Development of agricultural vocabulary to enable simple and effective searching for agricultural information

Akiko Okabe¹, Akane Takezaki¹ and Mitsuyuki Saito¹

1 Tsukuba Office, Agriculture, Forestry and Fisheries Research Council Secretariat, Japan, olive@affrc.go.jp

Abstract

This paper describes the development of agricultural vocabulary in both Japanese and English to enable simple and effective searching for agricultural information. In 2001, we cooperated with the Food and Agriculture Organization (FAO) in the AGROVOC project. AGROVOC, which was created by the FAO and the Commission of the European Communities, is a multilingual structured thesaurus containing terminology for all subject fields in agriculture, forestry, fisheries, food and other related domains. The terms in the thesaurus are organized on the basis of three types of relationships (namely, hierarchical, equivalent, and associative relationships), and it is published in 16 languages. In 2002, we translated approximately 27,300 English AGROVOC terms into Japanese, and submitted the list to the FAO. However, we faced several problems while translating the English terms into Japanese. Initially, it was required that the priority of some Japanese terms be determined because an English name often corresponded to more than one Japanese term. Moreover, there are not many Japan-specific agricultural terms in AGROVOC. Therefore, we have developed the Japan agricultural thesaurus (JAT) in order to add the relevant Japanese terms to AGROVOC, since in its current state, the Japanese version of the AGROVOC thesaurus cannot be used for retrieving Japanese agricultural information. The JAT includes Japan-specific agricultural terms (e.g., mulberry field), Japanese places (e.g., Ariake Sea), and common terms (e.g., abscission layer), which are presently not included in AGROVOC. We manage the JAT through a thesaurus management tool. JAT contains approximately 48,000 terms written in both Japanese and English. For each term, a word block is displayed showing its hierarchical and equivalent relations to other terms. We will use JAT to conduct a morpheme analysis of agricultural information.

Keywords: AGROVOC, Japan agricultural thesaurus, agricultural vocabulary, searching for agricultural information

Introduction

We, Tsukuba Office, Agriculture, Forestry and Fisheries Research Council Secretariat, Japan (hereafter referred to as Tsukuba Office), have compiled a database by collecting information necessary for research and development in the fields of agriculture, forestry and fisheries, and have provided access to the database via the Internet. The database is available free of charge to those involved in the agricultural, forestry and fisheries industries, providing the public with bibliographic information and full-text data.
In addition, we participated in a project to develop a multilingual agricultural thesaurus (hereafter referred to as AGROVOC) under the initiative of the Food and Agriculture Organization (FAO) of the United Nations by translating AGROVOC into Japanese.

However, the process of preparing a Japanese edition of AGROVOC revealed that AGROVOC had a relatively small number of scientific terms unique to Japanese agriculture, forestry and fisheries. In light of this, Tsukuba Office developed the Japan Agricultural Thesaurus (hereafter referred to as JAT), where terms specific to Japanese agriculture, forestry and fisheries are added to the Japanese edition of AGROVOC.

Here we describe these linguistic resources (i.e., AGROVOC in Japanese and JAT as an extension of AGROVOC in Japanese) that enable simpler yet more efficient searches of agricultural information.

**Overview of AGROVOC**

AGROVOC is a multilingual thesaurus developed by FAO jointly with the Commission of the European Communities (CEC) at the beginning of the 1980s, and it has been updated by FAO as a result of international cooperation (Aoki, 2004). It is based on versions in English, French and Spanish, and is also used for searching databases and reports produced by FAO. Appending AGROVOC terms to information as keywords enables the simultaneous retrieval or classification of data in multiple agricultural databases.

AGROVOC is available in 16 languages including Czech, German, Hindi, Hungarian, Italian, Japanese, Persian, Polish, Portuguese, Slovak and Thai in addition to the five official languages at FAO of Arabic, Chinese, English, French and Spanish.

