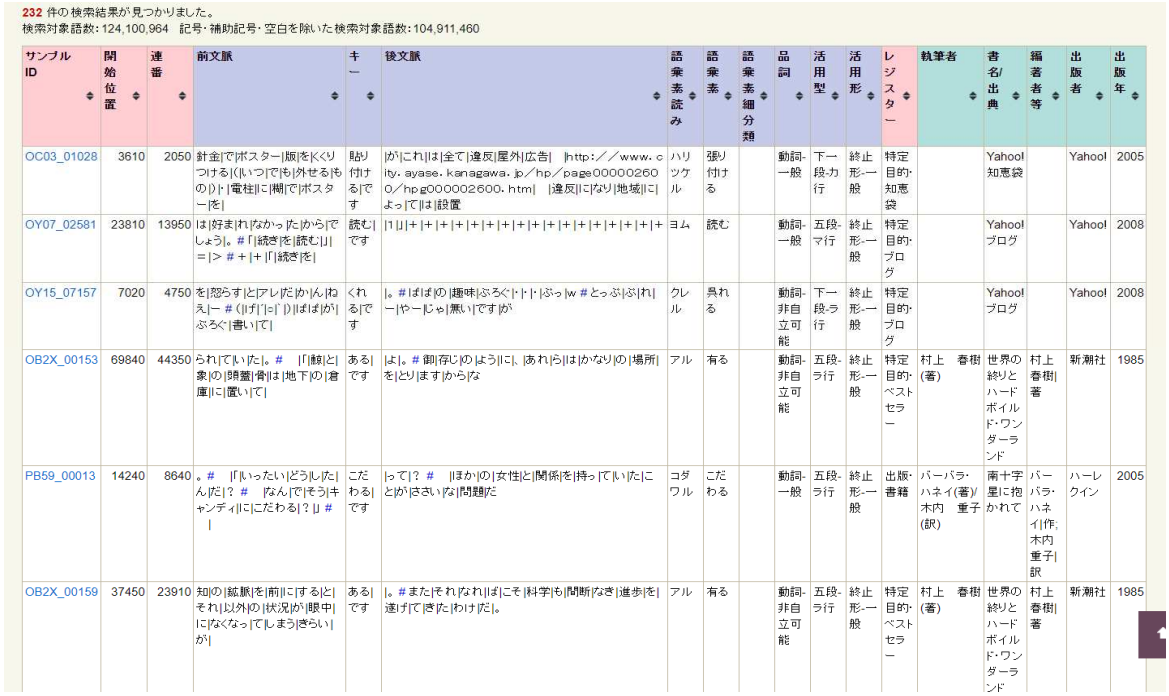
Before finishing the section for corpus annotation, I would like to touch upon two more points. First, annotation is by no means limited to linguistic annotation. There are many non-linguistic annotations that will be of great use for corpus analyses. An example is the so-called "impression rating score" (IRS) given to the CSJ. IRS is a collection of a series of subjective judgments on the impressions that native speakers perceive by listening to the samples of the CSJ. Twenty raters participated in the IRS annotation. They listened to many speech samples of about one minute long extracted quasi-randomly from the CSJ, and evaluated them using the five- or seven-stage rating with respect to various impressionistic bipolar scales like "formal–casual", "polite-impolite", "fast-slow", "relaxed-tensed", etc. The IRS is utilized in studies of phonetics and sociolinguistics for various purposes. There is also IRS-like annotation of written texts for a subset of the BCCWJ [9].

Figure 2: Snapshot of the UniDicExplorer

Second, there is also so-called metadata information, which is indispensable in many fields of corpus analyses. We are striving to provide as much metadata as possible without violating the writers' and speakers' privacy. The metadata in the BCCWJ includes the author's

name, gender and birthyear, the publication year, the publisher, genre of publication, the book/article title, the volume and number, the name of editor(s), and more.





