Analysis of Discussion Contributions in Translated Wikipedia Articles

Ari Hautasaari
Department of Social Informatics
Kyoto University
Yoshida-Honmachi, Sakyo-Ku,
Kyoto-Shi, 606-8225, Japan
arihau@ai.soc.i.kyoto-u.ac.jp

Toru Ishida
Department of Social Informatics
Kyoto University
Yoshida-Honmachi, Sakyo-Ku,
Kyoto-Shi, 606-8225, Japan
ishida@i.kyoto-u.ac.jp

ABSTRACT
Translation of articles in Wikipedia is one of the most prominent methods for increasing the quality of different language Wikipedias. Discussion pages in Wikipedia contribute to a large portion of the online encyclopedia, and are used by Wikipedia contributors for communication and collaboration. Although the discussion pages are the main channel between Wikipedia contributors all over the world, there have been relatively few in-depth studies conducted on communication in Wikipedia, especially regarding translation activities. This paper reports the results of an analysis of discussions about translated articles in the Finnish, French and Japanese Wikipedias. The results highlight the main problems in Wikipedia translation requiring interaction with the community. Unlike in previous work, community interaction in Wikipedia translation activities focuses on solving problems in the translation of proper nouns, transliteration and citing sources in articles rather than mechanical translation of words and sentences. Based on these findings we propose directions for designing supporting tools for Wikipedia translators.

Author Keywords
Wikipedia; translation; discussion;

ACM Classification Keywords
H.5.3. Group and Organization Interfaces: Web-based interaction.

General Terms
Human Factors; Design; Languages.

INTRODUCTION

Wikipedia is the largest collaboratively edited online encyclopedia available. Currently there are close to 19 million articles in 280 languages, and 29 million registered users in the multilingual Wikipedia. The English Wikipedia is currently the largest in terms of the amount of articles (3,6 million) and active users (145,000) followed by the German and French Wikipedias [23].

The overall growth of the English Wikipedia has slowed down in recent years due to problems in coordination, growing resistance to new content and tools available for editors and administrators [17]. However, coordination of activities in the non-encyclopedic pages, such as the Wikipedia article discussion pages, has continued to increase [12].

Consequently, one of the biggest issues in Wikipedia is making information available in all languages. The English Wikipedia is often used as the source language for translation activities aimed to enhance the quality of the multilingual Wikipedia. Translation activities in Wikipedia may be aimed at creating new articles in the target language Wikipedia, or increasing the quality of existing articles.

Few supporting tools tailored for Wikipedia translators exist today. For example, the Language Grid is an online infrastructure, which provides tools and resources for supporting Wikipedia translation activities [5, 6]. The language services, such as machine translators and multilingual dictionaries, available through the Language Grid are used for multilingual discussion support as well as for article translation with the aim of improving the quality of the multilingual Wikipedia and making information available in all languages.

Discussion in Wikipedia

Wikipedia discussions have a clear goal - reaching consensus within the community, and improving the Wikipedia article quality. Every Wikipedia page includes a "discussion" (or "talk") page for the purposes of interaction between Wikipedia contributors. In case of popular or controversial articles, the discussion pages may grow considerably in size, even exceeding the related article in terms of number of edits and length [18]. Where the article specific discussion pages include discussions mostly related to the corresponding article, discussions about policies and coordination of activities can span to various non-encyclopedic pages in the multilingual Wikipedia. For example, policies and guidelines in Wikipedia are created by the community members, and discussions about specific

ICIC'12, March 21–23, 2012, Bengaluru, India.

policies can include hundreds of contributions in the corresponding discussion pages[1].

Contributors outside the immediate working group of a particular article also use the discussion pages as a means of asking orienting information, and offering assistance for the article development [16]. For example, controversial articles often receive interest from casual editors (e.g. domain experts who edit articles anonymously). In the case of controversial articles, the anonymous contributions on discussion pages have been shown to increase animosities among the discussants, but anonymous article edits often have a positive effect on the article quality [10]. This suggests that problem and conflict resolution in Wikipedia discussions requires a vastly different set of strategies compared to other domains.

Besides the discussion pages related to articles, communication in Wikipedia is also conducted through specific WikiProjects and personal pages of registered users. The WikiProject pages include discussions ranging over multiple topics and activities, not necessarily related to particular articles, but often act as hubs for contributors interested in the same domain. For example, WikiProject:Japan[2] aims to improve the quality of articles related to Japan and Japanese culture in the English Wikipedia. In popular WikiProject pages, discussion contributions may often also be in foreign languages (e.g. Korean discussion contributions in the English language WikiProject), making it hard for users without the sufficient language ability to participate in the discussion.

In this study, we observe the communication and collaboration between Wikipedia contributors in relation to collaborative article translation in the Finnish, French and Japanese Wikipedia discussion pages. The aim of this paper is to identify the type of community interaction needed for successfully creating or amending an article via Wikipedia translation activities. Furthermore, we discuss the type of community interaction in relation to the different stages of a translated article. Based on these findings we propose directions for designing supporting tools for Wikipedia translators.

RELATED WORK

Previous research on communication in Wikipedia has focused on the correlation between discussion contributions and article quality [9, 10], but also on the analysis of content [15, 16, 18] and the structure of discussion pages [12]. Even though the content and interaction patterns have been observed in previous research, there are few studies available in the multilingual context. Especially considering the slowing growth of the English Wikipedia, translation activities in other language Wikipedias should be given more attention in order to support activities aiming to increase the quality of the multilingual Wikipedia.