Approximately 28,000 descriptors and around 10,900 non-descriptors are currently recorded in AGROVOC. This AGROVOC database is open to the public on the FAO Web site and users can search for information by entering a query in the search window (FAO, 2008) (Fig. 1). AGROVOC can be downloaded free of charge from the FAO Web site unless it is to be used for commercial purposes.
Tsukuba Office has participated in FAO’s AGROVOC project since 2001. In 2002, a Japanese edition of AGROVOC, which recorded approximately 27,300 technical terms translated into Japanese from all subject fields in agricultural, forestry, fisheries, food and related domains, was proposed to FAO.

However, several problems occurred when translating AGROVOC terms into Japanese because of the difference in concepts of agriculture between Western countries and Japan. Some of examples are described below (Takezaki, 2008).

1. Selection of descriptors and non-descriptors

A living organism often is known by several names and sometimes has different Japanese terms in different scientific fields—even if it has only a single name in English. Therefore, it was required that the priority of these terms be determined because an English name often corresponded to more than one Japanese term. For this reason, a term with the highest priority based on the rules of relevant academic societies and the prevalence of the term was defined as a descriptor. That said, the priority of terms needs to be regularly revised because it may change with the trends of the times.

In addition, we had no choice but to translate multiple English terms into a single term in Japanese. When only one name was a descriptor and the others were non-descriptors among...
these English terms, it was determined that there was no problem in translating all the English terms into a single term in Japanese. When all (or some of) such English terms are descriptors, however, an identical term in Japanese cannot be applied, because it fails to provide a one-for-one translation. Further consideration is required to resolve this problem.

(2) Notation of terms with multiple meanings

In AGROVOC, when a word in English can have more than one meaning, each meaning of the word is described in parentheses after the word. We also described differences in a Japanese term having multiple concepts in the same manner. However, when there is an equivalent relationship between English terms, for example, a scientific name and a common name, or a formal name and an abbreviation, no explanations in parentheses were provided. Since terms in AGROVOC have been increasingly provided with a scope note in recent years, the present descriptions in parentheses should be changed into the form of a scope note in the future.

(3) Shortage of scientific terms unique to Japanese agriculture, forestry and fisheries

Terms that are only used in Japan, such as “natto”, are not included in the AGROVOC. This is a common problem among countries participating in the AGROVOC project. For other examples, the term kou-chi “cultivated land” is recorded in AGROVOC as a narrower term for nou-chi “farmland”. Kou-chi “cultivated land” also has the narrower term sui-den “paddy field”, and “paddy field” has further narrower terms of roukyu-ka-sui-den “degraded paddy field”, rou-sui-den “leaking paddy field”, and yatsu-da “paddy field in ravine” in Japanese, but AGROVOC does not include these narrower terms. Thus, terms specific to Japanese agriculture, forestry and fisheries that were not included in AGROVOC had to be added.

**Development of JAT**

We found that terms included in the Japanese version of AGROVOC were not necessarily useful for searching Japanese agricultural information because of the lack of terms specific to Japan. Accordingly, we developed JAT, which includes scientific terms specific to Japanese agriculture, forestry and fisheries based on the hierarchical and equivalent relationships with terms included in AGROVOC.

JAT was developed by collecting scientific terms from the indices of academic dictionaries and encyclopedias used by researchers in the fields of agriculture, forestry and fisheries including Nougaku Dai-Jiten [Encyclopedia of Agriculture] (Yamazaki, et al., 2004), Encyclopedia of Forest and Forestry (The Japan Forestry Technology Association, 2001) and Japanese-English English-Japanese Dictionary of Fisheries (Kaneda, 1999), and by referring to various glossaries of technical terms developed by Japanese scientific societies related to agriculture, forestry and fisheries. Scientific fields that contributed new words to the JAT include biology, breeding, cultivation, disease, insect pest, food processing, chemistry, soil science, environment, animal husbandry, agricultural civil engineering, agricultural implement and machinery, agricultural economy, fisheries and forestry. Associative relations that have no clear definitions in ISO 2788 were not included in JAT.