サンプルID	開始位置	連番	前文脈	キー	後文脈	語彙素読み	語彙素	語彙素細分類	品詞	活用型	活用形	レジスター	執筆者	書名/出典	編著者等	出版者	出版年
OC03_01028	3610	2050	針金|で|ポスター|版|を|くくりつける|(|いつ|で|も|外せる|もの|)|・|電柱|に|糊|で|ポスター|を|	貼り付ける|です	|が|これ|は|全て|違反|屋外|広告| |http://www.city.ayase.kanagawa.jp/hp/page000002600/hpg000002600.htm| |違反|になり|地域|に|よっ|て|は|設置	ハリツケル	張り付ける		動詞-一般	下一段-カ行	終止形-一般	特定目的-知恵袋		Yahoo!知恵袋		Yahoo!	2005
OY07_02581	23810	13950	は|好ま|れ|なかっ|た|から|で|しょう|。|#|「|続き|を|読む|」|	読む|です	|1|」|+|+|+|+|+|+|+|+|+|+|+|+|+|+|+|+|+	ヨム	読む		動詞-一般	五段-マ行	終止形-一般	特定目的-ブログ		Yahoo!ブログ		Yahoo!	2008
OY15_07157	7020	4750	を|怒らす|と|アレ|だ|か|ん|ねえ|ー|#|(|げ|゛|=|゛|)|ば|ば|が|ぶろぐ|書い|て|	くれる|です	|。|#|ば|の|趣味|ぶろぐ|・|・|・|ぶっ|w|#|とっぷ|ぶ|れ|ー|やー|じゃ|無い|です|が	クレル	呉れる		動詞-非自立可能	下一段-ラ行	終止形-一般	特定目的-ブログ		Yahoo!ブログ		Yahoo!	2008
OB2X_00153	69840	44350	られ|て|い|た|。|#|「|顔|と|象|の|頭蓋|骨|は|地下|の|倉庫|に|置い|て|	ある|です	|よ|。|#|御|存じ|の|よう|に|、|あれ|ら|は|かなり|の|場所|を|とり|ます|から|な	アル	有る		動詞-非自立可能	五段-ラ行	終止形-一般	特定目的-ベストセラー	村上 春樹(著)	世界の終りとハードボイルド・ワンダーランド	村上春樹|著	新潮社	1985
PB59_00013	14240	8640	。|#|「|いったい|どうした|ん|だ|?|#|なん|で|そう|キャンディ|に|こだわる|?|」|#|	こだわる|です	|って|?|#|「|ほか|の|女性|と|関係|を|持っ|て|い|た|こと|が|ささい|な|問題|だ	コダワル	こだわる		動詞-一般	五段-ラ行	終止形-一般	出版・書籍	バーバラ・ハネイ(著)/木内 重子(訳)	南十字星に抱かれて	バーバラ・ハネイ|作,木内 重子|訳	ハーレクイン	2005
OB2X_00159	37450	23910	知|の|統|脈|を|前|に|する|と|それ|以外|の|状況|が|眼中|に|なくなっ|て|しまう|きらい|が|	ある|です	|。|#|また|それ|なれ|ば|こそ|科学|も|間断|なき|進歩|を|遂げ|て|きた|わけ|だ|。	アル	有る		動詞-非自立可能	五段-ラ行	終止形-一般	特定目的-ベストセラー	村上 春樹(著)	世界の終りとハードボイルド・ワンダーランド	村上春樹|著	新潮社	1985

Figure 3: Query condition setting in Chunagon (upper panel) and the results of the query (lower panel)

2.7 Public availability

By public availability is meant that corpora, or other language resources, are not exclusively occupied by a closed group of researchers, typically the developers and the stakeholders.

Some people take it for granted that openness implies "for free", but personally, I think this is too strong a claim. Very often, corpus development requires considerable cost, so it is permissible to recover a part of the development cost and use it for further development. At the same time, however, the price should be affordable for ordinary researchers.

In our case, all corpora developed in the NINJAL are available for free on the web via a web application called Chunagon（中納言), which is a web interface for the query of morphological information. There are currently about 30,000 registered users of Chunagon, and the users are constantly increasing. Figures 3 shows two snapshots of Chunagon.

The upper panel shows the window where query conditions are specified. In this case, any verb that is in its ending form and followed immediately by an auxiliary verb /desu/ is specified. The lower panel shows the results of the query. Here the 24 hits of verb+/desu/ are shown with the preceding and following contexts, POS information, and metadata. Users can download these results, and the query condition is automatically saved in the system for reuse in future.

