# Community-based construction of draft and final translation corpus through a translation hosting site Minna no Hon'yaku (MNH)

**Takeshi Abekawa[†], Masao Utiyama[‡], Eiichiro Sumita[‡] and Kyo Kageura[♯]**

[†] National Institute of Informatics
[‡] National Institute of Information and Communication Technology
[♯] University of Tokyo
[†] abekawa@nii.ac.jp, [‡] {utiyama,eiichiro.sumita}@nict.go.jp, [♯] kyo@p.u-tokyo.ac.jp

**Abstract**

In this paper we report a way of constructing a translation corpus that contains not only source and target texts, but draft and final versions of target texts, through the translation hosting site Minna no Hon'yaku (MNH). We made MNH publicly available on April 2009. Since then, more than 1,000 users have registered and over 3,500 documents have been translated, as of February 2010, from English to Japanese and from Japanese to English. MNH provides an integrated translation-aid environment, QRedit, which enables translators to look up high-quality dictionaries and Wikipedia as well as to search Google seamlessly. As MNH keeps translation logs, a corpus consisting of source texts, draft translations in several versions, and final translations is constructed naturally through MNH. This corpus can be used for self-learning by inexperienced translators on MNH, and potentially for improving machine translation.

## 1. Introduction

Recently, the importance of not only bilingual parallel corpora but also monolingual parallel corpora has been recognised more and more strongly (Barzilay and McKeown, 2001; Tono, 2009). A corpus that contains both draft and final translations made by humans, together with their source texts is at once a bilingual and monolingual parallel corpus. This type of corpus is especially useful and important for two reasons. First, it can be used for the training of inexperienced translators. For instance, the MeLLANGE corpus, which contains different versions of translation, is well known for its usefulness in translator training (MeLLANGE, 2009). The use of such corpora in translator training is becoming all the more important as we are witnessing the growth in the number of translations by volunteer translators. For instance, many NGOs and NPOs depend heavily on volunteer translators to meet their translation needs, and the training of these volunteer translators remains an arduous task for them. If corpora consisting of draft and final translations become widely available, self-training of translators will be promoted. Secondly, this type of corpus can provide a useful information for improving the performance of machine translation and translation-aid systems. With these applications in mind, the authors have been constructing a corpus consisting of source texts, their draft translations, and the final translations (henceforth "SDF corpus" for succinctness) with detailed tags for modification patterns (Abekawa and Kageura, 2008a).

There are, however, several problems in constructing and making use of such a corpus: (i) the data are not readily available, because human translators in general are reluctant to make draft translations accessible, even when they keep them (a reasonable amount of data should be obtained, part of which should preferably be tagged, for it to be useful for improving NLP technologies (Abekawa and Kageura, 2008b)); (ii) the mechanisms that enable human translators to make use of these translations are not readily available, so translators do not recognise the benefit of making draft and final translations available. In order to solve these problems, at least partially, we are taking advantage of the translation hosting site Minna no Hon'yaku (MNH), which has been publicly available since April 2009 at `http://trans-aid.jp/`. This paper introduces how an SDF corpus is constructed on MNH and how it is used by online volunteer translators on MNH. It also provides the basic statistics of the current data in the SDF corpus accumulated on MNH. Throughout this paper, we focus on English to Japanese translation. This is because, although MNH is essentially multilingual (some users work on Japanese to English translations, and we are planning to provide resources for English to Chinese, Chinese to English, Japanese to Chinese and Chinese to Japanese translations within 2010), as of February 2010 the majority of translations are made from English to Japanese.

## 2. Minna no Hon'yaku (MNH)

Minna no Hon'yaku (MNH) is an online translation hosting site (Figure 1), which provide the following functions (Utiyama, et. al., 2009):

1. Anybody can register with MNH anonymously. The registered user is provided with her/his own personal space for managing translation documents.
2. The registered users can publish their translations on the MNH site, in the same manner as a community-based news site, if copyright of the original text permits.
3. A range of social networking functions is provided on MNH, including message exchange, question and answer, translation request submission, communication by noticeboard, and social tagging.
4. Terminology and translation memory management functions, including individual registration of terms, batch upload of terminologies, registration of parallel texts.

