Constructing a corpus that indicates patterns of modification between draft and final translations by human translators

Takeshi Abekawa and Kyo Kageura

Library and Information Science Course
Graduate School of Education, University of Tokyo
{abekawa,kyo}@p.u-tokyo.ac.jp

Abstract

In human translation, translators first make draft translations and then modify and edit them. In the case of experienced translators, this process involves the use of wide-ranging expert knowledge, which has mostly remained implicit so far. Describing the difference between draft and final translations, therefore, should contribute to making this knowledge explicit. If we could clarify the expert knowledge of translators, hopefully in a computationally tractable way, we would be able to contribute to the automatic notification of awkward translations to assist inexperienced translators, improving the quality of MT output, etc. Against this backdrop, we have started constructing a corpus that indicates patterns of modification between draft and final translations made by human translators. This paper reports on our progress to date.

1. Introduction

In accordance with the rapid growth of information available on the Internet in an increasing numbers of languages, overcoming language barriers has become a keen concern all over the world (Cronin, 2003). Accordingly, the activity of volunteer translators is gaining in prominence. For instance, in an English-to-Japanese context, Translators United for Peace (TUP), which consists of about a 20 volunteer translators, won the Civic Media Award of the Japan Congress of Journalists in 2004 for its role in distributing important information online about the US attack on Iraq (TUP, 2008). In a more multilingual setting, PaxHumana provides news and information related to peace and war in French, German, Spanish and English (PaxHumana, 2008). Global Voices, an originally Harvard-based project, relies on volunteer translators to multilingualise blog texts written in a variety of languages (Global Voices, 2008). Many projects like these exist all over the world. In a different arena, localisation of open source software such as Mozilla or Wiki is also supported by volunteers.

Within this context, we are currently developing a translation aid system aimed at assisting volunteer translators working online, focusing on English-to-Japanese translation (Abekawa and Kageura, 2007a; Abekawa and Kageura, 2007b). One of the important requirements for such a system is to incorporate a mechanism that helps inexperienced translators put together high-quality translations. In addition, there is a necessity to attract more people to translation work, and the number of inexperienced translators working voluntarily is expected to grow. We have consulted with some 20 volunteer translators, and found that about half are professional or highly experienced translators devoting their free time to voluntary translation, while the other half have little experience in translation. These inexperienced translators have a sufficient command of the source (English) and target (Japanese) language, but lack specialized knowledge of translation. Given this situation, a crucial factor in realising a system that helps inexperienced translators is to clarify the gap between inexperienced and experienced translators and to describe the knowledge of experienced translators.

Though there are translation textbooks and practical guide books describing how to make good translations (Baker, 1992; Anzai, 1995; Kawamoto and Inoue, 1997), the descriptions in those books, though very useful as a guideline, assume a substantial amount of human knowledge and are not formalised in such a way that this knowledge can be transformed straightforwardly into a computationally tractable descriptions. For our ultimate aim of developing a system that notifies inexperienced translators of awkward translations, further information is required . Fortunately, in the process of translation, translators first make draft translations and then examine and edit them, often repeatedly. Thus there are normally at least two versions of the translation of a given text, i.e. the draft and the final translations. In commercial translation environments, it is sometimes the case that texts are first translated by inexperienced translators and then edited by experienced translators. We have thus obtained translation data (a triplet consisting of original English texts, draft Japanese translations and final Japanese translations) for several books from commercial publishers and are currently constructing a corpus that indicates patterns of modification between draft and final translations (the POMDAF corpus), in consultation with translators. As there is no corpus of this type as far as we know, and as the implicit knowledge of experienced translators has not been clarified so far, the first stage of corpus construction has been devoted to defining the basic types of information to be included in the corpus, which involves the clarifying translators' implicit knowledge. The work reported in this paper is work from this stage. We believe that the corpus, when completed, will provide the translation research community with a corpus that indicates patterns of modification by highly skilled human translators, and will constitute an important language resource for exploiting the implicit knowledge of translators.

