Data Formats and Management Strategies from the Perspective of Language Resource Producers

– Personal Diachronic and Social Synchronic Data Sharing –

Kazushi Ohya
Tsurumi University
Tsurumi 2-1-3, Tsurumi-ku, Yokohama, Japan
oya-k@tsurumi-u.ac.jp

Abstract

This is a report of findings from on-going language documentation research based on three consecutive projects from 2008 to 2016. In the light of this research, we propose that (1) we should stand on the side of language resource producers to enhance the research of language processing. We support personal data management in addition to social data sharing. (2) This support leads to adopting simple data formats instead of the multi-link-path data models proposed as international standards up to the present. (3) We should set up a framework for total language resource study that includes not only pivotal data formats such as standard formats, but also the surroundings of data formation to capture a wider range of language activities, e.g. annotation, hesitant language formation, and reference-referent relations. A study of this framework is expected to be a foundation of rebuilding man-machine interface studies in which we seek to observe generative processes of informational symbols in order to establish a high affinity interface in regard to documentation.

Keywords: Language Documentation, Data Sharing, Description Model

1. Introduction

A study of making, sharing, and using language resources began from a digitization project conducted by R.Busa, 1949(Busa, 1951), from then, the digitization of language resources has presented many challenges(Greenberger, 1964), and this has become a central theme of our research. Nowadays most information is born-digital, and language resources also have already been digitized. In this digital era, the main focus of language resource studies seems to have been changing from creating language resources to exploring how they can be shared and used. However, even if information is pervasive in a digital format, we can only generate information via a physical interface. Digital environments have developed and improved the quality of output information from a system and consequently also the output interface from the system. However, we need a physical interface to enter data and, as is often the case, the interface is not designed with a view of users in a digital system and we are required to fit with the interface a system provides.[1] In this case, we have to change in our plans, data, processes, and sometimes philosophy on entering data, in conformity with requirements from the system. This is one of the reasons why language documentation has been expected in linguistics(Gippert et al., 2006). Language documentation is not an activity that people unacquainted with technologies or computers try to learn them with support from computer scientists, but a study that linguists seek for a way of using computers as data-input tools that computer scientists have overlooked for a long time unconsciously.

In this paper, firstly, we report our projects on language documentation from 2008 to 2016 and, as findings from the studies, importance of personal diachronic data management. Following this, we present two data formats for personal diachronic data management and some criteria as practical measures for descriptive linguists. And finally, we report our on-going research on description environments and a study of a description model.

[1] In the mechanical era, man-machine interfaces had been cultivated through many experiments. However, in the digital era, the quality of man-machine interface, tends to be out of the main focus of academic research and, unfortunately, seems to be a field for gimmicks.

2. Experiments of Language Documentation Research

From 2008, we began research in language documentation, especially on North-East Eurasian languages, under the project LingDy supported by Tokyo University of Foreign Studies.[2] The purpose of this project was to help linguists start learning data management in accordance with language documentation research(Gippert et al., 2006), and to address problems in the adaptation of the protocols proposed in language documentation studies. The objective of this project was in a way to promote data sharing among an academic society and language communities. We made two types of original XML schemes, stored language data into a database server with the Berkeley DB XML engine, and checked feasibility of the schemes as standard formats of language resources used by descriptive linguists(Ohya, 2009; Ohya, 2011). At the end of this project in 2010, we concluded that the multi-link-path data scheme is not beneficial for linguists or data managers because data conversion for the data in this scheme requires more steps than the data conversion in old or simple formats. These findings were unexpected, because the schemes have been proposed by authentic international organizations. The multi-link-path scheme in XML has been regarded as having the advantages of separating definitions of units and structures of data, and has been recommended as a pivotal data scheme that contributes to data sharing. However, as we demon-

[2] http://lingdy.aacore.jp/en/

strated in (Ohya, 2015b), this type of scheme presents a difficulty in multi-link-path management, a primary data constraint and an obstacle to reusing primary or terminal data and annotation or non-terminal data. This results in one-way only data conversion from personal data in multiple formats to the pivotal data stored in archives, which means the opposite directional data conversion is more difficult than the pivotal format advocators indicate. We regarded it as a serious problem that language data in this scheme is almost impossible to be reused in personal data management. In public archives, re-usability of stored data should be ensured not only for other researchers but also for producers of language resources. Without the benefits of conducting archiving processes in linguists' data management, it would be hard for linguists or producers of language resources to have a prospect to integrate their personal data management mechanisms with public archives.[3]

