# "Could Someone Please Translate This?" – Activity Analysis of Wikipedia Article Translation by Non-Experts

**Ari Hautasaari**
Department of Social Informatics
Kyoto University
Yoshida-Honmachi, Sakyo-Ku,
Kyoto, 606-8501, Japan
arihau@ai.soc.i.kyoto-u.ac.jp

## ABSTRACT
Wikipedia translation activities aim to improve the quality of the multilingual Wikipedia through article translation. We performed an activity analysis of the translation work done by individual English to Chinese non-expert translators, who translated linguistically complex Wikipedia articles in a laboratory setting. From the analysis, which was based on Activity Theory, and which examined both information search and translation activities, we derived three translation strategies that were used to inform the design of a support system for human translation activities in Wikipedia.

### Author Keywords
Wikipedia; translation; activity analysis; non-experts.

### ACM Classification Keywords
H.5.2 [Information Interfaces and Presentation]: User Interfaces – Evaluation/methodology; J.5 [Computer Applications]: Arts and Humanities – Language translation.

### General Terms
Human Factors; Design.

## INTRODUCTION
Wikipedia is the largest collaboratively edited online encyclopedia available, and the sixth most visited website in the world[1]. Unsurprisingly, almost 60% of the traffic is directed to the English language Wikipedia. The English Wikipedia is the largest in terms of number of articles (close to four million), and active contributors. Together with the German, French and Dutch Wikipedias, the four language versions are the only ones with over one million articles. Combined, there are currently close to 22 million articles in 285 languages, and 34 million registered users in the multilingual Wikipedia[2].

Currently, 6,912 living languages, including sign languages, exist in the world, and in some countries, such as Indonesia, hundreds of languages are spoken by the people [12]. In this light, aiming to provide information equally in all languages seems like a daunting task to undertake. The only way to realistically achieve such a goal would be to rely on volunteer contributors, such as in Wikipedia, to translate existing content. However, even when focusing on the currently existing language versions of Wikipedia, the number of active contributors, especially in the smaller Wikipedias with fewer than 1,000 articles, is very low. The question is; how can we support the Wikipedia contributors translating articles in the multilingual Wikipedia?

### Wikipedia Translation
First, we need to define what we mean by Wikipedia translation. In this study, "Wikipedia translation" refers to the activities related to translating Wikipedia article pages. Wikipedia translation can also include the translation of non-encyclopedic pages, such as WikiProject pages, and even discussion pages attached to each Wikipedia article [15].

Wikipedia articles can be overtly complex in linguistic terms, as is evident from the existence of the Simple English Wikipedia[3], which includes English Wikipedia articles written in simplified language. In previous research [31], the main differences identified between the Simple English Wikipedia articles and the corresponding English Wikipedia articles were the less frequent use of complex words, and decreased variety of syntactic structures in the Simple English versions. Furthermore, the language used in conceptual articles was found to be more complex compared to biographical and object-based articles [31].

The language complexity in Wikipedia articles becomes increasingly apparent in translation tasks. As Wikipedia is a collaboratively edited encyclopedia, translation of articles is also a collaborative effort. In previous research, this aspect of Wikipedia translation was studied through the analysis of discussion contributions in translated Wikipedia articles


*CSCW '13,* February 23–27, 2013, San Antonio, Texas, USA.


[1] http://www.alexa.com/siteinfo/wikipedia.org

[2] http://meta.wikimedia.org/wiki/List_of_Wikipedias

[3] http://simple.wikipedia.org/wiki/Main_Page

[14]. The main problem during a translation task that required the translators to consult the Wikipedia community for help was identified as the translation of names, proper nouns and transliteration, as opposed to resolving the syntactic and semantic structure of the translated text.

Various supporting systems have been developed for the Wikipedia translation community, ranging from machine translation supported article editors, such as Microsoft's WikiBhasha[4], to translation workflow management [9, 23]. Machine translation has also been proposed as a viable solution to overcome some challenges when translating complex Wikipedia articles to multiple languages with the help of monolingual domain experts [15]. Machine translation techniques also enable cross-language information retrieval from websites [2], and multilingual collaboration in online environments [30]. "The aim of using computers for translation is not to emulate or rival human translation but to produce rough translations which can serve as drafts for published translations, as gists for information gathering, and as cross-language communication aids" [17]. For high quality translation a human translator is still required to be kept as part of the translation process.