In section 2, we introduce the basic data we have collected, on the basis of which the patterns of modification are to

be defined. In section 3, we describe the basic framework within which modification patterns are recognised and described. In section 4, we report the actual modification patterns we defined using the data. Section 5 is devoted to a discussion of the theoretical status of the work, which has become clear through our consultations with translators during the process of corpus construction.

2. The data

The data for the POMDAF corpus consists of triplets of the English original, draft Japanese translation and final Japanese translation of books and online articles, which were provided by two publishers and several translators. We currently have this three-part data for seven books (consisting in total of about 40,000 sentences) and six online articles (consisting of about 800 sentences). Among these, we selected three books (Leggett, 2005; Chomsky, 2004; Harvey, 2005) for the initial stage of the corpus construction. These books have a common feature, i.e. they were first translated by less-experienced translators and then checked and corrected by experienced translators. Table 1 shows the size of each book in terms of the number of sentences, for the English original , the draft Japanese translation and the final Japanese translation.

Table 1: Size of the three books

	Leggett	Chomsky	Harvey
Original	4,515	1,615	3,072
Draft	4,622	2,374	3,121
Final	4,644	2,468	3,155

We extracted 50 sentences (based on the final Japanese) from the data, as the first step in analysing patterns of modification between the draft and final translations. The amount of data currently analysed is very small, because it is critical to carry out an in-depth analysis of modification patterns in order to properly reflect translators' implicit and explicit knowledge in the corpus. In the analysis, we enlisted the help of a linguist and a translator, and examined the nature of modifications observed in the data.

3. The basic framework

Sentences are used as the basic unit of analysis. There are a small number of cases in which there is not an exact one-to-one correspondence between the sentences in the three texts (for example, a single English sentence being translated into two sentences in the Japanese draft and/or final). In these cases, we take the longest sentence among the triplet as the basic unit. Though most experienced translators judge the smoothness or awkwardness of draft translations in the context of paragraphs or larger discoursal units, the actual unit whose smoothness or awkwardness is judged is the sentence. In other words, expert translators modify draft translations basically sentence by sentence, while judging their smoothness or awkwardness within the context of a paragraph or several paragraphs. There can be more than one units of modification in one sentence.

An interesting and important point related to the data we are dealing with is its duality: while the data can be considered a "gold standard" in the sense that it consists of translations that have already been published and well accepted, it is not a gold standard in the sense of being on "average" that all translators aim to achieve, because each translation is unique and singular and could have turned out rather differently if it had been put together by a different translator. This point can be easily appreciated of one considers the fact that there are sometimes more than one high-quality translation for a source text, especially in the area of literary translations. We delve deeper into this topic in section 5.

After consulting the translator and the linguist, we distinguished the following four levels at which modification patterns are identified and described.

1. *Reasons for modification*: The reason that translators have in mind when making a modification. At the broadest level, we identified six reasons, which will be elaborated on in the next section. In addition, we asked the translator and the linguist to freely describe the reasons for the modification. This level corresponds to the judgement of draft translations by translators. Thus the reasons for modification are essentially attributed to the draft, conditioned by the English original. We identified this level as separate from the next level for two reasons: (i) though there is a logical relation between the judgment of the draft and the direction for modifications, the relation is not necessarily one-to-one or fixed; (ii) translators sometimes use different words for describing the reasons for modifications and the aims of linguistic operation, which may reflect different levels of implicit knowledge.
2. *Aims of linguistic operations*: This level was introduced to connect the reasons for modification and the description of linguistic operations applied to make the final translation. As such, the aims of linguistic operations are essentially an attribute of the relation between the draft translation and the final translation. About twenty categories were introduced, which will be elaborated on in the next section. This level is of critical importance in clarifying the implicit knowledge translators make use of when modifying the draft and making good translations. More than one aim of linguistic operation can correspond to one reason for modification and vice versa, both typewise and tokenwise.
3. *Linguistic operations*: This level is defined by means of linguistic or grammatical terms, such as "change of tense", etc. More than one linguistic operation may constitute an aim of linguistic operations. We tried to define linguistic operations in a sufficiently formal manner for two reasons: (i) to distinguish the interdependent or superimposed operations while at the same time identifying their relationships, and (ii) to enable further breakdown of modification phenomena to the surface primitive operations.
4. *Primitive operations*: This level is defined by means of surface terms of operation, such as "insertion of postposition ('no')" etc. The aim of setting this level

is twofold: (i) to define operation patterns in as detailed and formal a manner as possible so that they can be defined in a computationally tractable manner, and (ii) to enable human and computational analysis of the relevant features at work in the human modification process. More than one primitive operation may constitute a linguistic operation.