Advanced users who want to use the whole body of the BCCWJ and/or CSJ can purchase the whole corpora. The fee for academic use is 50,000 JPY for each. Commercial use is also permitted with these corpora but the fee for profit-making use is set to be more expensive. However, many companies, beyond our expectation, both home and abroad purchased these corpora for commercial use. For the past several years, the corpus sales increased rapidly and reached the level of some tens of million yens. This is due to today' AI boom.

3. Some practical issues in corpus design and compilation

In the previous section, I discussed the basic issues of corpus design mainly from a linguistic point of view. In this section, some practical issues of corpus design and compilation that were not covered in the previous section are discussed.

3.1 Balance in spoken corpus

As mentioned earlier, it is generally very difficult, if not entirely impossible, to define the population of naturally spoken language from which samples are to be drawn. There are, however, some trials to overcome the difficulty.

In the 1960s, some predecessors of the NINJAL designed and conducted so-called "24 hours survey". This is a sociolinguistic survey that recorded and transcribed all utterances uttered by a subject in a day [10]. Although this survey is well known in the history of Japanese sociolinguistics, the same kind of survey has never been tried since then. It is because the research cost was too high compared to the value of the data. It is not therefore

recommendable to construct a “24 hours” corpus. The best thing we can do in this field is to design a well-balanced spoken corpus. But this is also a difficult task, since we don’t know the real inventory of spoken registers in our daily lives and the frequencies with which the registers are used.

Recently, my colleagues made a valuable attempt to solve this problem [11]. Their goal is to construct a new spoken corpus (known as the Corpus of Everyday Japanese Conversation or CEJC) with which they can analyze people’s everyday verbal communication in naturalistic environments. This corpus is expected to deepen our understanding of spoken Japanese by compensating the monologues data of the CSJ.

Instead of immediately starting the corpus design, my colleagues first launched a social survey in which 243 informants balanced in their age and gender were recruited to record the properties of their daily conversations for three days including two weekdays and a holiday. Each conversation was recorded in a questionnaire and classified with respect to the three classes of properties, i.e. A) the attributes of conversation, B) the attributes of conversational situation, and C) the attributes of informants.

The first class is concerned with the attributes like the form of conversation (chat, business talk, meeting, class lessons, lecture), length of conversation, number of interlocutors, the human relationship of the speakers (family members, relatives, teacher/student, business colleague, friend, acquaintance, unknown etc.), the mode of conversation (telephone, net conversation including video), inclusion of foreign speakers, and the language used. The second class is concerned with the attributes like the time (morning, afternoon, etc.) and place (home, workplace, public place etc.) of conversation, and the social activities in which the conversation took place (meal, housekeeping, business, study, recreation, etc.). And, the third class consists of the attributes of the gender, age, and occupation.

It turned out by the survey that, among other things, it is the occupation that caused the largest variability in the attributes of conversation listed above. Panels in the Figure 4 show the variabilities in the form, length, number of interlocutors, time, place, and social activities as a function of the occupation of subjects.

In constructing the CEJC, 600 hours of conversations were video recorded with the help of many volunteers in various situations in their daily lives, but not all of them are included in the corpus. Samples are chosen from the collection so that the distributions of their attributes maximally match the distributions observed in the foregoing survey. The final version of the CEJC will include 200 hours of samples thus chosen. By conducting sampling based on such in-depth preliminary surveys, CEJC has become the world's most balanced conversational corpus.