While Wikipedia article translation is a collaborative effort, there has been little previous work done regarding the activities of individual volunteer translators. This study focuses on the individual actions of Wikipedia translators as components of collaborative translation work in Wikipedia, and gives a first glimpse on how the needs of individual non-expert translators can be accommodated in a Wikipedia translation support tool [13].

## RESEARCH QUESTIONS

The research questions in this study were derived from the existing work regarding Wikipedia translation [14]. In the previous work, the discussion contributions of Wikipedia contributors were analyzed to identify what type of problems in article translation process require community interaction in the Wikipedia discussion pages. After further investigating the existing literature, we found that there is very little research focusing on the translation process of individual Wikipedia translators.

Although translation studies are by no means a new field, studies conducted with non-professional translators tend to include students of translation as subjects. This led us to conclude that we do not know *how* volunteer Wikipedia contributors translate articles. Hence, our initial research question was:

> RQ1. How do non-expert Wikipedia translators conduct translation work?

The second and third research questions were clarified during a preliminary data analysis, where we noted the extensive need to access online resources during the translation process.

> RQ2. What are the information needs of non-expert translators?
>
> RQ3. How are the information needs resolved in an online environment?

This study is exploratory in the sense that we do not take a strict standpoint of translation studies, but focus on the human-computer interaction in Wikipedia article translation. Through an Activity Theoretical approach, we aim to explain the translation process of individual non-expert translators, and the process of resolving information needs during a task of Wikipedia article translation. Based on the analysis results, we draw design implications for a Wikipedia translation support tool.

## RELATED WORK

Previous studies have concluded that professionals and non-professionals show different types of behavior during a translation task [8, 18, 21]. In [26], Lörscher gives a brief comparison of the differences in professional and non-professional translation processes. The main differences found were that non-professionals tend to be more form-oriented, focus on translation as a problem solving task, and tend to include formulating problems in the translated text [27]. In [32], the cost of crowdsourcing translations is shown to be lower than when using professional translators, while still keeping the quality up to par.

Different language versions of a Wikipedia article can include a great deal of unique information. Omnipedia[5] is one project that aims to bring the information in the multilingual Wikipedia articles available to users in their native language. Previous studies have revealed that monolingual users are not aware of the multitude of information not available in their own language, while also highlighting the importance of culturally-aware hyperlingual applications [3, 6, 16]. In [6], the authors recommend a user-driven approach to Wikipedia translation, where cultural differences in different language Wikipedias are respected, and mechanical translation is supported by machine translation.

## METHODOLOGY

### Theoretical Background – Activity Theory

According to Activity Theory (AT), human activity is always directed towards an object, which exists outside the human. Activity Theory defines a three-level hierarchy for human activities: Activity-Action-Operation. Activities are always directed towards a motive. Activities are divided in

[4] http://www.wikibhasha.org

[5] http://omnipedia.northwestern.edu

to Actions, which are directed towards specific goals. Actions are again divided to Operations, which are lower-level units executed and adjusted to specific conditions of the context, and often internalized to a point where people may not be aware how they are performed.

Human activity is mediated through artifacts, with which humans interact with objects. Artifacts can range from physical objects, such as carpenter tools, to psychological artifacts, such as language and signs. Human activity is also collaborative. Humans do not act in a vacuum, but engage in collaborative processes with other people. In collaborative activities, different people conduct different actions belonging to an activity, aligning the goals of each action towards the objective of the activity through communicative artifacts. Activity Theory recognizes the evolving nature of human activities: Activities, actions and operations, the relationships between the activities, mediating artifacts, and socio-cultural context are constantly changing. This means that an activity can become an action, and an action can become an operation when the goals change [25].

For an in-depth introduction and discussion on Activity Theory and it's application to HCI design see: [11, 22, 24].

### Activity Analysis

Activity Analysis aims to provide a detailed description of human activity conducted within the constraints and resources of the real-world setting. As detailed in [4], Activity Analysis consists of a qualitative study, an observation phase, and a detailed analysis of activity patterns based on Activity Theory.