Previous studies on Wikipedia translation have focused on supporting multilingual discussions between Wikipedia translators with machine translation tools [2, 3], creating a conversation control system developed from collaborative Wikipedia translation protocol [7], and supporting collaborative translation and editing of wiki-content with machine translation [1]. However, to best of our knowledge, there have not been studies focusing on the type of community interaction, mono- or multilingual, needed to solve problems specifically in Wikipedia translation activities.

HYPOTHESES

The main research questions in this paper are:

RQ1: What are the main tasks and problems requiring community interaction in Wikipedia article translation?

RQ2: Are there differences in collaborative translation and editing practices in different stages of the article evolution?

Firstly, we expected the majority of discussions to be about coordination between contributors related to editing the article and the article content [15]. Furthermore, we expected to find a high frequency of discussion contributions regarding the content of the partly or completely translated articles [18].

H1: Discussions about editing a translated article have a high frequency of contributions regarding the content of the article.

Secondly, we expected to find discussions about translation specific activities, such as help requests on how to translate certain words or sentences. More specifically, we expected a high frequency of help requests directed at domain experts regarding specific words and expressions in the translated articles [1].

H2: Discussions about translating an article have a high frequency of contributions regarding translation of specific words and expressions.

H3: Differences in the distribution of discussion contributions exist between the different stages of article evolution.

Differences in collaborative editing practices from a cultural point of view are discussed in [13]. Based on this work, we expected to find differences in the collaborative work practices in Wikipedia translation requiring community interaction in the three language Wikipedias.

---

[1] Policies and Guidelines project page in the English Wikipedia http://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines

[2] WikiProject:Japan in the English Wikipedia http://en.wikipedia.org/wiki/Wikipedia_talk:WikiProject_Japan

H4: There are differences in the type of community interaction conducted in discussion pages between Finnish, French and Japanese Wikipedias

DATA COLLECTION

In previous work, content analysis on the Wikipedia discussion pages has been done mostly in the English Wikipedia, focusing on a selected sample of general article pages [12, 15]. The data for this study was collected from the Finnish, French, and Japanese Wikipedias. The Finnish Wikipedia and the Japanese Wikipedia both represent a language group, which is dominant only in one country, whereas the French Wikipedia represents a language group ranging to multiple countries and cultures. Translations in the target Wikipedias are often conducted from the English Wikipedia due to the availability of new information [1].

For this study, we chose a data set from the categories listing partly or completely translated articles in each language [20, 21, 22]. The categories do not include all translated articles, but are a representative set of partly or completely translated articles in the target Wikipedias. Each identified category was mined for articles with contributions in the related discussion pages. We limited the discussion pages to be added in the data set to the articles in the categories of partly or completely translated articles, excluding related project and user pages.

We extracted 228 discussions pages with 720 discussion contributions from the Finnish Wikipedia, 93 discussion pages with 644 discussion contributions from the French Wikipedia, and 94 discussion pages with 330 discussion contributions from the Japanese Wikipedia (N = 1694).

ACTIVITY, CONTEXT AND ACTION IN WIKIPEDIA DISCUSSIONS ABOUT TRANSLATION

The individual contributions in the discussion pages were each categorized in three dimensions (activity, context and action) in order to identify the types of community interaction related to translating Wikipedia articles. The categorization used in this study was created after carefully reading the contributions once and definitions clarified before actual annotation.

In terms of article evolution in Wikipedia, it is important to identify the different stages of a translated article. More precisely, articles may be translated partly, completely or extended through translation activities. Partly translated articles can be further edited using target language sources. Thus, it is reasonable to differentiate between the discussion contributions related to these specific activities.

In the analysis of the discussion pages, two main *activity* categories emerged, where the discussion was either about *editing* a translated article (with no regard to the original source article), or about *translating* an article. As mentioned above, the activities of the Wikipedia contributors in different stages of the article evolution are distinctive, including the community interaction aspect. Furthermore, discussion pages, where the general topic is about translation, tend to focus only on translation. Similarly, discussion pages about editing a translated article focus on editing activities. Hence, every discussion contribution was categorized as part of an *editing* activity or a *translation* activity.

Categories for Message Context and Action for Discussion Contributions

Six categories for the *context* of the discussion contributions were identified, answering the question *"What is he/she talking about?"* The following categories indicate the main context of individual discussion contributions in relation to the corresponding article:

- Content: Discussion about the content of the article.
- Layout: Discussion about the layout of the article, including links, figures, templates, ect...
- Sources: Discussion about citing sources in the article.
- Naming: Discussion about naming of the article, sections and sub-sections, use of names and proper nouns, and transliteration in the article.
- Significance: Discussion about the significance of the article, section or sub-section.
- Wording: Discussion about how words, phrases and grammar are used in the article.

In this categorization, *naming* and *wording* are the two identified, mutually exclusive, topics of discussion about the use of language in the translated articles. *Wording* indicates, for example, a discussion about grammatical errors in sentences. *Naming,* on the other hand, includes discussions about article titles, or how to correctly refer to well known events (e.g. Watergate scandal).

As Wikipedia can be edited by anyone, *significance* of Wikipedia content is a common topic. A number of policies and guidelines have been established on whether content is notable enough to be included in Wikipedia (e.g. guidelines for notability of web content[3]). In this categorization, *significance* indicates discussion contributions about whether an article, or specific part of the content, is considered significant enough to be included in Wikipedia, or in the corresponding article.