Figure 1. The MNH main page

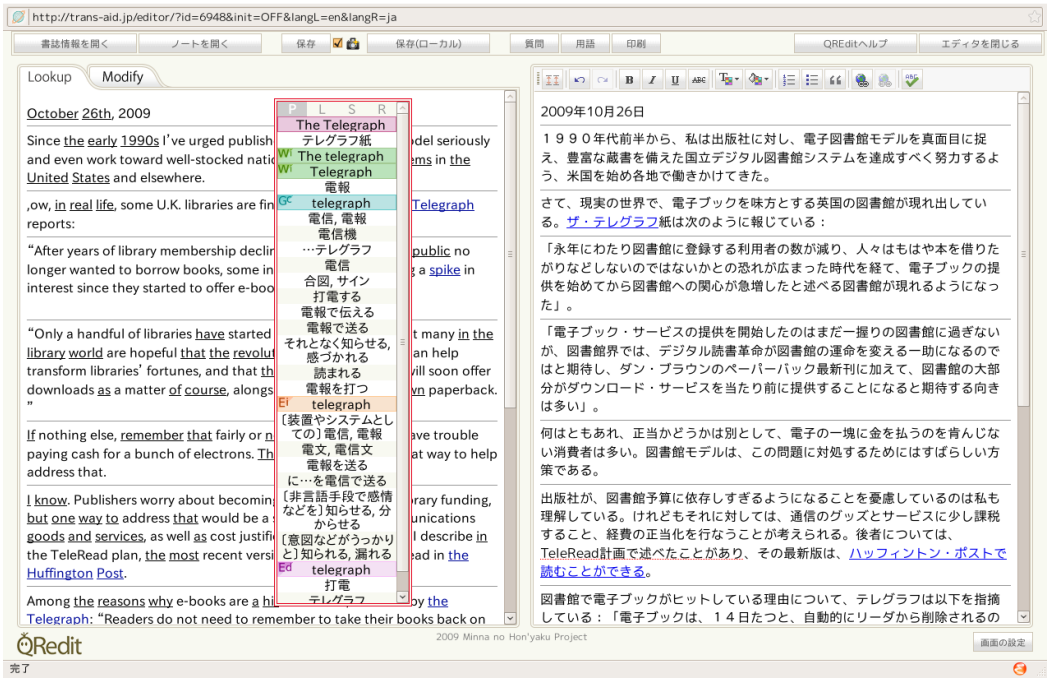

Figure 2. The integrated translation-aid editor environment QRedit

5. A series of search functions is provided, including the search of translation texts by content words or by tags, of translated text pairs (TM), of registered terms, of users, and of questions.

What makes MNH especially important and attractive for online translators is its built-in integrated translation-aid editor environment, QRedit (Abekawa and Kageura, 2007a; Abekawa and Kageura, 2007b). QRedit is a two-pane editor which provides the following functions for online translators (Figure 2):

1. flexible (idiom variations can be looked up) and stratified (important or difficult multi-word elements are highlighted) lookup and copy-and-paste functions of dictionaries and terminological resources (Takeuchi, et. al., 2007), including a high-quality English-to-Japanese dictionary widely used by translators (Sanseido, 2001) ;

2. seamless connection to Wikipedia bilingual and monolingual entries, as well as to Google search;

3. function to register term translations in the process of translation and their immediate lookup;

4. an easy-to-use and effective interface which enables users to maintain their rhythm in producing translations.

MNH thus accumulates a translation corpus naturally. At the time of writing this paper (February 7th, 2010), MNH has 1,061 registered users (of whom 42 publish their translations on MNH) and 3,301 translated documents (of which 1,720 are published). Some prominent Japanese NGOs, such as Amnesty International Japan and Democracy Now! Japan, and translators from civic translation groups such as the Japanese translation team of GlobalVoices Online and Translators United for Peace (TUP) are using MNH. In addition, several university professors use MNH for their seminar in reading and translating English news in their domains. We shall come back to the statistics in section 5, when we describe the current status of the SDF corpus on MNH.

## 3. Construction of the SDF corpus

### 3.1. Basic mechanism

The basic mechanism for accumulating draft and final translations is very simple. Translators using MNH save their translations to keep the data when they finish the translation. MNH keeps a log of up to 10 versions of translation for each document. At first when we designed the mechanism, we only had a single save mode: each time the translator saved the translation, it was saved as a new version of translation. After we made MNH publicly available and users started using the system, however, we noticed an obvious problem. Translators do not only save their translations to update the translation logs, but they often save them in the midst of translation to avoid data loss. Some users save their translations every now and then, and the first draft translations can be easily lost because the system keeps only 10 versions, while the translators keep saving each and every small step in their translation process.