JAT currently has approximately 48,000 words. It also includes Japanese agricultural terms (e.g., mulberry field), Japanese place names (e.g., Ariake Sea) and other common terms (e.g., abscission layer) (Fig. 2). Hierarchical relationships among terms were constructed by adding the new terms to the lowest layer or the intermediate layer among terms that have
already been included in AGROVOC (Fig. 3). Many of the words added to the intermediate layers are terms whose concepts have been present, but the terms themselves were not recorded in AGROVOC.

We are going to propose these additional terms to FAO in order to improve AGROVOC so that to become a thesaurus that can enhance the precision of searching literature related to Japan.

mulberry field, autumn declined paddy, Bai-u front, carbon dioxide generator, oil cakes, sugar-acid ratio, abandoned cultivation, Ariake Sea, abscission layer, Accipiter gentilis, acetalated wood, Acheilognathus, acid rain, air-drying effects on ammonification, Acipenser medirostris, actinomycetes, active transport, active fault, active oxygen, Adenophora, Agano River, adenovirus, adhering water, air masses, Anguilla japonica, aerobic condition, Adoxophyes honmai, adsorption water, Agricultural Chemicals Inspection Station, aerial roots, aerosol, agar medium, agricultural water use, agaricus bisporus, aging society, aggregate structure, air dry soil, a crossbreed between the mallard and the domestic duck, Agrobacterium, algorithms, allicin, cross pollination, alluvial fans, alpine lake Oze-numa

Fig. 2. Examples of terms added to JAT

Campanulatae

Compositae

Arctium

Arctium lappa

Passeriformes

Corvidae

Corvus

Fig. 3. Examples of hierarchical relationships in JAT

New words (in boxes with dotted lines) are added to the lowest layer or the intermediate layer of the AGROVOC words (in boxes with solid lines).
Management of JAT

Although JAT was previously managed using Microsoft Office Excel, file sizes grew with the increase of glossaries included, and the complexity of operation also grew. Accordingly, we introduced a thesaurus management tool (DicBox ED;Runet) that enabled us to visualize hierarchical relationships in the thesaurus and accomplish easier revision of contents (Fig. 4). Both Japanese (expressed in kanji and hiragana) and English terms are included in this management tool.

Each term is expressed in a hierarchical structure and the depth of the hierarchy is indicated by a black-colored numerical figure at the head of a descriptor. A red mark is placed at the head of a non-descriptor. When you click a non-descriptor, its descriptor is also expressed in gray. Information on terms can be changed on an individual term basis. Revision of hierarchical relationships can also be easily performed by simply dragging and dropping relevant terms.

Fig. 4. Information management screen of JAT
Future uses of JAT

With the prevalence of the Internet, there has been an increased trend toward obtaining information from Web sites. To enable more precise information retrieval, there is an increased necessity for more substantial language resources such as dictionaries and thesauruses.

At the Tsukuba Office, we have developed a JASI database, which lists bibliographic information on literature published in Japan covering agriculture, forestry and fisheries. It is open to the public free of charge. We have applied JAT to a morpheme dictionary and incorporated JAT into the automatic index term extraction system of JASI. Although the system is currently under verification, it is expected that it will be applied to analyzing retrieval trends and searching related terms.

We also plan to provide JAT to researchers looking to develop effective search systems. We will be able to improve JAT to make it more reliable and expand the fields to which JAT is applied by providing JAT to researchers engaged in the development of retrieval technologies, and by receiving feedback on the results of the use of JAT from these researchers.

Thus, JAT, a linguistic resource developed by Tsukuba Office, has the potential for a wide variety of uses, but it has also the possibility of losing its value because it may become obsolete given the rapid advance in scientific technologies. Therefore, JAT requires the constant addition of new terms, and the revision and deletion of old terms—as do all thesauruses. These steps will require labor, money, and a wide range of expertise covering all the subjects of agriculture, forestry and fisheries. It is essential to construct an inspection system that enables the regular evaluation of data in JAT.

References