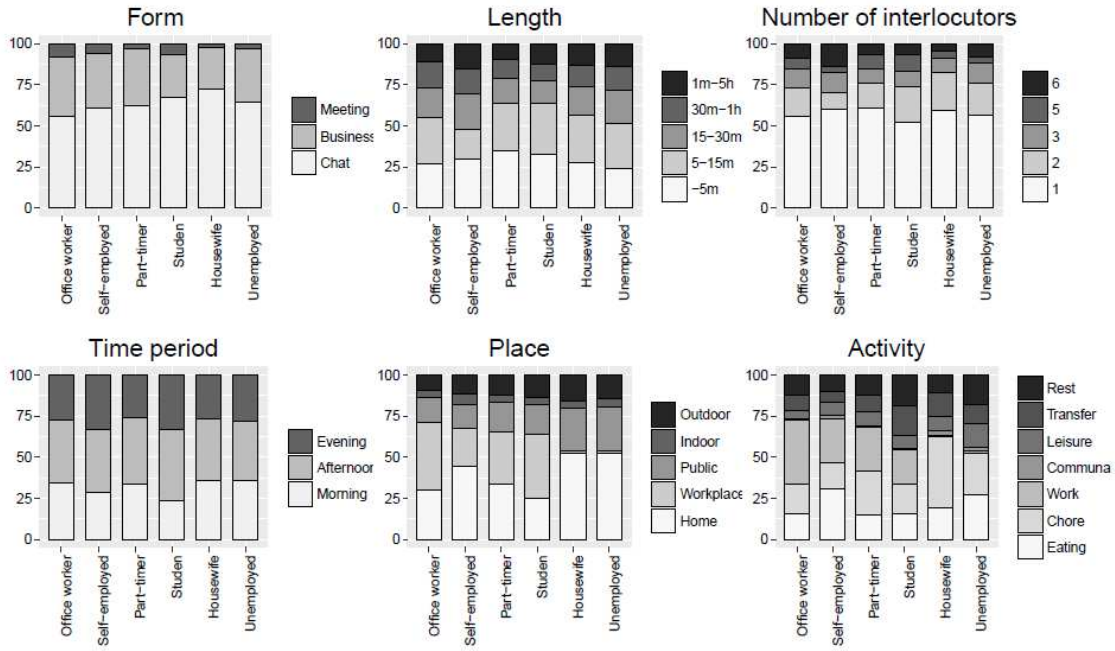


Figure 4: Variability of conversational attributes due to occupation (Cited from [11])

3.2 Copyright clearance

Clearance of the copyright-related issues are the largest bottleneck in the compilation of a contemporary corpus. Although copyright-related issues exist both in spoken and written corpora, it is comparatively easier in the spoken corpora; because the samples of spoken corpora are often recorded specifically for the corpora, it is therefore possible to obtain the permission from the speakers at the time of recording.

Compared to this, treatment of the written samples is quite complex and time-consuming. In the case of the BCCWJ, we treated the total of 24,000 copyright-protected samples in the single register of books; in the end of five-year project, we could obtain permissions from the authors (or copyright holders) of 17,000 samples. It is to be noted that the remaining 7,000 samples were not rejected. The rejection rate was less than 5%. The largest cause of failure in copyright handling is the lack of a way to reach the copyright holders; the problem of so-called "orphan works" [12].

Success rate of copyright clearance was lower in magazine samples than in books. Many magazine articles were written by contract journalists, and the typical contracts in this industry do not include the article about copyright. So, according to the Japanese copyright law, the copyright of the article belongs to the contract journalist who wrote the article. But publishing companies are generally very reluctant to provide us with the contact information of their contract journalists. Behind this reluctance was the influence of the Personal Information Protection Law that prohibited the disclosure of personal information held by

companies without the individual's permission. This law was put in force in 2006, the year when the "Japanese Corpus" project started. The BCCWJ was born under a bad sign!

Earlier in this section, I wrote that copyright clearance was comparatively easy in spoken corpus. There are, however, cases where the clearance becomes extremely difficult. Material of TV broadcast belongs to such cases. At the time I designed the CSJ, I wanted to include samples of TV news program spoken by news casters, so I started negotiation with a broadcasting company; but it was only to find that TV program is a melting pot of copyrights. It was not only the TV casters and the writers of the news article who had the right; all people who were concerned with the program—including the stylist who arranged the outfit of the casters and the florist who arranged the flowers on the caster's desk, and so on and on—had their rights. I don't know if it's true, but anyway, this was the TV station's statement. I had no choice but to give up.

The legal treatment of copyright issues including orphan works differ depending on countries. In the United States, the orphan works will be treated following the legal principle of fair use, which is a principle that if you use a copyright protected work fairly it will not infringe the copyright even if you don't have the permission from the copyright holder. But, unfortunately, such principle does not exist in the Japanese copyright law. It is well known that the Japanese copyright law strongly respects the rights of copyright-holders; that makes the lives of corpus compilers miserable.

I don't know what the situation in Taiwan is like, but I suppose it is not very different from that of Japan. If so, I would like to utilize this opportunity to emphasize the necessity of new copyright law. It is well known that the basic form of today's copyright law was formed in the 19th century, where the creation of copyright-protected materials is a proprietary matter of a limited number of cultural elites. Today, on the other hand, we live in a society where every person has the chance of becoming a copyright-holder due mainly to the Internet.