In order to avoid this problem, we introduced two modes for saving translations in November 2009: ordinary mode and snapshot mode. The translation version saved in the ordinary saving mode is overwritten when the translator saves the new translation. If the translation is saved in the snapshot mode, the version is kept unless 10 more versions are newly saved in the snapshot mode. The selection of the mode is straightforward: users can simply check or uncheck "snapshot" indicated by the camera icon next to the "save" button when they save the translation in QRedit. Although users should become conscious of the saving modes, we found it preferable not only from the point of view of keeping due versions but also from the point of view of making translators use the SDF corpus for their own benefits. Each user can check their translation logs starting from the very first draft translation to the final version, the details of which will be described in section 4.

### 3.2. Community-based accumulation

MNH can collect not only draft and final translations made by a single translator, but also those made by different translators. MNH has a function that enables users to give permission for other translators registered with MNH to edit their original translations. Such permission can be open-ended, or restricted to a particular group of users. This function is of particular importance for NGOs, NPOs, university classes and other groups involved in group-based translation.

In these translation groups as typically represented by NGOs, it is a normal procedure in translation that a draft translation is produced by some inexperienced translators and then modified and finalised by experienced translators. Incidentally, the division of labour between a (draft) translator and a reviser is, or should be, common to translation activities in general, including those carried out by translation companies, as can be from the requirements given in the EU standard for translation services EN-15308 (Robert, 2008).

Thus if an inexperienced translator gives permission for editing his/her draft translations to experienced translators in the group and the experienced translators revise the translations, the translation logs – which include the draft and final versions – made by different translators are kept on the server. When the draft translator has the original ownership of the document, s/he should give final permission for publication if the translation is to be published. Thus draft translators inevitably have to at least see the document if they wish to publish the translations edited by more than one translator. The more groups that use MNH for translations, therefore, the more chances there are that draft and final translations made by different translators are accumulated on MNH as part of the SDF corpus.

The translation group management on MNH tries to strike a balance between openness and security among registered translators. Two steps are needed to activate group editing. First, each user has to register the other users who are to constitute the group and to define the group. Then, for each document which the user wants other translators in the group to edit, the user gives them editing permission, which can be done by choosing the editing status when s/he saves the translation. Some group users claim this procedure of defining a group is cumbersome; we are currently examining the optimal way of providing the function of defining a group on MNH.

## 4. Use of the SDF corpus by human translators

The corpus consisting of source texts and different versions of translations (including draft and final) can be used for self-training by inexperienced translators, as well as offering a basic corpus for improving MT and related technologies (Abekawa and Kageura, 2008b). Though the latter aspect, i.e. to incorporate human expertise in MT, has gained importance especially recently (Casacuberta, et. al., 2009), we discuss here the former aspect. To take full advantage of the SDF corpus in improving MT performance, further steps, including relevant tagging, are expected to be necessary, which we have not yet pursued fully.

The SDF corpus is especially effective for NGOs and other groups relying on volunteer translators in training inexperienced translators. Many NGOs constantly face the problem of a dearth of good and/or reliable volunteer translators. Due to the low retention rate of volunteer translators, the


Figure 3. A comparative view of the draft and final translations

core, experienced translators become busier, leaving them no time to give advice to inexperienced translators, which further reduces the retention rate – a vicious circle. Thus mechanisms to enable inexperienced volunteer translators to train themselves are urgently needed. These mechanisms should be integrated into the process of translation to which they contribute as a volunteer; otherwise they cannot maintain enough motivation.

To enable inexperienced translators to self-train and check their translations, MNH provides a comparative view mode that displays any two versions of translations kept in the SDF corpus on MNH. Thus inexperienced translators can easily check where and how their translations are modified by experienced translations. Figure 3 shows a comparative view of draft and final translations. The differences between two versions are captured by a javascript library google-diff-match-patch provided by Google (Google , 2009). In Figure 3, the parts deleted from the draft translation are indicated in green, with lines overwritten, while the parts added to the modified translation are indicated in red. Incidentally, the comparison can be made from the other side, if the user wishes.

Because MNH supports and facilitates the natural cycle of community-based translations by providing a translation-aid environment for making translations and enabling modification of translations, including by other translators, the comparative view of the SDF corpus provided by MNH also constitutes a part of the translation process. After the draft translation is modified by an experienced translator, the inexperienced translator who made the draft can see which parts have been modified and in what way, this within the same environment that s/he uses for producing the draft translations, as though the draft is checked by experienced translators for educational purposes. There are SNS sites explicitly for language learning, such as Lang-8 (Lang-8, 2009). What characterises MNH's translation training function through SDF corpus is the fact that it is a side effect of the natural cycle of group-based translations and constitutes an integrated part of the actual translation process.