In addition to the message context, the main intended *action* of a discussion contribution was coded, answering the question *"What does he/she do, or want others to do?"* Seven categories for *action* were identified:

- Help request: User asks for help directly or indirectly.

[3] http://en.wikipedia.org/wiki/Wikipedia:Notability_(web)

- Help provide: User provides help spontaneously or by request.
- Edit request: User requests for an edit on a specified topic, or for the whole article.
- Edit notice: User notifies that an edit has been conducted on a specified topic, or on the whole article.
- Critique: User provides critique regarding the article without directly prompting for action.
- Coordination: User coordinates actions with other users.
- Policy: User refers to an established Wikipedia policy calling for action, or calls for creation or modification of a local policy.

In cases where no clear categorization could be established for a discussion contribution, category "Other" was used. This category includes mostly personal insults and spam, which are not regarded as part of the discussion about article translation, and thus excluded from the analysis set.

Each discussion contribution was carefully read and categorized in three dimensions based on the main *activity, context* and *action*. In some cases discussion contributions could be very long and elaborate including multiple contributions in one. The coders would in these cases choose the most appropriate categorization representing the main intended contribution.

Inter-annotator agreement was tested with a reliability set of 90 items by two additional reliability coders in each language. The reliability coders were trained with a training set of 30 items in the target language not included in the reliability set. Fleiss' kappa was calculated for each language set for *activity, context* and *action*. The inter-annotator agreement in all languages is reported in Table 1. As the *kappa* may be unproportionally low with higher amount of categories, the overall agreement percentage is also reported in all languages.

	Kappa			
Language	**Activity**	**Context**	**Action**	**Total**
Finnish	0.74	0.82	0.65	84%
French	0.88	0.75	0.68	89%
Japanese	0.64	0.69	0.56	78%

Table 1. Inter-annotator agreement for activity, context and action in each language.

Examples of Discussion Contribution Categorization

Example 1 - activity is *editing*, context is *content* and action is *critique* without explicitly requesting an edit:

"*Limiting imaginary line. What is that supposed to mean?"* (Translation from Finnish by author)

Example 2 - activity is *editing,* context is *layout* and action is *edit notice*:

"*The links were directing only to a template, so there is no need for modification. If the author approves, it is OK to request a removal, but since it is fine to leave them here I will make them redirect.*" (Translation from Japanese by author)

Example 3 - where activity is *translation,* context is *sources*, and action is *critique* without explicitly requesting an edit:

"*I cannot support this request for fear that it imposes a non-neutral image in the article. I read the German article and its main concern is to criticize the basic supporters of the initiative in assuming a breach of international law and so on. So I'm sure we can find better sources*." (Translation from French by author)

Example 4 - activity is *translation,* context is *layout,* and action is *policy:*

"[…] *I added the translation template after the translation.* […] *If adding the template after the translation is done is not allowed, it can be removed."* (Translation from Finnish by author)

Example 5 - activity is *translation,* context is *content,* and action is *edit notice:*

"I added the English version. I will probably redirect it to the summer and winter [articles] [...]." (Translation from Japanese by author)

DISCUSSION ABOUT EDITING A PARTLY OR COMPLETELY TRANSLATED ARTICLE

The results in the *editing* activities were similar to previous studies on the discussion page contributions in terms of the frequency of discussions about content and coordination [15, 18]. In discussions about *editing* activities, the majority of contributions were about the *content* and the *layout* of the related Wikipedia article in all three languages. Figure 1 represents the distribution of discussion contributions in the three language Wikipedias regarding *editing* activities. In the Finnish Wikipedia (N = 404), discussions about *content* and *layout* each comprised 25.25% of the discussion contributions (50.50% in total). Similarly, in the French (N = 405) and the Japanese (N = 112) Wikipedias the majority of discussion contributions were about *content* and *layout* (72.09% and 68.75%, respectively). In this data set, the Japanese Wikipedia was the only one with more discussion contributions about *layout* than *content* when the discussion was about *editing* activities (40.18%).

The French Wikipedia had the highest frequency of contributions regarding *content* (58.02%). This can be partly explained by the amount of active contributors and articles, as well as the relative age of the French Wikipedia. The results are also similar to the English Wikipedia discussion contributions reported in [11]. In the older, or larger, Wikipedias, practices and policies are likely to be better established than in the younger, or smaller,

Wikipedias leading to a lower frequency of discussions about *layout*. Currently, the French Wikipedia ranks the third in number of articles and active users after the English and German Wikipedias [23].

Discussions about citing sources were relatively common in the Finnish and French Wikipedias (18.81% and 12.35%, respectively). In the Japanese Wikipedia, sources were less common with 7.14% of all discussion contributions regarding *editing* activities.