Based on these observations and results, we started a second project supported by Grants-in-Aid for Science Research in Japan from 2010 to 2014(Ohya, 2011). In this project, we used FLEx[4] as an application for entering language resources instead of ToolBox[5] used in the previous project, because it seemed to be a trend among descriptive or field linguists at that time. We examined linguists' field notes in order to ascertain behaviours in data management processes. We expected to find clues to support personal data management. One result, which is actually a by-product, is that we (re)confirmed that linguists have many language resources which are potentially available to other researchers but actually not yet ready to be accessed. Language resources we can use form only the tip of the iceberg that is stored in linguists' personal archives. So far this state of data has not been regarded as a problem with systems. Of course, in any research activities, there is much more working data than published data. In a science study, it is a rare case that primary data is itself published, because it is often huge, sometimes includes noise, and is hard to be interpreted directly. Secondary data usually works as prime data in publication. On the other hand, in a language study, primary data itself is prime data in publication. Especially, in descriptive linguistics, sound data recorded in a field and the dictated texts are primary data in a study and also prime data in publication. As we confirmed in the previous project, the formats proposed as standards have serious drawbacks and, to compensate for them, linguists are required to fulfill many requirements in their actions, e.g. converting their data into other data formats, providing metadata, documenting file naming rules, converting their data to match with a standard, using a common platform, uploading their data to archives immediately, and so on(Thieverger, 2012). It is often a laborious process. The ways proposed by computer scientists or standard creators seem to lead linguists into new and difficult territory. Indeed, their proposals seem to fail to improve linguists' personal data management environments. Therefore, in order to make the language resources in personal environments publically available, to be used in language processing research, we suggest the need to seek a way to establish a consistent data management model from entering data through personal data management to storing it in public archives. In other words, supporting personal diachronic data management may lead to fostering social synchronic data sharing. Thus, in our next project, began in 2014, we have studied not only a data format but also a data management model.

Note that we do not take a stance that language resources can be fully generated without linguists. Some researchers in language processing studies seem to regard language resources as being generated automatically without human intelligence. Such a viewpoint may certainly appear reasonable in cases where targeted language has a community on a large scale, and research targets can be inferred according to the mean of distribution, or when research interest is not in language itself but in human behaviors. However, on the other hand , we are fundamentally interested in language itself and the way it is realized.

3. Data Format Model for Personal Diachronic Data Sharing

In the on-going project, supported by Grants-in-Aid for Science Research in Japan from 2014 to 2016, we have studied a data format for language resources, especially (a) annotated text data and (b) time information, to connect (a) and (b) in field notes or working data in application software. From interviews with linguists, we present the following findings: (1) Linguists are using well-used applications such as ToolBox, FLEx, ELAN[6] and such to produce final data to be issued or recorded to confirm their research attainment. (2) A large proportion of data except that which is stored in applications, is preserved in field notes, or is recorded as plain text data when the linguist has considerable knowledge about computers. And other ordinary linguists are eager to know the ways to handle language data with computers, and consult a language documentation study. (3) However, even linguists using text data for their primary data are uncertain of a way to write annotation in text data and relate it to the referred description. They use applications such as Excel, ToolBox, FLEx and such to the extent of needing connectivity or validity as processable data(Durand et al., 2014). (4) The relation between sound and the text data is kept or ensured by linguists' recognition according to these file names. It is a rare case to make a file of metadata to describe the relation. In a sense, new digital tools may make linguists confused and may not contribute to the promotion of data sharing.

From these observations, we set the following targets to research: (1) To ensure that text data which linguists produce is available as it is or with a little additional process in converting into other data formats in the future, we must show linguists simple design criteria in order to check their data structure or alignments of data units in the text data. (2) To ensure the connection of sound and text data without

[3]Here we do not discuss a problem relating to responsibility linguists have for language communities, or for openness of peer-review systems. This is a problem on a data-flow mechanism of information systems.

[4]http://fieldworks.sil.org/flex/

[5]http://www.sil.org/computing/catalog/show_software.asp?id=79

[6]https://tla.mpi.nl/tools/tla-tools/elan

depending on specific applications, we have to establish a simple system of data management especially for time information and IDs.