This is not a place for legal discussion, but I would like to express my belief that copyright law needs to be updated to match today's digital society, and the revised law should be based on automatically trackable posthumous copyright registration system. This is not a wild dream of a layman. Maria A. Pallante, the former United States Register of Copyrights, expressed similar opinion in her famous paper entitled "The next great copyright act" [13]. I will return to this issue in the last section of this talk.

3.3 Documentation

The third issue is the necessity of proper documentation. By documentation is meant the planned recording activity to make the corpus compilation reproducible. Without such an activity, information about the construction of the corpus is rapidly lost over time. Such information is beneficial for designers of similar corpora needless to say, but it also provides essential information for the analysis of the present corpus. Documentation of a project is

something that corresponds to the so-called “lab note” in natural sciences. While it is too obvious a requirement for all scientific activities, it is rarely handled properly, especially in the field of humanities.

The compilation team of the BCCWJ consisted of four subgroups each responsible for the sampling, text formatting, morphological analysis, and copyright clearance; and during the course of the five-year project, we had regular monthly meeting of these groups. Each time, leaders of the subgroups reported their progresses and problems; in addition, they were requested to write down a manual that describing their work. These manuals were shared by all groups and contributed a lot for the mutual understanding. The manuals were updated from time to time. In the end of the project, we were able to publish twenty manuals of about 2,000 pages. These manuals are publicly available on the Web [14].

One common misunderstanding about the documentation is that documentation is something that we do in the end of a project. On the contrary, it has to be done while the project is going on. In so doing, we can find potential problems of the ongoing project, and the time needed for solution can be shortened by sharing the problem with all members of the project. Documentation can be a tool for problem finding and sharing.

4. On Taiwanese national corpora project

In this last section, I would like to put my personal thoughts on the Taiwanese national corpora project based upon my own experiences. By doing this I may run the risk of being arrogant or misleading, but I think it is part of my task as the keynote speaker of this forum.

4.1 Basic spirit of the two new laws

A simple internet query reveals that there are at least twenty national corpora in the world, including British English, American English, Australian English, Welsh, Irish, Russian, Armenian, Slovak, Bulgarian, Czech, Polish, Croatian, Albanian, Greek (Hellenic), Hungarian, Turkish, Tatar, Thai, Ossetic, and Georgian. In addition to these, languages like Chinese, Korean, and Japanese have corpora that can be regarded as national corpora. But all of these seem to be monolingual corpora (To be honest, I haven’t had chance to check all these corpora, but from the information that I could find in the Web, they are most likely to be monolingual). Here, it is important to note that some of the nations that the corpora listed above represent are multilingual nations (like Russia, Turk, and Australia). Despite this fact, the single-language corpora built in these nations were called “national.”

According to my understanding, the two new Taiwanese laws of language policy, i.e., 客家基本法 and 原住民族語言發展法 stand in exactly the opposite spirit, claiming that all languages should be recognized equally as national languages. And these laws also require that all national languages should have their national corpora.

4.2 Goals of the corpora

It is important, however, to note that none of these laws refers to the goal of the corpora. It seems to me as if the compilation itself is the goal. One may say that the corpora are to be designed as general-purpose corpora. But what is a general-purpose corpus? The answer to this question differs from one designer to another unless there are some external restrictions. Without such restriction, the corpora run a risk of serious lack in the interoperability, which is not a desirable consequence both for the linguistic studies and language policies.

My advice for the researchers who are in charge of the national corpora is the importance of setting concrete goals for each corpus; also, if the corpora are compiled independently by different research groups, it is extremely important to have communication among the compilation groups. You also had better have open discussion with the people in the governmental and administration sections before starting the compilation, and while the project is going on. Having the government and administration sections properly understand the project details can be of crucial importance to the maintenance of the corpus. I will return to this issue later in this section.

4.3 The size

As discussed in section 2.4, corpus size is determined by many practical factors including the size of research budget, the goals of the corpus, and the design (especially that of annotations). But there is one more issue that you have to determine before you start the project. That is the relative sizes of various "national" corpora. According to Wikipedia, the number of Hakka speakers in Taiwan is about 3 million, while the number of Ami (or Pangcah) speakers is about 200,000. Do these facts justify that the size of the Ami corpus is one-fifteenth of the size of Hakka corpus?