In this study, the qualitative phase included a video recording of the activity. The focus of this study is on the activity system of Wikipedia article translation, and we began the analysis by collecting data on the following:

**Activity** – The Wikipedia article translation activity, actions, operations, actors, and artifacts used in the translation activity.

**Person** – Activities and actions of individual non-expert translators conducting the translation activities.

**Artifacts** – The use of specific tools for mediating activities and actions in Wikipedia article translation.

In the observation phase, we created a coding schema for the analysis of the collected data based on Activity Theory. The schema was focused on the following characteristics of the Wikipedia translation task:

**Activity** – Identifying all the activities included in the translation task based on the motive and object of work.

**Action** – All actions performed as a part of an activity.

**Operations** – Manual operations performed as a part of an action.

**Context** – Context of each action including the time of action occurrence during an activity, and the artifacts used to mediate the action.

**Actors** – Non-expert translators conducting the activities. In this study, each translator was working individually.

The created coding schema allowed the identification of activities, actions, the starting time and end time of each subsequent action, details on the operations, tools used to perform and actors performing each action.

The specific activities were distinguished from actions, and other activities, by observing the *object* of the task in the given context. In Article Translation Activity, the object is the translated article. As discussed in detail later, the object of the task could change to a single word, phrase or information item, and the activity was then defined as a Information Search Activity.

## EXPERIMENTAL DESIGN

The task for the experiment participants was to translate an English Wikipedia article to Chinese. The criteria for choosing the articles for this experiment were:

1. There was no existing translation in the Chinese language Wikipedia [6] or in the Chinese collaboratively edited online encyclopedia, Baidu Baike[7].
2. Length of each article/passage is no more than 500 English words.
3. Each article represents a different domain (i.e., no two articles are about computer science).
4. The article content is representative of the general Wikipedia article language complexity, and does not have a version in the Simple English Wikipedia[8].

The length of each translation task was limited to no more than one hour thirty minutes. To fit this timeframe, the length of each English passage was set to a maximum of 500 words. "Akan Volcanic Complex" and "Ying Huang" were complete Wikipedia articles, whereas "Crowdsourcing" included two sections of the full article not available in the Chinese language Wikipedia or Baidu Baike to match the set article translation length (Table 1). The source articles were uploaded to a Wiki-server dedicated for this experiment.

### Participants

The experiment participants were university graduate students of various fields hired specifically for this task. The group consisted of 15 Chinese native speakers, with at

[6] http://zh.wikipedia.org

[7] http://baike.baidu.com

[8] http://simple.wikipedia.org/wiki/Main_Page

least a score of 61 in TOEFL iBT[9] or 500 in TOEFL PBT[10] academic English aptitude test.

| Article | Number of words | Average translation time |
|---|---|---|
| Akan Volcanic Complex http://en.wikipedia.org/wiki/Akan_Volcanic_Complex | 490** | 93 minutes |
| Ying Huang http://en.wikipedia.org/wiki/Ying_Huang | 466** | 82 minutes |
| Crowdsourcing* http://en.wikipedia.org/wiki/Crowdsourcing | 412** | 93 minutes |

** "Web-based Crowdsourcing" and "Collaboration".*
*** On December 12th 2011.*

**Table 1: Articles for translation.**

In this paper, the term "non-expert translator" refers to the experiment participants who were A) not professional translators, or have not received training in translation, and B) not domain experts with extensive knowledge on the topics of the articles being translated. With this choice of terminology, we aim to differentiate the subjects in this study from non-professional translators, who have often been subjects in previous translation studies (e.g., [1, 20]). The participants in this study represent bilingual "Site Visitors", or casual users, who are potential contributors and one target user group for Wikipedia translation support tools [9].

### Experiment Task

We asked each translator to create a translated Chinese language Wikipedia article from the source article. The participants were only given basic instructions on how to create a new Wikipedia article page and a brief tutorial to the interface options in the MediaWiki software.

Each participant was randomly assigned to one of the three articles as a translator. Hence, each article was translated by five individual non-expert translators during the experiment. The participants were given identical laptop computers, which were cleaned of any page history, cache and auto-filling options.