3.1. Model for Interlinear Style

As an idea of coping with the aforementioned first requirement, we set simple design criterion for making text data with a simple syntax. The simple design criterion is a model of data formats. In this sense, we take a stance to advocate sharing models or semantics instead of schemes or syntax in order to ensure personal diachronic data sharing.

As a model for a basic form used in a so-called interlinear style familiar to linguists, for example, we propose the followings.

Syntax:

$$
\begin{aligned}
&CORPUS := SNT+;\\
&SNT := AL, ANN+, AAL+, ER;\\
&ANN := (\{\text{any strings}\}, DLM)+;\\
&AL|AAL := \{\text{any strings}\};\\
&ER := \{\text{empty record}\};\\
&DLM := \{\text{delimiters such as space}\};
\end{aligned}
$$

Constraint: any units of the same type in a parent unit have the same number of child units.

SNT is a data unit at a level of sentences, which consists of annotated line(AL), annotations(ANN), alternative annotated line(AAL), and an empty record(ER). In this model, ANN is list data, in which each cell is manifested by a delimiter(DLM). According to the constraint, in each sentence level, the numbers of sub-parts of an annotation are the same in all the annotations of the sentence. And, the number of ALL is the same in all the SNTs.

This kind of simple criteria is easy for linguists to adopt in their activities of making and checking their data. Furthermore, providing an application to check this constraint will help them concentrate on their own work. This constraint ensures their data to be used to do ordinary requirements they expect even in any kind of schemes the data has. However, this criterion does not apply to the full processes that would be requested in the future. For example, this criterion does not ensure a process to specify the order of units or to select the same type of morphemes. If there is a possibility that more sophisticated processes are expected in the future, linguists need to adopt more constraints, e.g. rules for ID management. Preparing this kind of sets of syntax models and criteria is expected to succeed in reassuring linguists about the lifespan of their data. Computer scientists' work is providing multiple kinds of conversion services based on the models, the criteria, and actual data shcmes linguists use to ensure the lifespan of the data.

3.2. Model for Time Information

As an idea of coping with the aforementioned second requirement, we set a simple model criterion for time information that is reported in (Ohya, 2015a). This format for time information is tentatively called GIST; a format of general information of sub-time for linguistics. This format is a set of records consisting of identifiers of time-based objects in super-set order. Each identifier of time-based objects is a super or sub element of the adjacent elements, thus this format can be regarded as for indicating sub-time information. The identifiers are an ID or an equivalent of it such as a file name, or a pair of time information with start and end time-stamp.

3.2.1. Syntax

A syntax of GIST is defined as follows.

$$
\begin{aligned}
&GIST := I+;\\
&I := (N|T)+;\\
&N := NAME;\\
&T := (TIME, TIME);\\
&TIME := hh : mm : ss([.,]d+)?;\\
&NAME := \{\text{any strings}\};
\end{aligned}
$$

N and T are identifiers(I) of time-based objects. N is a name and T is a pair of time-stamps represented by [hh]:[mm]:[ss] style that is similar to a part of times in ISO 8601 format, and [ss] can be extended with a comma or dot and decimal fractions. A name for N is a string of characters for the time being.

3.2.2. Semantics

Semantics of GIST is simple: a left object is a super object of a right object in a sequence of objects. An object is indicated by an identifier, which can be a name(N) or a pair of time-stamps(T). Given an object is denoted by an identifier(I), semantics of GIST is defined as follows.

$$[\![I_1 I_2]\!] := I_1 \supseteq I_2$$

If I is expanded into N or T, expressions with a name and a pair of time-stamps are as follows.

$$
\begin{aligned}
&[\![N_1 N_2]\!] := N_1 \supseteq N_2\\
&[\![NT]\!] := N \supseteq T\\
&[\![TN]\!] := T \supseteq N\\
&[\![T_1 T_2]\!] := T_1 \supseteq T_2
\end{aligned}
$$

3.2.3. Sample Implementation

As sample applications using this GIST format, we made software Sclip(sound clip)[7] that cuts out a part of the sound according to instructions in a GIST format. Provided the following instructions and a shared sound data file are there,

```
00:00:01.2,00:00:50.3,file1.wav
00:00:01.2,00:00:50.3,00:00:00,00:00:15,file2.wav
00:00:01.2,00:00:50.3
00:00:01.2,00:00:50.3,00:00:00,00:00:15
00:00:02.2,00:00:50.3,file4.wav,00:00:01.2,00:00:10.32
```

the sclip makes a new sound file named file1.wav from the shared sound file according to the first record, and a new sound data of file2.wav from 00:00:01.2 to 00:00:16.2 in the shared sound data according to the second record. If there is no name at the final item in a record, the sclip defines a file name automatically and saves a result of partial

[7] https://sites.google.com/site/lingdytextarchive/software

sound data. The functions of extracting a part of sound from a sound object and giving a name to a part of sound are iteratively activated while corresponding to rules in GIST.