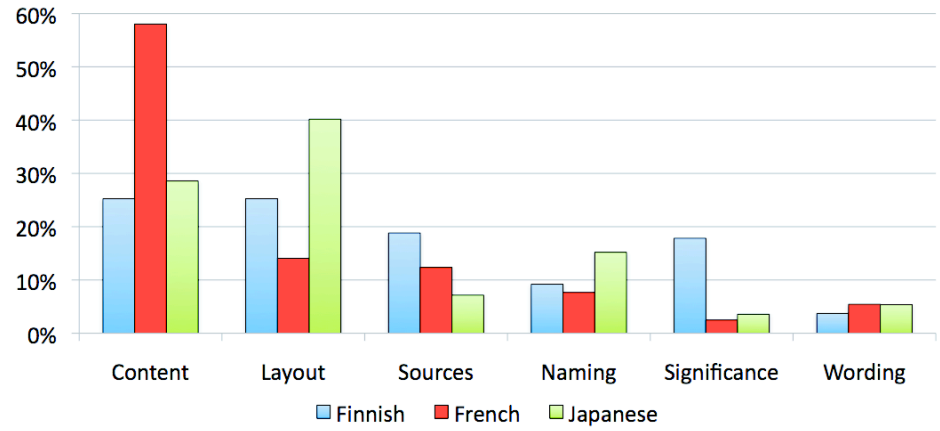

Figure 1. Distribution of *context* in discussion contributions related to *editing* activities in Finnish (N = 404), French (N = 405) and Japanese (N = 112).

DISCUSSION ABOUT TRANSLATING AN ARTICLE

In discussions about *translation* activities, the majority of discussion contributions were about *naming*. As described above, *naming* here refers to resolving the proper form for the title of the article, section or sub-section, names or proper nouns, and transliteration in the corresponding article. The context of these contributions is notably different from *wording,* as the category does not include phrasing or resolving proper translation of individual words or expressions.

Figure 2 represents the distribution of discussion contributions in the three Wikipedias regarding *translation* activities. In the Finnish Wikipedia (N = 302), *naming* is most common by 53.97%, whereas only 12.91% of contributions are about *wording*. Similarly, the French (N = 190) and Japanese (N = 217) discussion contributions are mainly about *naming* (54.21% and 49.31%, respectively), with only a small portion of contributions regarding *wording* (11.05% and 12.44%, respectively).

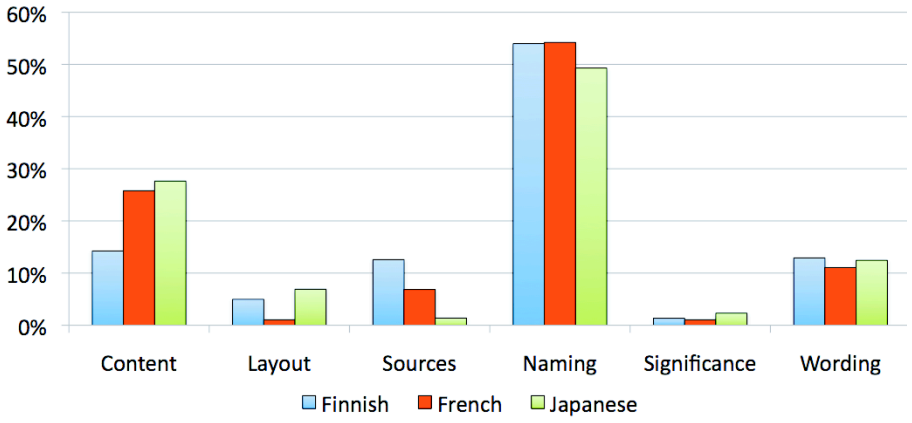

Figure 2. Distribution of *context* in discussion contributions related to *translation* activities in Finnish (N = 302), French (N = 190) and Japanese (N = 217).

The results were unexpected as we assumed a high frequency of discussion contributions regarding *wording* (*H2*). The hypothesis was based on the assumption that Wikipedia translators would ask for help from domain experts regarding domain specific words [1] and article content [18]. However, in *translation* activities community interaction was required most frequently when resolving problems in *naming* based on other language Wikipedias and external sources, such as target language media.

Compared to discussions about *editing* activities, discussions about *translation* activities include far less contributions regarding *sources*. The tendency was consistent across all languages, with 12.6% in the Finnish discussions, 6.8% in the French and 1.4% in the Japanese discussions.

Example 6 is a typical case of discussion about a *translation* activity. The discussion contributor makes a direct reference to the translation of the article name. Example 6 also illustrates a recurring phenomenon in discussions about *naming*. The reference for using the version in a Finnish language newspaper or other media is often associated with *naming* practices. This is related to the policy stating that Wikipedia contributors do not create content, be it the name of an incident or transliteration, but use outside resources for accurate encyclopedic articles.

Example 6 - activity is *translation*, context is *naming*, and action is *edit request*:

"*I think the article name is fairly bad. Presumably, 'Norwegian*

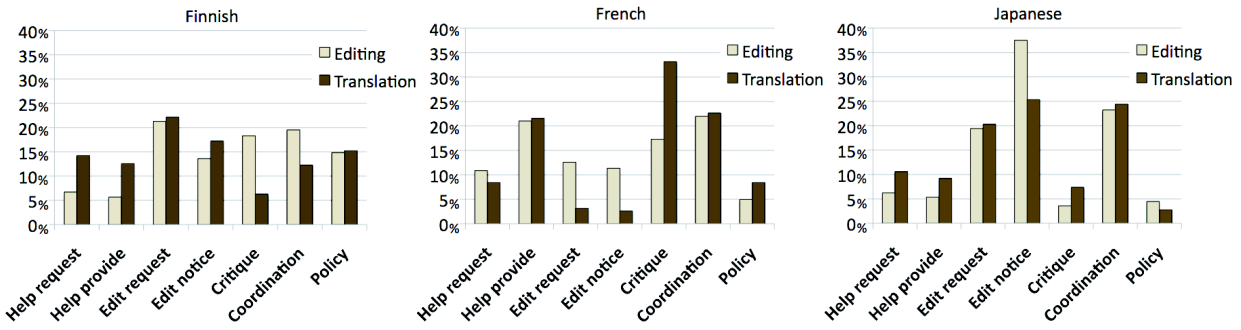

Figure 3. Distribution of *actions* in discussion contributions in Finnish, French and Japanese.