In addition to these, we are planning to add information concerning the necessity of modifications, because some translators see that some modifications are not really necessary. The three levels of necessity are "obligatory," "preferable," and "optional".

4. Patterns of modification

This section elaborates the four levels of information we are currently assigning to the corpus.

4.1. Reasons for modification

In the modification process, the translator first recognises one of a number of *states* in a draft translation, which may or may not trigger modifications. This is often carried out unconsciously. When the draft translation is modified, we can observe the *reasons for modification* that correspond to the states. We take the unit in which a single reason for modification is identified by a translator as the basic *unit of modification*. So in our corpus construction process, one reason for modification correspond to one unit of modification. At the broadest level, we classified the reasons for modification into six categories. These are shown in Table 2 shows them. Although these reasons are conceptually clear and thus can be used as a guide for further analysis of the data, it is not necessarily the case that translators can judge the reasons for a particular modification clearly and consistently, because judging a sentence as being "natural" or "confusing", for instance, is not a binary process but a graded one, and the distinction among different reasons is often not immediately clear in actual translations.

Table 2: Reasons for modification

1. Mistranslation
2. Translation is confusing and the original meaning is not clearly expressed.
3. Translation is unnatural or awkward.
4. Against the overall style and guidelines
5. Against the taste of the translator who modifies the draft
6. Others

All in all we identified 181 reasons for modification. These correspond to the basic unit of modification. For each basic unit of modification, more than one linguistic operation is applied.

4.2. Aims of linguistic operations

As was discussed in the previous section, while the reasons for modification are defined on the basis of the draft translation from the translators' point of view, the aims of linguistic operations are defined on the basis of linguistic operations applied to the draft to make the final. Table 3 shows these aims. Note that some of the aims listed in Table 3 are not mutually exclusive. As we are in the preliminary and experimental stage of constructing the POMDAF corpus, we deliberately left room for these ambiguities, because these are the expressions that most straightforwardly come out of translators looking at the data.

Table 3: Aims of linguistic operations

1. To make the expressions more fluent
2. To use more suitable expressions
3. To reduce the complexity of a sentence
4. Orthographic changes
5. To add information (content words)
6. To change the focus of topicalisation
7. To make expressions less complicated
8. To add or delete subjective expressions
9. To delete redundant expressions (content words)
10. To change the temporal relation between the speaker and the event that is talked about (tense, aspect)
11. To clarify the relation between two elements
12. To avoid repetition of the same element
13. To replace an anaphoric expression with a concrete expression
14. To add or delete zero pronouns
15. Balancing phrases
16. To make expressions more formal
17. Balancing clauses
18. Removing ambiguous structures
19. Others
20. None (in the case of mistranslation)

4.3. Linguistic operations

At the level of linguistic expression, the most natural way to classify modification patterns is by means of basic linguistic labels such as "change of voice" or "change from nominal modification to adverbial modification". In fact, many translation textbooks give translation tips by explaining reasons for modification, aims of linguistic operations and linguistic operations at the same time. Table 4 shows examples of linguistic operations. From the point of view of NLP, these modification patterns consist of one or more *basic operations*. For instance, a "change of voice" may consist of such primitive operations as "changing the case-marker of the subject," "swapping the position of subject and object," etc. We therefore defined linguistic operations hierarchically, in accordance with the hierarchy of the linguistic units involved.