Of course, there is no simple answer. All we can do is to try to size each corpus based on some objective criteria so that there is no noticeable inequity between languages. Here too, open discussions with the administration and government sections is important.

One issue that attracts my attention at this point is the treatment of the Taiwanese (i.e., Southern Min). Building a corpus of this big language used by some tens of million speakers in Taiwan within the limit of the fixed research budget means diminishing the size of the corpora of the Hakka and indigenous languages. But without it, the "national" corpora of this country are imperfect.

4.4 The balance

The next issue is the balance within a corpus. As for this problem, the first decision should be the ratios of spoken and written registers. In principle, the ratio of spoken and written materials should be variable in each language depending on the sociolinguistic characteristics of the language. And this decision depends heavily upon the goal of a corpus. Personally, I

think that it is better to set a high ratio of spoken language in the case of the indigenous language corpora. There are two reasons for this. First of all, most indigenous languages are mostly spoken rather than written. Second, the collection of recorded speech with proper transcription is something that information processing industry needs badly. The presence of such data will contribute greatly to the solution of so-called under-resource language problems.

4.5 Possibility of a corpus of language behavior

So far in this talk, the word "corpus" is used exclusively to refer to a collection of samples taken from a single language. However, we can also think of a corpus such that whose statistical population is the whole utterances of people in a given society. The CEJC that I referred to in section 3.1 is an example.

If you construct such a corpus in Taiwan, the corpus will inevitably be a multilanguage corpus. In the survey of 陳[15], 92%, 9%, and 3% of Taiwanese people responded respectively that they were fluent in Mandarin and Southern Min, Mandarin and Hakka, and Mandarin and indigenous language. This is the case of people in their thirties and early forties (30-45), but similar patterns are observed in other age groups.

陳 also reported that the choice of language differ systematically according to the social settings like in "religious place", "home", "with friends", "school or public places", and "working place". In the case of people in their thirties and early forties, 32% and 63% choose "台語" (i.e., Southern Min and indigenous languages) respectively in "school or public places" and "home"; Mandarin was chosen by 84% and 75% respectively in the same settings.

These surveys show the presence of diglossia in the Taiwanese society; "diglossia" means a social situation where two language variants or languages are used with different functions in a society [16]. To be more exact, the case of Taiwan had better be called multiglossia. Corpus-based analysis of multiglossia will be quite beneficial not only to linguistics but also for the language policy studies.

4.6 Collaboration with information sciences

Today corpus compilation is not a matter of linguistics alone. Compilation of large-scale corpora is simply impossible without the help of information processing technologies. At the same time, however, the goals of corpus compilation can be different between linguists and information scientists/engineers. The general tendency is that while linguists put more emphasis on the quality of corpora, engineers put more emphasis on the quantity. One important task of the leader of a big corpus project is to find out a reasonable point of compromise between the two groups of researchers.

4.7 Necessity of a long-term support

Lastly, I would like to put emphasis on the fact that the compilation of "national" corpus is, in principle, an ever-lasting project. Because a living language does not stop changing, its corpus needs to be updated continuously or with an appropriate interval. For example, we can't find スマホ ("smartphone") in the BCCWJ (one hundred million words) whose samples were collected mostly from the time period between 2001-2005; on the other hand in the 25.8 billion words NWJC, whose samples were collected between 2011-2014, we find 315,276 instances. This word, which we encounter very frequently every day in all medias, did not exist in 2005 but was disseminated quickly after its birth. I think that the BCCWJ must be updated by the end of 2020s at the latest. And I'm afraid your Sinica Corpus has similar problem.

It is accordingly very important that the national corpora project is financially supported not only in the time of its initial construction but also after its release, for the sake of update. I would like to emphasize this necessity to the esteemed persons in the Taiwanese government who joined the forum today.

Another important governmental support is the revision of the copyright law. As I mentioned above, copyright clearance is the largest bottleneck in the compilation of contemporary corpora. So, if we can compile corpora on the basis of fair use of copyright, it will greatly reduce the burden of corpus compilers.