### Data Collection

We recorded the computer screen of each participant for the full duration of the experiment task. In addition, we collected the browsing history and the final translated Chinese article from each participant. In previous studies, the think-aloud method has been used to collect data from translators during a translation task, but we opted to rely on the video recording due to previous inconclusive evidence on the effect of think-aloud method on the translation product [19]. Overall, we collected approximately 18 hours of video material, 15 fully translated articles, and the browsing history of each of the 15 participants.

### Data Annotation

After the experiment, each video recording was annotated for activities, actions and operations during the article translation process. Each operation was identified by a mouse-click (such as changing a browser window) or keystroke (such as typing a letter), which were highlighted in the video log, but not on the participants' screen. Each action was given a timestamp based on the video recording (excluding administrative tasks). In total, we identified 16 actions during a translation activity, and 23 operations belonging to one or more actions.

## ACTIVITY ANALYSIS OF WIKIPEDIA ARTICLE TRANSLATION BY NON-EXPERTS

### Translation Strategies

We annotated the individual actions of each experiment participant during the Article Translation Activity, where the object is defined as the translated article. The annotation schema is illustrated in Table 2. Based on the Activity Analysis, we derived three distinct strategies used by non-expert Wikipedia translators.

| | Action ID | Action Description |
|---|---|---|
| C | 14 | Refine machine translation input |
| | 13 | Machine translate words/phrases |
| | 12 | Refine word/phrase in dictionary search |
| | 11 | Search word/phrase in dictionary |
| | 10 | Refer to knowledge resources (Wikipedia, online articles, publications…) |
| | 9 | Scan search engine results |
| | 8 | Refine search engine search terms |
| | 7 | Search engine query |
| B | 6 | Post-edit machine translated sentence in target article |
| | 5 | Compare machine translation result to original text |
| | 4 | Machine translate full sentences |
| | 3 | Read own translation (proofread) |
| | 2 | Translate sentence |
| | 1 | Read source article |
| A | -1 | Open resources (text editor, web pages…) |
| | -2 | Create target language page |

**Table 2: Activity Analysis coding schema for Actions in Wikipedia article translation. (A) Tool management. (B) Sentence translation. (C) Information search.**

[9] Test of English as a Foreign Language: Internet-based test

[10] Test of English as a Foreign Language: Paper-based test

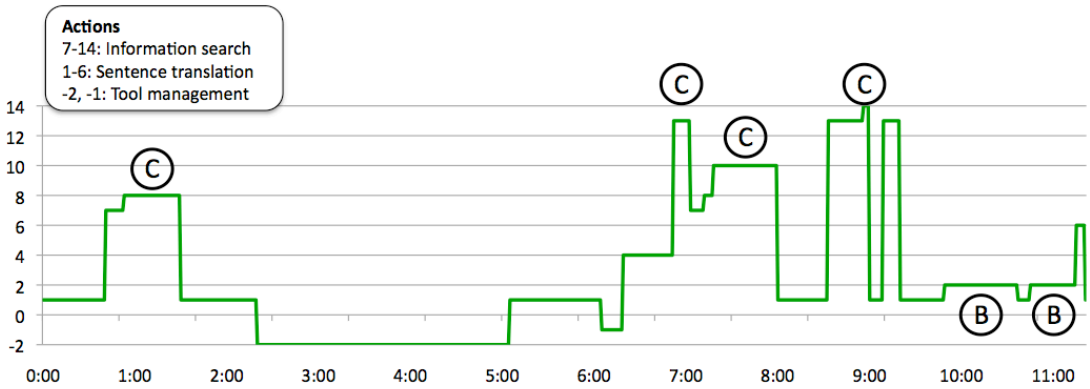


**Figure 1: Temporal action sequence of *Search First Then Translate* strategy. (C) Search for word translations, or information about source language words or phrases (B) Translate sentences based on previous search results in target page.**

Each participant was exclusively identified to adopt one of the three strategies in different stages of the article translation process. The identified translation strategies have some similarities with the translator types described in [7]. However, in [7], dictionaries and online resources were not allowed for the translators, and the strategies described in this paper are representative of individual non-expert Wikipedia translators.