## 5. Current state of the SDF corpus on MNH

### 5.1. Overall statistics

As of February 7, MNH has 1,061 users (of whom 42 users publish their translations) and 3,301 translations (of which 1,581 are published), as mentioned in section 2. Table 1 shows the bi-monthly statistics of the number of users and that of translated documents on MNH since April 7, 2009, when MNH was made public. The number of registered users as well as that of documents have steadily increased. The increase in the number of published documents is faster than the increase in all the documents, thus the ratio of published documents increases.

Table 2 shows the number of translations for the number of versions (up to 10) retained in MNH. Table 3 shows the same data for published translations. Figure 4 visualises the

| | No. & per centage of documents | | | | | |
|---|---|---|---|---|---|---|
| Versions | Apr. 09 | Jun. 09 | Aug. 09 | Oct. 09 | Dec. 09 | Feb. 10 |
| 1 | 46 (10.80) | 346 (27.75) | 611 (34.52) | 831 (37.77) | 1149 (41.66) | 1451 (43.96) |
| 2 | 227 (53.29) | 347 (27.83) | 398 (22.49) | 447 (20.32) | 527 (19.11) | 619 (18.75) |
| 3 | 95 (22.30) | 222 (17.80) | 248 (14.01) | 277 (12.59) | 312 (11.31) | 359 (10.88) |
| 4 | 25 (5.87) | 86 (6.90) | 100 (5.65) | 118 (5.36) | 135 (4.89) | 152 (4.60) |
| 5 | 8 (1.88) | 57 (4.57) | 72 (4.07) | 79 (3.59) | 88 (3.19) | 100 (3.03) |
| 6 | 9 (2.11) | 37 (2.97) | 52 (2.94) | 57 (2.59) | 66 (2.39) | 73 (2.21) |
| 7 | 4 (0.94) | 24 (1.92) | 41 (2.32) | 50 (2.27) | 56 (2.03) | 60 (1.82) |
| 8 | 2 (0.47) | 14 (1.12) | 23 (1.30) | 28 (1.27) | 32 (1.16) | 34 (1.03) |
| 9 | 2 (0.47) | 15 (1.20) | 24 (1.36) | 29 (1.32) | 34 (1.23) | 39 (1.18) |
| 10 | 8 (1.88) | 99 (7.94) | 201 (11.36) | 284 (12.91) | 359 (13.02) | 414 (12.54) |
| Multiple | 380 (89.20) | 901 (72.25) | 1159 (65.48) | 1369 (62.23) | 1609 (58.34) | 1850 (56.04) |
| Total | 426 | 1247 | 1770 | 2200 | 2758 | 3301 |

Table 2. The number of documents by the number of versions (all)

| | No. & per centage of documents | | | | | |
|---|---|---|---|---|---|---|
| Versions | Apr. 09 | Jun. 09 | Aug. 09 | Oct. 09 | Dec. 09 | Feb. 10 |
| 1 | 32 (33.68) | 214 (48.31) | 371 (52.11) | 471 (50.70) | 656 (52.31) | 817 (51.68) |
| 2 | 29 (30.53) | 81 (18.28) | 110 (15.45) | 149 (16.04) | 214 (17.07) | 282 (17.84) |
| 3 | 20 (21.05) | 71 (16.03) | 89 (12.50) | 108 (11.63) | 133 (10.61) | 170 (10.75) |
| 4 | 5 (5.26) | 23 (5.19) | 32 (4.49) | 44 (4.74) | 53 (4.22) | 67 (4.24) |
| 5 | 2 (2.11) | 11 (2.48) | 21 (2.95) | 25 (2.69) | 30 (2.39) | 40 (2.53) |
| 6 | 2 (2.11) | 10 (2.26) | 15 (2.11) | 17 (1.83) | 19 (1.51) | 23 (1.45) |
| 7 | 0 (0.00) | 3 (0.68) | 8 (1.12) | 12 (1.29) | 16 (1.28) | 18 (1.14) |
| 8 | 0 (0.00) | 2 (0.45) | 6 (0.84) | 8 (0.86) | 10 (0.80) | 12 (0.76) |
| 9 | 0 (0.00) | 3 (0.68) | 7 (0.98) | 8 (0.86) | 10 (0.80) | 14 (0.89) |
| 10 | 5 (5.26) | 25 (5.64) | 53 (7.44) | 87 (9.36) | 113 (9.01) | 138 (8.73) |
| Multiple | 63 (66.32) | 229 (51.69) | 341 (47.89) | 458 (49.30) | 598 (47.69) | 764 (48.32) |
| Total | 95 | 443 | 712 | 929 | 1254 | 1581 |