As I checked in the Internet, Taiwanese copyright law has the article of fair use (the article 65). Since I'm not a specialist in the legal field, I'm not sure if language corpora can be compiled under the Taiwanese principle of fair use. In case the principle does not help the compilers, then the Taiwanese government should support them by making exceptional treatment.

In Japan, recently, the copyright law was revised so that the business of electronic overseas dissemination of modern Japanese books by the National Diet Library is an exception to copyright protection. Corpus compilation by the NINJAL is not included in the revision, unfortunately. But I think the Taiwanese national corpora project is very well worth exceptional treatment, because the project is motivated by the two laws.

5. In place of conclusion

In this talk, especially in the last section, I made comments on the Taiwanese "national corpora" project based upon my limited understanding of the project per se and the social background of the language issues. If there is something ridiculous in my comment, that is entirely my responsibility.

The Taiwanese project is a challenging one from a view point of an expert of language resources. But I believe at the same time that the Taiwanese people's spirit, which is embodied in the two new laws, is wonderful and highly commendable. As a matter of fact, if the Japanese

government enacted such a law one hundred years ago, the fate of the Ainu language might have changed.

Finally, I would like to express my deep gratitude for the people of the Academia Sinica who invited me here. I hope that my talk is helpful to you in some way, and I also hope that ILAS and NINJAL will continue to make useful academic exchanges in the coming years.

References

[1] Maekawa, Kikuo. “Corpus of Spontaneous Japanese: Its Design and Evaluation”. *Proceedings of ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, Tokyo, pp.7-12, 2003.

[2] Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. “Balanced corpus of contemporary written Japanese”. *Language Resources and Evaluation* 48 (2), pp.345-371 (DOI 10.1007/s10579-013-9261-0), 2014.

[3] Asahara, Masayuki, Kikuo Maekawa, Mizuho Imada, Sachiko Kato, and Hikari Konishi. “Archiving and Analyzing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL, Japan”. *Alexandria*, 25 (1/2), pp.129-148, 2014.

[4] 前川喜久雄「コーパスの存在意義」前川喜久雄(編)『コーパス入門』（講座日本語コーパス第 1 巻）.東京：朝倉書店, pp.1-31, 2013.

[5] Biber, Douglas & Susan Conrad. *Register, Genre, and Style*. Cambridge Textbooks in Linguistics. Cambridge: Cambridge Univ. Press, 2009.

[6] 小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕『「現代日本語書き言葉均衡コーパス」形態論情報規定集第 4 版（上・下）』 国立国語研究所, 2011. Downloadable from https://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-02.pdf and https://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-02.pdf

[7] Mecab official website. https://taku910.github.io/mecab/

[8] 小木曽智信・中村壮範「『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用」自然言語処理, 21 (2), pp. 301-332, 2014.

[9] 柏野和佳子・立花幸子・保田祥・飯田龍・丸山岳彦・奥村学・佐藤理史・徳永健伸・大塚裕子・佐渡島紗織・椿本弥生・沼田寛「BCCWJ 図書館サブコーパス全テキストへの文体情報付与結果の分析」第 3 回 コーパス日本語学ワークショップ予稿集, pp.63-70, 2013.

[10] 国立国語研究所『待遇表現の実態：松江 24 時間調査資料から』国立国語研究所報告 41, 秀英出版, 1971. Down loadable from https://db3.ninjal.ac.jp/publication_db/item.php?id=100170041

[11] Koiso, Hanae, Tomoyuki Tsuchiya, Ryoko Watanabe, Daisuke Yokomori, Masao Aizawa, and Yasuharu Den. “Survey of Conversational Behavior: Towards the Design of a Balanced Corpus of Everyday Japanese Conversation.” *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, pp. 4434-4439, 2016.

[12] 前川喜久雄「コーパス構築と著作権処理」人工知能学会誌, 25 (5), pp. 628-632, 2010.

[13] Pallante, Maria A. “The next great copyright act.” *The Columbia Journal of Law and the Arts*, 36 (3), pp. 315-344, 2013.

[14] Downloadable from https://pj.ninjal.ac.jp/corpus_center/bccwj/doc.html

[15] 陳淑嬌「台灣語言活力研究」鄭錦全他(編)『語言政策的多元文化思考』所載. 中央研究院語言学研究所, 2007.

[16] Ferguson, Charles A. “Diglossia”. *Word*, 15, pp.325-340, 1959.