In the following sections, the Wikipedia translation strategies are illustrated with an action sequence of three different translators translating the title and the first paragraph in the "Akan Volcanic Complex" article:

***Akan Volcanic Complex***
*Akan Volcanic Complex is a volcanic group of volcanoes that grew out of the Akan caldera. It is located within Akan National Park, about 50km Northwest of Kushiro in eastern Hokkaidō, Japan.*

*Search First Then Translate Strategy*
Figure 1 represents the action sequence for a translator type who searches unknown words in the source article first, then translates whole sentences or paragraphs. (C) indicates the information search actions during a translation task. In this example, the translator uses a search engine to find a definition for a word, machine translates words and phrases using Google Translate, and finally when the translator has resolved the meaning and Chinese translation for all the unknown words in the paragraph, conducts the sentence translation action (B).

*Translate-Search-Translate Strategy*
Figure 2 represents the action sequence for a translator type who translates sentences in increments, while looking for word translations or related information in between sentence translation actions. (C) indicates the actions during which the translator conducts a dictionary search, translates words in a machine translator or searches the meaning of an English word in a search engine. As illustrated in Figure 2, the information search actions are done in parallel with the translation of sentences (B). In other words, the translator would conduct an information search action whenever he/she encountered an unknown word in the source text while translating the same sentence.

*Paraphrase Machine Translation Strategy*
Figure 3 represents the action sequence for a translator type who paraphrases machine translation results as the main translation action. (B1) indicates the sequences where the translator uses a machine translation tool to translate complete sentences. (B2) indicates an action where the translator paraphrases the machine translation result to increase the quality of the translated sentence. The translator evaluates the machine translation quality during both actions, and conducts information searches for word translation or for additional information about the words or phrases when the machine translation result alone does not suffice for an accurate translation.

**Learning Effects on Choosing a Translation Strategy**
The translators spent the longest amount of time on actions related to sentence translation. These included reading the source article, paraphrasing machine translation results as well as translating the sentences by hand. Similarly to previous work conducted on professional translators [10], the non-expert Wikipedia translators required a wide range of tools to perform the task. In this experiment, translators spent on average 24.22% of total article translation time conducting information search actions with no significant difference between strategies (Table 3).

| *Activity* | *Actions* | *% of article translation* |
|---|---|---|
| Sentence translation | (1-6) | 66.46% |
| Information search | (7-14) | 24.22% |
| Tool management | -1, -2 | 9.32% |

**Table 3: Time spent on different activities during Wikipedia article translation.**

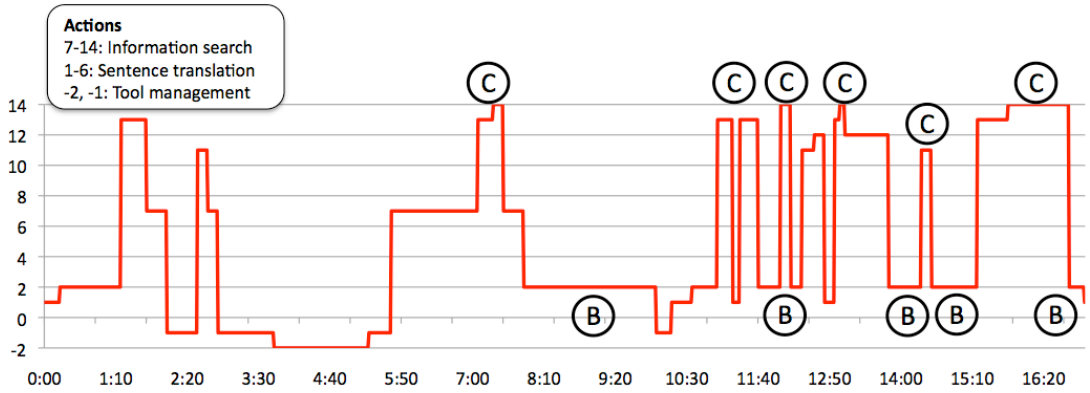


**Figure 2: Temporal action sequence of *Translate-Search-Translate* strategy. (C) Search for word translations, or information about source language words or phrases (B) Translate sentences based on previous search results in target page.**