Table 3. The number of documents by the number of versions (published)

| Date | Users | All docs | Published docs (%) |
|---|---|---|---|
| Apr. 7, 2009 | 49 | 426 | 95 (22.30) |
| Jun. 7, 2009 | 545 | 1247 | 443 (35.53) |
| Aug. 7, 2009 | 666 | 1770 | 712 (40.23) |
| Oct. 7, 2009 | 867 | 2200 | 929 (42.23) |
| Dec. 7, 2009 | 978 | 2758 | 1254 (45.47) |
| Feb. 7, 2010 | 1061 | 3301 | 1581 (47.90) |

Table 1. The growth in the number of documents on MNH

statistics given in Tables 2 and 3 by barplots. We can regard the number of documents that have more than two versions as constituting the SDF corpus – or perhaps better still – the "raw" SDF corpus.

From Tables 2 and 3 and Figure 4, we can observe the following tendencies in the development of the SDF corpus:

1. As can be seen from the line "Multiple" in Table 2 as well as from the top left panel of Figure 4, the number of translations with more than two versions is steadily increasing.

2. As can be seen from the line "Multiple" in Table 3 as well as from the top right panel of Figure 4, the number of published translations – which can be more safely regarded as "complete" translations – with more than two versions is also steadily increasing.

3. The ratio of translations with more than two versions is in decrease over time, and it seems to be still decreasing within the time span of the data.

4. The ratio of published translations with more than two versions decreased from April to August 2009, then seems to have converged to slightly below 50 per cent.

5. The ratio of published translations with more than two versions is lower than that of all the translations with more than two versions. This in fact is contrary to our initial expectation, because we expected that the published translations would be reviewed carefully. One reason for this may be that those who publish translations are relatively experienced, and they aim at distributing information in a timely manner rather than making sophisticated translations. This needs further investigation.

6. There are a good number of translations with 10 versions, the ratio of which is comparable to that of translations with three versions. However, the ratio is recently in decline, most probably because we introduced two saving mode in November 2009. That the ratio of translations with 10 versions in published translations is consistently lower than that of all the translations may reflect the fact that translators are

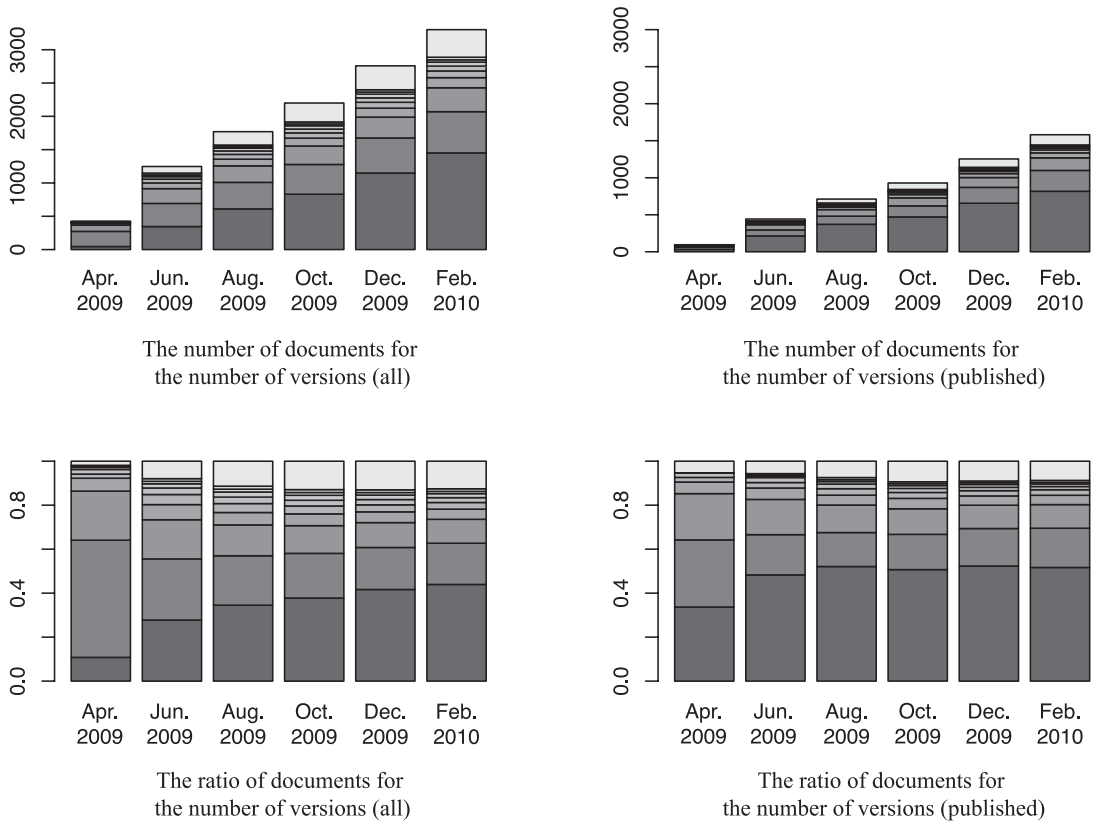


Figure 4. The number and ratio of documents for the number of versions

more conscious of the translation versions for published translations. This, again, needs further study.