3.2.4. Inconvenience of GIST

Using this GIST format has the inconvenience of requiring a system of ID management to realize the expected environment of data management. For example, a metadata file that indicates information about connection between a file in a GIST format and the targeted annotation data file requires an implementation to resolve IDs used in the file. So far, it has been taken for granted that software of multimedia players is used to ensure the relationship. However, in order to make sound data as primary data in linguistics, sound data should ideally be independent from any application environments. Thus, we adopt a plain text format and minimum data scheme in GIST. But, on the other hand, our solution with GIST presupposes the existence of ID-resolving services. So, it can be said that this approach is the second best solution to seek for trade-offs between actual computational environments and the needs of linguists. And, a trade-off point usually changes as an environment does.

4. Environment of Description

In our three projects we have confirmed that using a common application among linguists is an unrealistic solution and using a shared scheme such a standard proposed as multi-link-path model, is also unrealistic or at least unsuitable for linguists. Thus, we have proposed the necessity of data conversion services to support linguists instead of a common standard format, and an idea of sharing a common model to ensure data sharing. However, it is also true that our proposals do not seem to be perfectly in tune with our purposes of supporting personal data management activities that ensures long-life data, or diachronic data sharing, because any ID-resolution system needed for a proposal is difficult to be regarded as a lifelong service. Therefore, we decided to reconstruct our research approaches from scratch. We changed our research focus from a data management system to the way how linguists produce language resources.

In the first place, we started examining how linguists, especially engaged in descriptive or field linguistics, denote, define, and enter language data in field notes and on computers. In interviewing informants and observing language activities, linguists in a field often record language information on a field note as in Fig.1.

In the field note we can find traces which show how linguists struggled with keeping the information they noticed or their thinking at a particular moment, e.g. precise dictation, clear correction, distinctively noted comments, working records, and unconscious or uncoded conscious limitation of information on a page. These ways and techniques are usually mastered by linguists through their experiences in a field and from a limited amount of knowledge gathered from books or lectures. These kinds of traces can be seen not only in a field note, but also in the behavior of active readings. For example, in reading books, we sometimes underline a word or a phrase to indicate something, and if

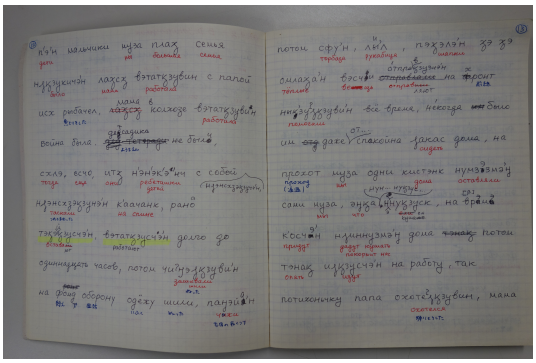

Figure 1: A page of a field note

we want to clear up the reason, we add special marks like asterisks or comments near to the line, or we make a line as a link to indicate the connection(Marshall, 2010; Pearson et al., 2014). Within the limitation of two dimensions we seem to be using common strategies to denote information. If we can set a frame for overall language activities to observe states of information generation, it will be possible to capture a wider range of language activities with a simple principle.

In all these states of description, we can see types of annotations and ways of generating annotations. In the present project, we are seeking a general description model that reflects all these states of description based on a principle that has been tested in a study of markup or meta languages. From a linguistic viewpoint, language data based on this model can be expected to reflect a principle of language dynamics or movements(Coseriu and Geckeler, 1981). From a viewpoint of an annotation study, this general model can bridge the gap between the so-called data-centric approach and the so-called document-centric approach(Landow and Delany, 1993). This model can make it possible to record diachronic data movements and also provide criteria to estimate what function the present plain text will be applied to after additional processing in some applications, a process which is important for language documentation.