Rocket Incident' is a direct translation from the English Wikipedia, but [the title] *has not been used anywhere. It would be better to find what the incident is called in the* [Finnish language] *newspapers."* (Translation from Finnish by author)

ACTIONS IN DISCUSSION CONTRIBUTIONS

The distribution of *actions* in each language regarding both *editing* and *translation* activities is illustrated in Figure 3. The frequency of discussion contributions in the French Wikipedia were slightly higher in the *help provide* (21.18%) and *critique* (22.35%) categories compared to Finnish and Japanese Wikipedias. On the other hand, the frequency of *edit requests* (21.67% and 20.06%) and *edit notices* (15.16% and 29.48%) were slightly higher in the Finnish and Japanese Wikipedias, respectively.

The discrepancies between the types of *actions* can be explained by slightly different practices expressed in the discussion contributions. While it was common in the French Wikipedia to provide help unsolicited, the contributors' aim would be to increase the quality of the article. Similarly, the Finnish and Japanese contributors' intention was to increase the article quality by asking for a revision or an edit on a particular subject, which would prompt an *edit notice* from other contributors. The trend was also consistent between *editing* and *translation* activities in each language (Figure 3).

The age and size of a given Wikipedia may be a factor in the type of behavior expressed in discussion contributions. The Finnish Wikipedia, being the smallest in terms of number of contributors, and youngest in terms of number of articles, may not have as well established practices within the community as in the French and Japanese Wikipedias. This is also observable in the frequency of discussion about *policy* (15.01%) as opposed to the French (6.0%) and Japanese (3.3%) Wikipedias.

RESULTS

Regarding *editing* activities in the translated articles, our results were consistent with previous research, with a high frequency of discussion contributions about *content* and *layout* (*H1*). However, the trend did not persist in discussions about *translation* activities. Our hypothesis was that there would be high amount of discussions about how to translate certain domain specific words and expressions. In other words, we expected a high amount of *help requests* regarding *wording* (*H2*).

The majority of discussion contributions regarding *translation* activities were about *naming*; transliteration, translation of names and proper nouns and resolving article, section and sub-section titles based on various language sources. *Help requests* on *wording,* translation of words and expressions, were observed, but not nearly as frequently as discussions about *naming*. The mechanical translation practices of Wikipedia contributors were not in the scope of this study, but based on these results we can conclude that the tasks or problems requiring community interaction are more often related to *naming* rather than mechanical translation of words and phrases.

One reason for the large amount of discussions about *naming* is the diversity in naming practices of events between different language sources, such as mass media. Especially in the Finnish Wikipedia, discussion about *sources* was common (16.15%). These two topics are loosely related, as direct translations of the names of well-known events are often not acceptable in the target language Wikipedia. In other words, if an event has had some media coverage, the commonly used name should be adopted also in the same language Wikipedia. The following example illustrates this observed behavior.

Example 7 includes a discussion about a *naming* of an article, where the community members dispute the article name choice based on the unavailability of target language references. The first discussion contributor has translated the article from English and chosen the title based on his translation, but it has been changed by another contributor (Example 7.2). In Example 7.1, the original translator notifies that he reverted the change. The second editor disputes the translated article name and implies an edit request referring to lack of sources. The conversation ends when the original translator is not able to produce a source

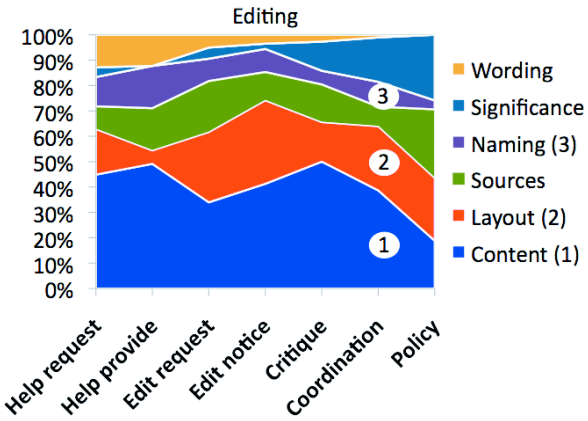

Figure 4. Distribution of discussion contributions related to *editing* activities in all three languages.

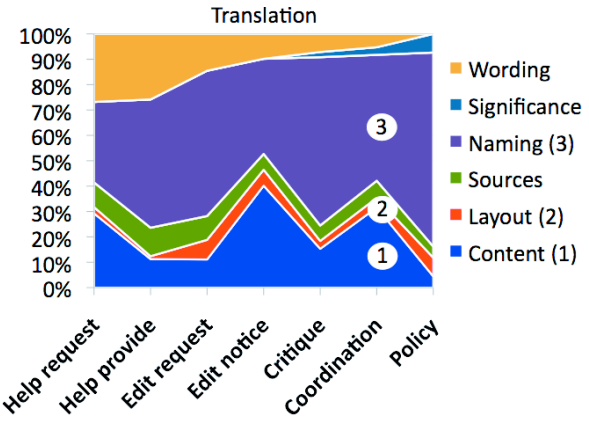

Figure 5. Distribution of discussion contributions related to *translation* activities in all three languages.

in the target language that includes a reference to the disputed article name (Example 7.3).