### 5.2. Documents edited by multiple translators

Table 4 shows the status of users who are involved in group translations and documents. Percentages are calculated for the corresponding figures given in Table 1. We witness a general growth in the number of users involved in group translation and the number of documents edited by more than one translator. The ratio is also on the increase in general.

Table 5 shows the statistics for the documents on MNH which are edited by more than one translator. Table 6 shows the statistics for the published documents on MNH which are edited by more than one translator. Note that the number of documents for one log is by definition zero, so it is omitted from the Tables. Figure 5 visualises the information in Tables 5 and 6. Although the number of documents edited by multiple users is still small, it is steadily increasing.

The following characteristics can be pointed out.

1. The number of documents translated by more than one translator is steadily increasing, although the pace of the increase is not stable.

2. The number of published documents with two versions is zero, reflecting the mechanism given by MNH in which the document owner, most probably the draft translator, needs to give permission for publication to the document at the final stage (see 3.2).

| | | Documents | |
|---|---|---|---|
| Date | Users (%) | All (%) | Published (%) |
| Apr. 7, 2009 | 2 (4.08) | 1 (0.23) | 1 (1.05) |
| Jun. 7, 2009 | 11 (2.02) | 11 (0.88) | 3 (0.68) |
| Aug. 7, 2009 | 24 (3.60) | 35 (1.98) | 11 (1.54) |
| Oct. 7, 2009 | 26 (3.00) | 54 (2.45) | 14 (1.51) |
| Dec. 7, 2009 | 52 (5.32) | 94 (3.41) | 38 (3.03) |
| Feb. 7, 2010 | 55 (5.18) | 110 (3.33) | 48 (3.04) |

Table 4. The growth in the number of documents edited by more than one user on MNH

3. Compared to all the documents shown in Tables 2 and 3 and Figure 4, the documents with more than four versions occupy a much higher ratio, which is natural.

4. The number and ratio of documents with 10 versions is particularly noticeable, constituting more than half at some points in time. The ratio is in decrease since we introduced the snapshot saving mode.

5. We do not observe a clear tendency for the published documents to be less edited than all the documents.

### 5.3. Issues

Although the number of documents with more than one translation version is increasing steadily, including those edited by more than one user, further clarifications are necessary to take full advantage of the corpus thus accumulated. We have not yet carried out in depth investigations of the quality of the translations and the nature of modifications made in each log for each document. In relation to

| | No. & per centage of documents | | | | | |
|---|---|---|---|---|---|---|
| Versions | Apr. 09 | Jun. 09 | Aug. 09 | Oct. 09 | Dec. 09 | Feb. 10 |
| 2 | 0 (0) | 2 (18.2) | 2 (5.7) | 4 (7.4) | 4 (4.3) | 5 (4.5) |
| 3 | 1 (100) | 1 (9.1) | 1 (2.9) | 3 (5.6) | 7 (7.4) | 12 (10.9) |
| 4 | 0 (0) | 1 (9.1) | 2 (5.7) | 5 (9.3) | 9 (9.6) | 9 (8.2) |
| 5 | 0 (0) | 1 (9.1) | 1 (2.8) | 2 (3.7) | 4 (4.3) | 6 (5.5) |
| 6 | 0 (0) | 2 (18.2) | 6 (17.1) | 8 (14.8) | 11 (11.7) | 15 (13.6) |
| 7 | 0 (0) | 1 (8.1) | 1 (2.8) | 1 (1.9) | 3 (3.2) | 3 (2.7) |
| 8 | 0 (0) | 0 (0.0) | 3 (8.6) | 3 (5.6) | 4 (4.3) | 4 (3.6) |
| 9 | 0 (0) | 0 (0.0) | 1 (2.8) | 1 (1.9) | 4 (4.3) | 5 (4.5) |
| 10 | 0 (0) | 3 (27.3) | 18 (51.4) | 27 (50.0) | 48 (51.1) | 51 (46.4) |
| Total | 1 | 11 | 35 | 54 | 94 | 110 |