5. Frame of Description Study

In order to establish a general description model, as a working hypothesis, we set three phases: (1) an embodying phase, (2) an engraving phase, and (3) an encoding phase.

In the embodying phase, we observe data-making processes in time sequence, like entering, revising, filling, normalizing, and so on. This phase can be regarded as a field for processes of how to determine data units. It is a diachronic study of descriptive examination. As a sub-field of this phase, we presume seven steps: entering data, setting a temporary data unit, examining data units, supplying missing data, setting a data structure, data validation, and data conversion. These kinds of processes in this phase have been studied in software engineering and system science.

In the engraving phase, we observe ways of adding information like establishing reference-referent relations, annotating, anchoring, and so on. This phase can be regarded as a field for ways to denote information. It is a synchronic

study of descriptive examination. As a sub-field of this phase, we presume five types: (a) annotation by characters, (b) annotation by non-characters, (c) indicating a place of annotation, (d) combinations of the previous type a, b and c, and (e) combinations of the type d and a reference-referent relation or link. In annotations as a result of active readings and in descriptions on fieldnotes, we can find traces of activities in this phase. Traces of these activities have been regarded as demonstrating personal techniques of their working style and have not been regarded as a target to examine a limit of information capacity on a medium or a limit of abstraction of informational representation on a medium.

In the encoding phase, we observe rules of descriptions like syntax, instance patterns, abstract data models, and so on. This is a study of markup language itself. It is just a field of a formal (meta)language study. As a sub-field of this phase, we presume four steps: syntax, descriptive rules, set phrases, and descriptive style(Ohya, 2001). This phase has been studied in part in meta-language research, or digital humanities especially text-encoding research. However, they have not regarded physical traces of annotation on paper as a kind of realization of a unit of information in an engraving phase and relating the syntax of annotation in an encoding phase. For example, reference-referent relations appear both in signs on a engraving phase and syntax of ID-IDREFS descriptions on a encoding phase, and they seem to be in the same kind of classification.

5.1. Example

As examples of analysis on the description frame, we observe three enlarged parts of the Fig.1. In Fig.2, we can see a phase of entering data and concurrent revision.

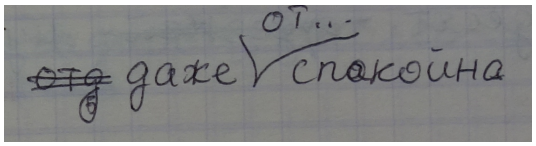

Figure 2: Sample 1

The first three letter are deleted after initial writing of dictation and at the same time the two new letters with dot leaders added with uncertainty as a kind of annotation. In terms of an embodying phase, we can see two kinds of entering data and one supplying data processes. In terms of an engraving phase, we can see the type a and c.

In Fig.3 we can see another step of entering data, in which a linguist adds annotations in red ink(the bottom-right) and revises the data denoted previously in red ink(the double strikeout).

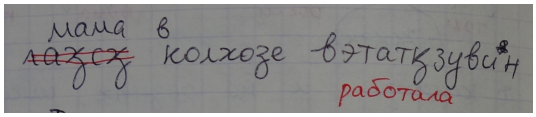

Figure 3: Sample 2

This linguist tends to use the same color to indicate the same stage of description level in a time sequence. In terms of an embodying phase, we can see two kinds of entering data processes. In terms of an engraving phase, we can see the type a and c.

The difference between descriptions in Fig.2 and 3 is not in a process level but in a semantic or model level. Fig.3 includes descriptions of so-called linguistic annotation at the second data entering process in an embodying phase. This observation indicates that in addition to models for description processes like an embodying phase and for appearances like an engraving phase, we need a model for semantic objects that are possibly targets of processes. However, since a data unit is usually defined through some processes in embodying and engraving processes, assignment from objects in the two phases to one in a semantic model has to be applicable to a change of unit or type.[8]

In Fig.4 we can see another new step of entering data that is another kind of annotation written in blue ink(the top-left third line). This linguist revises the second step descriptions at this third step with the third color(the bottom-left second line). And we can see a mark of place for something relating annotation, which is indicated as a line mark in yellow highlight ink(between the left second and third, and the right first and second lines from the bottom up).