4.4. Primitive operations

The fourth and final level of information is the description of the surface change for each *basic operation* defined in 4.3. We introduced four primitive operations of "insertion", "deletion", "replacement" and "transposition". The actual surface modification phenomenon is linked with one or more of the four primitive operations. Primitive operations are then linked to linguistic operations, to which the aim of operations is attached. Through the aim of linguistic operations, we can establish a thread of description from

Table 4: Example of linguistic operations

Unit	Linguistic operation
Voice	From passive to active voice
Modality	From supposition to concession
Punctuation	Emphasis of a parallel structure
Particle	Deletion of a case particle equivalent
Verb	Deletion of a redundant verb
Compound noun	From noun phrase to noun clause
Verb clause	From the end of the sentence to the beginning

the reasons for modification (translators' judgement) to the actual concrete modifications (target of linguistic processing). Table 5 shows a small part of the corpus, to which all the four levels of information are assigned.

5. Theoretical position of the work

In the process of corpus construction, some important and interesting points have become clearer in relation to the nature of translation. Though they still remain abstract and are not necessarily reflected in the corpus in a concrete manner, we summarise what we have found to be important points here.

5.1. The nature of translation activity

One of the most important points we found is that translation as perceived by experienced translators has little to do with the language that linguists and computational linguists see (Kageura, 2006). Translation is concerned first and foremost with individual texts. This fact introduces a historical dimension into the translation. If we adopt the distinction between the study of language and the study of *enoncé* as postulated by Foucault (1968),

> La question que pose l'analyse de la langue, à propos d'un fait de discours quelconque, est toujours: selon quelles règles tel énoncé a-t-il été construit, et par conséquent selon quelles régles d'autres énoncés semblables pourrairent-ils être construits? La description du discours pose une tout autre question: comment se fait-il que tel énoncé soit apparu et nul autre à sa place?

Translation is first and foremost concerned with énoncé as is dealt with in the study of énoncé, not the study of language. We can define a chain of concepts; "texts" – "énoncé" – "archive" – "monument". This is distinguished from the chain of concepts in linguistics, i.e. "texts" – "language" – "corpus" – "example". The task of translators is to give a position to the translated text within a given set of énoncé, which constitutes an archive in the target language, a position that is equivalent to the position of the original text, which is perceived as a singular monument within the archive of all other historical monuments to which the text is related. Put differently, translation is concerned with the range of *socially acceptable or preferable expressions* (i.e. those that are realistically possible) while linguistics is concerned with the range of well-formed utterances. Within this framework, linguistic operations are necessary as a prerequisite for but not the core of translation.

As both the source text and its translation are singular and occur only once in history, the translation calls for individual decision making. This is most typically observed in literary translations. For instance, Suzuki (2006), who translated Proust's *Á la recherche du temps perdu* into Japanese, depicted translation as a process of constructing language, trying to arrive at the place that the author aimed at in the original. For this type of task, each step involves making a new decision (cf. Munday, 2001; Venuti, 2004).

To end the discussion here, however, world be a grave mistake, because at the other extreme of translation, i.e. professional translation of highly technical documents, we find substantial convergence among translations made by different translators or translations made at different times by the same translator, which proves that the translation process can be, to a substantial extent, reduced to a standard procedure or computation in its broader sense. It should be emphasised again here that, for translators, this procedure is defined not as a linguistic operation but as a textual operation of assigning a position to the translation that is isomorphic to the position of the original. Figure 1, taken from Kageura (2007), illustrates this.

We therefore have at least the following three operational spheres in translation activity:

1. The linguistic sphere, as dealt with by linguists and computational linguists. Regarding textual data as corpora representing some aspects of language phenomena is included here.

2. The textual sphere, in which texts are perceived as historical products. Regarding texts as monuments and language expressions as an archive is included here.

3. The decision making sphere, which is dependent on individual translators who are singular and unique.

Within these three spheres, what we are trying to explore is the nature of operations in the textual sphere as reflected in the linguistic sphere, because after all what are immediately available to us are the given translation triplets, and not the archive within which the texts are situated. To what extent we can explore information inherently belonging to the textual sphere at the level of a given text depends on the extent to which textual sphere operations are projected on to phenomena in the linguistic sphere, which can only be revealed through the construction of this type of corpus.

5.2. Agreement of judgements among translators

The aforementioned nature of translation activity and the nature of the task we face in the construction of POMDAF impose certain peculiarities upon the nature of the corpus construction. It is commonly held that the inter-annotator agreement is of utmost importance in corpus construction in general. This is supported by the not-so-unnatural assumption about language that the same language spoken by different people is basically the same – an essentially Quetlet-isque assumption of the "average person".