Example 7.1 – activity is *translation,* context is *naming* and action is *edit notice:*

"*I know that* [name of article] *is its own separate concept. I just translated it from the English Wikipedia.*" (Translation from Finnish by author)

Example 7.2 – activity is *translation,* context is *sources* and action is *edit request:*

"*Could you present a source example where the Finnish concepts have the difference in meaning.*" (Translation from Finnish by author)

Example 7.3 – activity is *translation*, context is *naming* and action is *coordination*:

"*A book called* [reference]. *However it talks about* [unrelated subject]. […] *I have nothing against changing the name to* [original article name]" (Translation from Finnish by author)

Our results imply that the current approaches for supporting Wikipedia translation are not necessarily solving the main problems in Wikipedia translation. For example, a system introduced in [11] supports mechanical article translation with machine translation support, but does not address two key problems identified in this paper – target language *sources* and *naming*.

Effect of Activity Type

As predicted in *H3*, the type of *activity* related to the discussion contributions has an effect on the distribution of the messages in terms of *context*. In other words, when participants discussed about *translating* an article the distribution of discussion contributions was significantly different than when discussing about *editing* an article (χ^2 = 54.4, $p < .05$).

As illustrated in Figure 4 and Figure 5, the frequency of discussions about *content* (1) and *layout* (2) is much lower in discussions about translation. Furthermore, frequency of *naming* (3) is considerably higher when translating an article compared to editing activities. The shift in the distribution of discussion *context* between activities illustrates the most common types of collaborative work practices of the Wikipedia translators in different stages of the article evolution:

- Translating an article – Resolving translation of article title, subsection titles, names, proper nouns and transliteration. Resolving problems in translation of words and expressions. Finding target language sources.
- Editing a translated article – Discussing editing activities and resolving disputes regarding the article content. Citing sources in target language. Resolving problems in article layout.

Differences in the Distribution of Context and Action of Discussion Contributions Between Finnish, French and Japanese Wikipedias

As predicted in *H4*, differences in the distribution of *context* were found between the Finnish and French Wikipedias (χ^2 = 16.7, $p < .016$). In the distribution of *action,* there were differences between all languages (Fi-Fr: χ^2 = 23.0, $p < .016$, Fi-Ja: χ^2 = 18.5, $p < .016$, and Fr-Ja: χ^2 = 41.8, $p < .016$). However, the results here are not straightforward (i.e. clear cultural differences cannot be indicated). Firstly, it is not reasonable to generalize the results to a cultural level for a number of reasons. The results indicate that community interaction is indeed different in the different language groups in some aspects, but considering the scope of the data set, we cannot say if the differences rise from the group composition itself (i.e. Wikipedia contributors), language, culture or some other factors.

Secondly, there are observable similarities within the data set. For example, the distribution of *naming* and *wording* in discussions regarding *translation* of an article is consistent between the three languages (Table 2). This suggests that problems in mechanical translation of words and expressions, as well as translation of names and proper nouns are equally common between different language and different size Wikipedias. From a user point of view, the two translation problems require equal amount of community interaction regardless the target language of the translated article.

	Naming	Wording
Finnish	53.97%	12.91%
French	54.21%	11.05%
Japanese	49.31%	12.44%

Table 2. Frequency of naming and wording in discussions related to translation activities.

Thirdly, although there were some discrepancies between the different language Wikipedias regarding the distribution of *actions*, the results cannot be generalized. As we discussed above, the different behavioral aspects in discussion contributions all have the same goal – improving the quality of the related article. Furthermore, the French Wikipedia represents a language group that spans over multiple countries and regions. Hence, it is unreasonable to draw a generalization about the cultural aspects of the Wikipedia contributors based on this data set, but rather use the results as an indication of behavioral tendencies in discussions about translation in different language Wikipedias in future studies.

OPPORTUNITIES FOR DESIGN

The MediaWiki software, originally designed to run Wikipedia, has been constantly updated in terms of features, but the basic functionalities have stayed the same. A number of open source developers have created a vast variety of extensions for the software to increase its functionality in Wikipedia as well as in other sites. Based on our results, we propose directions for designing supporting tools for Wikipedia translation, especially through open source development of MediaWiki extensions[4].

Support for Consistent Translation of Names and Proper Nouns

The emergent problem identified in this study is the consistent translation of proper nouns and names, as well as transliteration. Firstly, the problem stems from the translators' inability to properly notate the given proper noun or name in the target language. Currently there are many solutions to overcome the problem of transliteration. For example, foreign names can be automatically transliterated to Japanese katakana characters, but in many cases there may not be a direct equivalent available, especially for more complex names [14]. Reversely, names written in kanji characters (e.g. Chinese) can be transliterated to English using machine translation techniques [19].

However, a more complex problem emerges when translating a name of a current event. In general, names of events have an established form, such as "Watergate scandal", but in the case of a very recent event, it might be difficult to determine the proper name. Especially in the case of Wikipedia, the naming of an article is important because of search engines; if an article is named improperly, users do not find the information with the common search words.