Table 5. The number of documents by the number of versions edited by multiple users (all)

| | No. & per centage of documents | | | | | |
|---|---|---|---|---|---|---|
| Versions | Apr. 09 | Jun. 09 | Aug. 09 | Oct. 09 | Dec. 09 | Feb. 10 |
| 2 | 0 (0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) |
| 3 | 1 (100) | 1 (33.3) | 1 (9.1) | 1 (7.1) | 2 (5.3) | 7 (14.6) |
| 4 | 0 (0) | 1 (33.3) | 2 (18.2) | 2 (14.3) | 4 (10.5) | 4 (8.3) |
| 5 | 0 (0) | 0 (0.0) | 0 (0.0) | 1 (7.1) | 2 (5.3) | 4 (8.3) |
| 6 | 0 (0) | 0 (0.0) | 1 (9.1) | 1 (7.1) | 3 (7.9) | 5 (10.4) |
| 7 | 0 (0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 2 (5.3) | 2 (4.2) |
| 8 | 0 (0) | 0 (0.0) | 1 (9.1) | 1 (7.1) | 2 (5.3) | 2 (4.2) |
| 9 | 0 (0) | 0 (0.0) | 1 (9.1) | 1 (7.1) | 2 (5.3) | 3 (6.3) |
| 10 | 0 (0) | 1 (33.3) | 5 (45.5) | 7 (50.0) | 21 (55.3) | 21 (43.8) |
| Total | 1 | 3 | 11 | 14 | 38 | 48 |

Table 6. The number of documents by the number of versions edited by multiple users (published)

applications for which SDF corpora are expected to be useful, the following points, among others, should be clarified or promoted.

First, for use in improving MT or in identifying patterns and establishing general rules of translation revision, we can potentially use all the accumulated SDF data. In order for that, however, detailed examinations of the nature of modifications are necessary. Though we are pursuing the research in this direction (Abekawa and Kageura, 2008a; Abekawa and Kageura, 2008b), we have not yet produced substantial results.

Second, although the corpus accumulated on MNH is currently used by human translators registered with MNH, the number of users who take advantage of this function is still relatively small. In terms of the number of documents, almost half are not revised at all, which indicates that many users probably do not regard revision as important or do not regard learning from the SDF corpus as something they can derive benefit from. The relatively small number of documents with multiple versions and users who take part in group editing may imply that users or potential users are not aware of the potential benefits they can obtain from the SDF corpus and the comparative view. This can be overcome in two respects, i.e. by helping potential users understand the functions and by improving the usability of the system, both of which remain to be done.

## 6. Conclusions

We have introduced a way of constructing a translation corpus (SDF corpus) that contains source texts, their draft translations, and their final translations, using the translation hosting site Minna no Hon'yaku (MNH), which is publicly available. The environment that MNH offers to translators provides them with a self-contained cycle of group-based translations, which includes making draft translations, modifying and finalising them, and giving feedback to inexperienced translators who produce draft translations. Thus the SDF corpus can be immediately useful for translators using MNH, which promotes the natural accumulation of SDF data in its turn.

MNH was made publicly available in April 2009 and the functions described in this paper is fully operational and are all provided to actual users. The statistics show that the size of the SDF corpus is increasing steadily on MNH, but also reveal some issues which should be addressed for the SDF corpus to be fully useful. We are now in the process of dealing with these problems.

At the time of writing, MNH only supports Japanese-to-English and English-to-Japanese translation, due to the limited availability of high-quality dictionaries. We have a development plan to provide English-to-Chinese, Chinese-to-English, Japanese-to-Chinese and Chinese-to-Japanese translation aid functions within 2010. Also, the interface of MNH is currently only in Japanese. The English interface is to be made public by April 2010.

## 7. Acknowledgements

This work is supported by the Japan Society for the Promotion of Sciences (JSPS) grant-in-aid (A) 21240021 "Developing an integrated translation-aid site which provides comprehensive reference sources for translators".