Table 5: Example of the POMDAF corpus

English: They are, to some degree, whistleblowers.
Draft : かれらは多少なりとも警報者なのである。
Final : かれらは、多かれ少なかれ、内部告発者なのである。

Modification 1
かれらは <1><\/1> 多少なりとも警報者なのである。
かれらは <1>、<\/1> 多かれ少なかれ、内部告発者なのである。

Reason	3. Translation is unnatural or awkward.
Aim of linguistic operation	To make the expression more fluent
Linguistic operation	<1>Punctuation : The emphasis of topicalisation
Primitive operation	<1>Insertion : “、”

Modification 2
かれらは <1> 多少なりとも <\/1><2><\/2> 警報者なのである。
かれらは、<1> 多かれ少なかれ <\/1><2>、<\/2> 内部告発者なのである。

Reason	3. Translation is unnatural or awkward.
Aim of linguistic operation	To make the expressions more fluent
Linguistic operations	<1>Adverb phrase : Decomposition
	<2>Punctuation : Inserted after adverbial form.
Primitive operations	<1>Replacement : “多少なりとも”(to some degree) → “多かれ”(much) + “少なかれ”(less)
	<2>Insertion : “、”

Modification 3
かれらは多少なりとも <1> 警報者 <\/1> なのである。
かれらは、多かれ少なかれ、<1> 内部告発者 <\/1> なのである。

Reason	1. Mistranslation
Aim of linguistic operation	Modification of mistranslation
Linguistic operation	<1>Noun phrase: Mistranslation
Primitive operation	<1>Replacement : “警報者”(admonitor) → “内部告発者”(whistleblower)

The problem with dealing with the translation process is that this is often not the case. Just like in research, the “average” product of the “average person” is not necessarily highly praised. Therefore the greater ability the translator has, the more individualistic the translation becomes; and the more individualistic the translation is, the higher it is valued. The corollary: disagreement among translators or annotators may not necessarily be a procedural problem but an essential aspect of dealing with the translation process.

Although we have so far analysed only a small amount of data, we have a rough idea that translators may agree to a substantial extent about the reasons for modifications or the problems in draft translations, because many of these are identified at the linguistic level. It seems more difficult for translators to reach the same conclusions about how they modified the draft. Though the corpus is given and thus for each instance the modification is fixed, similar instances may undergo different sorts of modifications, and the same modification may be judged by translators as having different aims. This is related to the fact that modifications as well as judgments about modifications are made vis-à-vis the textual sphere, which assumes an existing relevant set of documents that is in translators' heads but is external to the corpus. We optimistically assume that the disagreements among different translators and differences in judgement about the same part of a text made at different times by the same translator will not become too far-reaching, because if the judgment is done vis-à-vis the textual sphere, which can be huge and varied but is by definition finite, it is not unreasonable to assume that the judgment process in the textual sphere can be categorised to a substantial extent, which will limit the range of judgements made about given texts. With this in mind, we consciously try not to exclude different judgement by translators as something problematic at the moment.

6. Conclusions

Currently, we have just finished tagging 50 sentences extracted from the data described in section 2. The amount of tagged data is very small as it took us a long time (more than a two-person month) to finalise the basic framework for descriptions. This is because this process at the same time involves the clarification of translators' implicit knowledge, which has not been clarified so far. We have to date focused more on the clarification of translators' implicit knowledge rather than on the formalisation of the information and the tagset. What we have reported in this paper is the first stage of the POMDAF corpus construction. We are currently planning the corpus construction as a three year project.
At the moment, we are working on the POMDAF corpus in