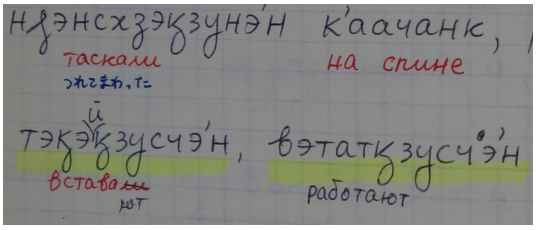

Figure 4: Sample 3

In terms of an embodying phase, we can see at least five kinds of entering data processes. In terms of an engraving phase, we can see the type a, c, and e.

Fig.4 is very informative for computer scientists and engineers. For example, in this case, the third data entering and abstraction level of semantics can be regarded as providing a clue about the timing of defining data units, and also new findings that have not been confirmed until then. At this time, a data unit is defined and the other would-be data units are supplied with this data unit.

At present, we are examining the embodying phase with types of annotations and ways of generating annotations observed in field notes and annotations during active reading.

6. Conclusion and Future Work

We have only began our examination of embodying and engraving phases of our description model in field notes. We realize the need for more variations of descriptions and more field notes we examine. However, this is not achieved

[8] We are now examining timing of defining data units in field notes. The timing of defining data units seems to be different in conjunction with a limit of abstraction level on semantics. This changes depending on linguists which means a style of descriptions.

by enlarging the sale of research. It requires careful consideration in order to accomplish it. Basically, field notes are usually personal memoranda, which include information that is not readily disseminated. For example, as we observed in Fig.4, sometimes an outer edge of linguist's knowledge is exposed in a field note. And, descriptions in a field note potentially contain a negative influence to field work in a questioning style and a social relation especially when informants of the language resources find out the descriptions, for they become to know intentions the linguist had at that time. Therefore, we need a circumspect plan to proceed this study. On the other hand, it is true that we have a good response from the observation and potential of this research frame to bring a clear perspective for generative states of information in descriptions. We plan to confirm sub-fields of each phase by the end of the current project.[9]

7. Bibliographical References

Busa, R. (1951). *Sancti Thomae Aquinatis Hymnorum Ritualiu Varia Specimina Concordantiarum – A first example of word index automatically compiled and printed by IBM punched card machines –*. Fratelli Boccao.

Coseriu, E. and Geckeler, H. (1981). *Trends in Structural Semantics*. Gunter Narr Verlag.

Jacques Durand, et al., editors. (2014). *The Oxford Handbook of Corpus Phonology*. Oxford University Press.

Jost Gippert, et al., editors. (2006). *Essentials of Language Documentation*. Mouton de Gruyter.

Martin Greenberger, editor. (1964). *Computers and the World of the Future*. The MIT Press.

George P. Landow et al., editors. (1993). *The Digital Word: Text-Based Compouting in the Humanities*. The MIT Press.

Marshall, C. C. (2010). *Reading and Writing the Electronic Book*. Morgan & Claypool Publishers.

Ohya, K. (2001). Necessities on a descriptive level for reusing metadata desctiptions. *Proceedings of DC2001*, pages 97–100.

Ohya, K. (2009). Data structure for minority language corpora (in japanese). *IPSJ Symposium Seriese*, 2009(16):115–122.

Ohya, K. (2011). Missing services in language documentation in terms of information processing – the report of lingdy project – (in japanese). *IPSJ Symposium Seriese*, 2011(8):59–66.

Ohya, K. (2015a). A general format for time information to the first-class data of general linguistics. *The 4th International Conference of Language Documentation and Conservation (ICLDC4)*, http://hdl.handle.net/10125/25368.

Ohya, K. (2015b). Corpus sharing strategy for descriptive linguistics. *Journal of Japanese Association for Digital Humanities*, 1(1):68–85.

Pearson, J., Buchanan, G., and Thimbleby, H. (2014). *Designing for Digital Reading*. Morgan & Claypool Publishers.

Nicholas Thieverger, editor. (2012). *The Oxford Handbook of Linguistic Fieldwork*. Oxford University Press.

[9]This study has been supported by many linguists. I would like to thank these researchers; Ritsuko Kikuzawa, Iku Nagasaki, Chikako Ono, Itsuji Tangiku, Yukari Nagayama, Linjing Li, Norikazu Kogura, and Kosei Otsuka.