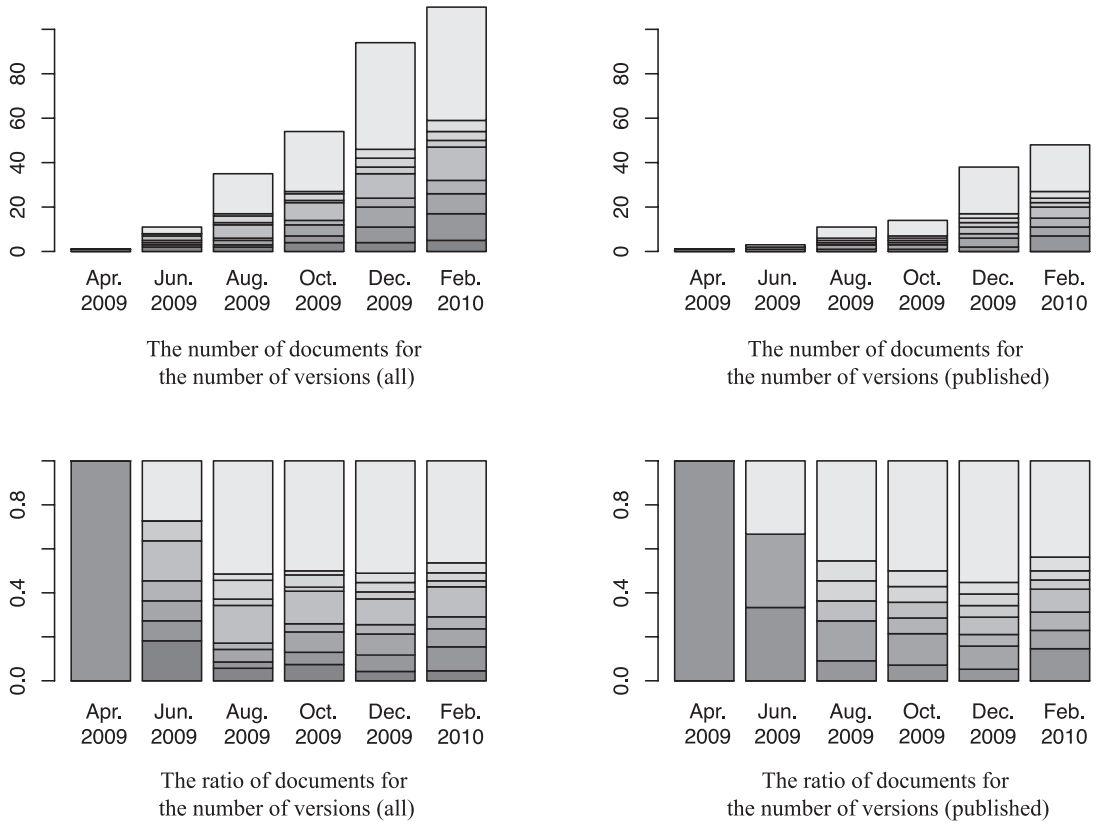


Figure 5. The number and ratio of documents for the number of versions

## 8. References


Abekawa, T. and Kageura, K. 2007a. QRedit: An integrated editor system to support online volunteer translators. In *Digital Humanities 2007*, p. 5–8.

Abekawa, T. and Kageura, K. 2007b. A translation aid system with a stratified lookup interface. In *Proceedings of ACL 2007 Demos and Poster Sessions*, p. 5–8.

Abekawa, T. and Kageura, K. 2008a. Constructing a corpus that indicates patterns of modification between draft and final translations by human translators. In *LREC 2008*.

Abekawa, T. and Kageura, K. 2008b. What prompts translators to modify draft translations? An analysis of basic modification patterns for use in the automatic notification of awkwardly translated text. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, p. 241–248.

Barzilay, R. and McKeown, K. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of ACL2001*, p. 50–57.

Casacuberta, F., Civera, J., Cubel, E., Lagarda, A. L., Lapalme, G., Macklovitch, E. and Vidal, E. 2009. Human interaction for high-quality machine translation. In *Communications of the ACM*, 52(10), p. 135–138.

Google 2009. google-diff-match-patch. `http://code.google.com/p/google-diff-match-patch/`

Lang-8 2009. `http://lang-8.com/`

MeLLANGE 2009. `http://corpus.leeds.ac.uk/mellange/ltc.html`

Robert, I. 2008. Translation revision procedures: An explorative study. In Boulogne, P. (ed.) *Translation and Its Others. Selected Papers of the CETRA Research Seminar on Translation Studies*. `http://www.kuleuven.be/cetra/papers/papers.html`

Sanseido 2001. *Sanseido Grand Concise English-Japanese Dictionary*. Tokyo: Sanseido.

Takeuchi, K., Kanehira, T., Hirao, K, Abekawa, T. and Kageura, K. 2007. Flexible automatic look-up of English idiom entries in dictionaries, In *MT Summit XI*, p. 451–458.

Tono, Y. 2009. The JEFLL Corpus Project. `http://jefll.corpuscobo.net/`

Utiyama, M. Abekawa, T., Sumita, E. and Kageura, K. 2009. Hosting volunteer translators. In *MT Summit XII*.