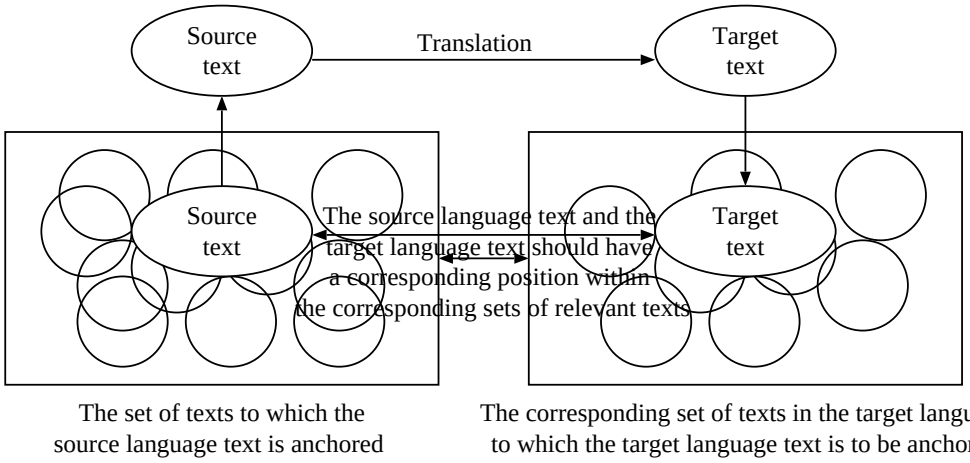

Figure 1: The framework of translation

three interrelated areas:

1. Further clarifying and elaborating on translators' implicit knowledge through construction of the corpus;

2. Increasing the amount of basic data and the size of the tagged corpus, while at the same time examining the validity of the general framework of the information we established for tagging;

3. Formalising the tagset and the format of the corpus.

In parallel with the construction of the POMDAF corpus, we are developing an experimental system that notifies inexperienced translators of awkward translations (Abekawa and Kageura, 2008).

Acknowledgements

This research is supported by a Hakuhodo "Language and Culture/Education" research grant. It is also partly supported by grant-in-aid (A) 17200018 "Construction of online multilingual reference tools for aiding translators" for the Japan Society for the Promotion of Sciences (JSPS). We would like to thank anonymous reviewers for their useful comments.

7. References

Abekawa, T. and Kageura, K. 2007a. QRedit: An integrated editor system to support online volunteer translators. In *Digital Humanities 2007*, p. 3–5.

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. 2008. 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.

Anzai, T. 1995. *Eibun Hon'yaku Jutu* (in Japanese). Tokyo: Chikuma.

Baker, M. 1992. *In Other Words: A Coursebook on Translation*. London: Routledge.

Chomsky, N. 2004. *Chomsky on Miseducation*. Lanham, MD: Rowman & Littlefield Pub Inc. [Terashima, T. et. al. trans. 2006. Tokyo: Akashi Syoten.]

Cronin, M. 2003. *Translation and Globalization*. London: Routledge.

Foucault, M. 1968. Sur l'archéologie des sciences: Réponse au cercle d'épistémologie. *Cahiers peur l'Analyse*, 9 p. 9–40.

Global Voices. http://www.globalvoicesonline.org/

Harvey, D. 2005. *A Brief History of Neoliberalism*. New York: Oxford University Press, USA. [Morita, N. et. al. trans. 2007. Tokyo: Sakuhinsya.]

Kageura, K. 2006. The status of corpora in human translation. In *Proceedings of the 12th Annual Meeting of the Association for Natural Language Processing*, p. 452–455.

Kageura, K. 2007. Terminological lexicons and terms in context: The translator's perspective. In *Proceedings of the 7th Conference of Terminology and Artificial Intelligence*, p. 1–10.

Kawamoto, K. and Inoue, K. 1997. *The Art and Craft of Translation* (in Japanese). Tokyo: University of Tokyo Press.

Leggett, J. 2005. *Half Gone*. London: Portobello. [Masuoka, K. et. al. trans. 2006. *Peak Oil Panic*. Tokyo: Sakuhinsha.]

Munday, J. 2001. *Introducing Translation Studies*. London: Routledge.

PaxHumana. http://paxhumana.info/fr.php3

Suzuki, M. 2006. There is no national border for stupidity. In Iwanami, ed. *The work of translators*, Tokyo: Iwanami, p. 209–214.

TUP. http://groups.yahoo.co.jp/group/TUP-Bulletin/

Venuti, L. 2004. *The Translation Studies Reader*. 2nd. ed. London: Routledge.