In case of breaking news events, such as major natural disasters, online communities engage in diverse forms of collective action to share the rapidly changing information [8]. However, if there is no established method to translate keywords in the disaster information, such as the common name of the event and location, it can be difficult for translators to make use of the emerging information in a timely fashion. Furthermore, it can be very hard for different language speakers to find information through search engines until the keywords included in the information, such as article topics, have a well established form.

To ensure that translation of names, proper nouns and transliteration is consistent over multiple translated articles, the following design aspects should be considered:

- Accessibility to a user editable multilingual dictionary resource (e.g. multilingual domain specific naming dictionary) for referencing established translations of names and proper nouns in the target language Wikipedia.
- Dictionaries and dictionary entries should be arranged according to the domain of the translated article (or sentence) to avoid mistranslations in unrelated articles (see e.g. [4]). In other words, when multiple entries exist, the domain of the translated article is used to resolve the disambiguation.
- Ability for users to browse/search the dictionary entries in Wikipedia as well as combine available dictionaries with other language resources.
- Ability for contributors to coordinate through discussion pages directly related to a specific dictionary or dictionary entry in order to resolve inconsistencies in a centralized repository.

There are still open questions on the use of dictionaries in translation, namely on how to decide which of multiple entries is the correct translation. Especially in the case of specialized words and expressions, interaction with domain experts is essential to avoid word-sense disambiguation in

[4] http://www.mediawiki.org/wiki/Manual:Extensions

user-edited domain specific dictionaries. A multi-language discussion platform developed for supporting communication between Wikipedia translators and domain experts in such tasks is introduced in [2].

Support for Citing Sources in Translated Articles

As discussed above, source citing in translated articles is another prominent problem. In general, the consensus in Wikipedia communities is to "*always cite sources in the target language*". However, finding the right sources in the target language can be a time consuming effort, especially in cases where the translator is not familiar with the topic (i.e. not a domain expert). Further, a source containing the particular information may not exist in the target language.

We propose two approaches for supporting Wikipedia translators in terms of source citing:

- Automated search for a translated source material in the target language available in online archives (e.g. with cross-referencing standard book numbering, closest match of machine translated title/author name, mining available web resources, ect.)
- Development of crowdsourcing translation tool for open content sources not available in the target language using machine translation as a supporting tool.

The problems in translation found in this study are likely also present in other translation domains. For example, crowdsourcing translation is becoming increasingly popular for cheap translation of large amounts of text. However, the problems in *naming* may be increasingly difficult in crowd translation services, if no communication channel between collaborators is available.

CONCLUSION

Translation activities in Wikipedia aim to improve the quality of the multilingual Wikipedia by making information available in all languages. Numerous communities and WikiProjects are working towards improving the articles in their domain of choice via article translation. Discussion pages are the main communication channel between Wikipedia contributors for organization and coordination of activities in the corresponding article, user pages and other non-encyclopedic Wikipedia pages.

In this paper, we observed the discussion contributions in three language Wikipedias, focusing on discussions about partly or completely translated articles. We identified two types of activities, *editing* and *translation*, reflected in the discussion contributions. This showed that Wikipedia article translation is a two-fold process, where an article is first partly or completely translated to the target language and edited by Wikipedia contributors as a separate activity, or an existing article is improved via translation activities.

The discussion contributions were further divided in to six categories based on the *context*, and seven categories based on the intended *action*. *Context* reflects the general topic of a discussion contribution and *actions* reflect the action requested or provided by the message contributor. In combination, these dimensions provided a distribution of discussion contributions, which revealed collaborative tasks and problems requiring community interaction specific to Wikipedia translation.

Our results were consistent with previous research in terms of type of discussions about *editing* a translated article, with high frequency of contributions regarding the *content* and *layout* of the related Wikipedia article. However, in all three Wikipedias, discussions about *translation* activities were most frequently about proper *naming* of articles, sections and sub-sections, transliteration, proper nouns and names. Our results show that community interaction is needed most frequently when translating or transliterating names and proper nouns based on target language sources, in all observed languages.

Our results imply that current approaches to Wikipedia translation may not answer the most pressing questions in the quality of translated articles in Wikipedia. For example, current machine translation tools available for Wikipedia translation do not accommodate adequate and consistent translation of names and proper nouns across several related articles. Furthermore, as seen in this paper, resolving the proper *naming* practices often requires community interaction centralized in one particular article's discussion page making consistent *naming* difficult in related articles.

Although we only found few multilingual discussion contributions in the data set, it is apparent that given the right tools, Wikipedia translators would greatly benefit from access to domain experts and contributors in other language Wikipedias. Our approach for overcoming inconsistencies especially regarding *naming* practices in multiple related articles is to implement a multilingual collaboratively edited domain specific dictionary as a supporting tool for Wikipedia translation. By providing a domain specific multilingual dictionary, discrepancies in *naming* could be lowered significantly without affecting articles outside the given domain. The concept was first introduced in [4], and an existing infrastructure for distributing language services, such as user created domain specific dictionaries, is described in [5, 6].

FUTURE WORK

This study focused on Wikipedia discussions in three language Wikipedias, which is more than most studies thus far have included. Even though the English Wikipedia is the largest, and moreover the best resource for data, there are cultural aspects in Wikipedia communities in different language Wikipedias, which are not observable in the English Wikipedia. Hence, a follow-up study including a wider variety of languages is needed.

A temporal study on the evolution of a translated Wikipedia article in relation to the community interaction including the mechanical aspects of Wikipedia translation will be reported in a future paper. We are aiming to support translation of Wikipedia articles to multiple languages, as well as supporting multilingual discussions between translators and domain experts with different language abilities using machine translation tools.

ACKNOWLEDGMENTS

This research was partially supported by Kyoto University Global COE Program: Informatics Education and Research Center for Knowledge-Circulating Society, and Strategic Information, and Communications R&D Promotion Programme (SCOPE) from Ministry of Internal Affairs and Communications of Japan.

REFERENCES

1. Desilets, A., Gonzalez, S., Paquet S., and Stojanovic, M. Translation the Wiki Way. In *Proc. of the 2006 International Symposium on Wikis,* ACM Press (2006), 19–32.
2. Hautasaari, A., Ishimatsu, M., Xia, L., and Ishida, T. Supporting Multilingual Discussion of Wikipedia Translation with the Language Grid Toolbox. *IEICE technical report. Natural language understanding and models of communication 109(390)* (2010), 67-72.
3. Hautasaari, A., Takasaki, T., Nakaguchi, T., Koyama, J., Murakami, Y., and Ishida, T. Multi-Language Discussion Platform for Wikipedia Translation. In *Ishida, T. (ed.). The Language Grid - Service-Oriented Collective Intelligence for Language Resource Interoperability*, Springer (2011), 231-244.
4. Hautasaari, A., and Ishida, T. Semantic Web Approach to Support Wiki-to-Wiki Translation Communities. In *Proc. JAWS 2009*, 483-488.
5. Ishida, T. *The Language Grid - Service-Oriented Collective Intelligence for Language Resource Interoperability*. Springer, 2011.
6. Ishida, T. Language Grid: An Infrastructure for Intercultural Collaboration. *IEEE/IPSJ Symposium on Applications and the Internet,* IEEE Computer Society (2006), 96-100.
7. Ishimatsu, M., Murakami, Y., Hautasaari, A., and Ishida, T. Supporting Wikipedia Translations Based on Protocol Analysis. *IEICE technical report 110(428*) (2011), 63-68. (In Japanese)
8. Keegan, B., Gergle, D., and Contractor, N. Hot off the wiki: dynamics, practices, and structures in Wikipedia's coverage of the Tōhoku catastrophes. In *Proc.* WikiSym '11, ACM (2011), 105-113.
9. Kittur, A., and Kraut, R.E. Harnessing the wisdom of crowds in Wikipedia: quality through coordination. In *Proc. CSCW'08*, ACM (2008), 37-46.
10. Kittur, A., Suh, B., Pendleton, B., and Chi, E. He says, she says: Conflict and coordination in Wikipedia. In *Proc. CHI'07,* ACM Press (2007), 453-462.
11. Kumaran, A., Saravanan, K., Datha, N., Ashok, B., and Dendi, V. Wikibabel: A wiki-style platform for creation of parallel data. In *Proc. of the ACL-IJCNLP 2009 Software Demonstrations,* ACL (2009), 29–32.
12. Laniado, D., Tasso, R., Volkovich, Y., and Kaltenbrunner, A. When the Wikipedians talk: network and tree structure of Wikipedia discussion pages. In *Proc. ICWSM*, AAAI Press (2011), 177-184.
13. Pfeil, U., Zaphiris P., and Ang, C.S. Cultural Differences in Collaborative Authoring of Wikipedia. *Journal of Computer-Mediated Communication, Volume 12, Issue 1* (2006), Blackwell Publishing Inc, 88–113.
14. Qu, Y., Grefenstette, G., and Evans, D.A. Automatic transliteration for Japanese-to-English text retrieval. In *Proc. SIGIR '03,* ACM (2003), 353-360.
15. Schneider, J., Passant, A., and Breslin, J.G. Understanding and Improving Wikipedia Article Discussion Spaces. In *Proc. SAC'11*, ACM (2011), 808-813.
16. Stvilia, B., Twidale, M.B., Smith, L.C., and Gasser, L. Information quality work organization in Wikipedia. *JASIST,* 56(6) (2008), 983-1001.
17. Suh B., Convertino, G., Chi, E., and Pirolli, P. The singularity is not near: slowing growth of Wikipedia. In *Proc. WikiSym '09,* ACM (2009), 1-10.
18. Viegas, F.B., Wattenberg, M., Kriss, J., and van Ham, F. Talk before you type: Coordination in Wikipedia. In *Proc. HICSS'07,* IEEE Computer Society (2007), 1-10.
19. Virga, P., and Khudanpur, S. Transliteration of proper names in cross-lingual information retrieval. In *Proc. of the ACL 2003 workshop on Multilingual and Mixed-Language Named Entity Recognition*, ACL (2003), 57-64.
20. Wikipedia - Finnish Wikipedia (Referred: April 2011): http://fi.wikipedia.org/wiki/Luokka:Käännetyt_lähteettömät _artikkelit/.
21. Wikipedia - French Wikipedia (Referred: April 2011): http://fr.wikipedia.org/wiki/Catégorie:Projet:Traduction:Articles_liés/.
22. Wikipedia - Japanese Wikipedia (Referred: April 2011): http://ja.wikipedia.org/wiki/Category:%E7%BF%BB%E8%A8%B3%E4%B8%AD%E9%80%94/.
23. Wikipedia - List of Wikipedias (Referred: 30.5.2011): http://meta.wikimedia.org/wiki/List_of_Wikipedias/.