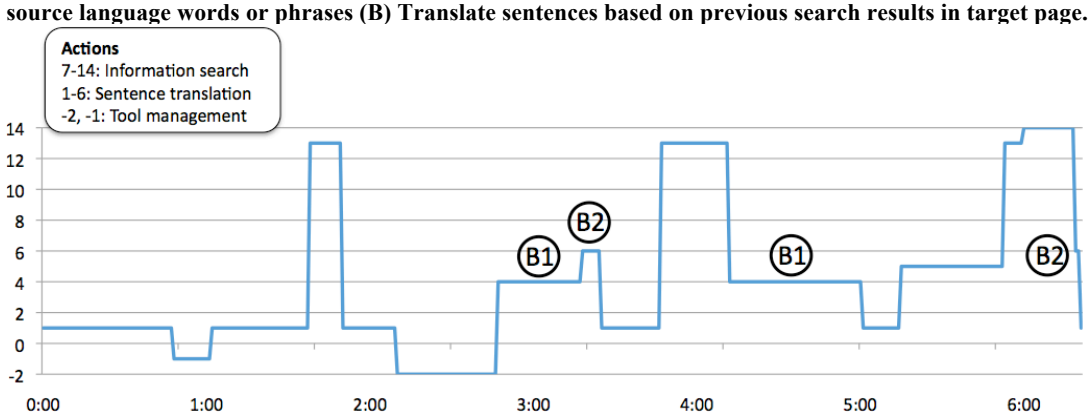


**Figure 3: Temporal action sequence of *Paraphrase Machine Translation* strategy. (B1) Translate full sentences in a machine translator. (B2) Paraphrase machine translation result in target page.**

There was no distinguishable trend in the choice of the initial translation strategy, and five of the participants did not modify their strategy during the translation task. However, three translators of the "Crowdsourcing" article began with *Translate-Search-Translate* strategy and later modified their strategy to *Paraphrase Machine Translation*. Two of the translators began with *Paraphrase Machine Translation* and did not modify their initial strategy (Table 4).

Based on the Activity Analysis, and the source article content, it became apparent that the "Crowdsourcing" article included longer paragraphs and more complete sentences than the other two articles. Furthermore, it was faster for translators to paraphrase the full machine translated sentences, which did not include a large amount of specialized words, or names and proper nouns. "Akan Volcanic Complex" and "Ying Huang" included over 40 words that could be considered specialized (incl. names and proper nouns), whereas the "Crowdsourcing" article included only 17 specialized words. Based on these observations, we concluded that *Paraphrase machine translation* can be a more desirable strategy for translating conceptual articles than biographical articles in Wikipedia, even though the English language might be more complex [31]. The effectiveness of each translation strategy could be determined from the translation time and quality of the translated article, but it is out of the scope of this paper.

| Translator ID | Translation Task | Initial Strategy | Modified Strategy |
|---|---|---|---|
| 1 | Akan Volcanic Complex | Paraphrase MT | Search First |
| 2 | Akan Volcanic Complex | Search First | Search First |
| 3 | Akan Volcanic Complex | Search First | Paraphrase MT |
| 4 | Akan Volcanic Complex | Search First | Translate-Search |
| 5 | Akan Volcanic Complex | Translate-Search | Translate-Search |
| 6 | Ying Huang | Paraphrase MT | Paraphrase MT |
| 7 | Ying Huang | Paraphrase MT | Paraphrase MT |
| 8 | Ying Huang | Translate-Search | Translate-Search |
| 9 | Ying Huang | Paraphrase MT | Translate-Search |
| 10 | Ying Huang | Translate-Search | Translate-Search |
| 11 | Crowdsourcing | Translate-Search | Paraphrase MT |
| 12 | Crowdsourcing | Paraphrase MT | Paraphrase MT |
| 13 | Crowdsourcing | Translate-Search | Paraphrase MT |
| 14 | Crowdsourcing | Paraphrase MT | Paraphrase MT |
| 15 | Crowdsourcing | Translate-Search | Paraphrase MT |

**Table 4: Learning effects in translation strategy choice.**

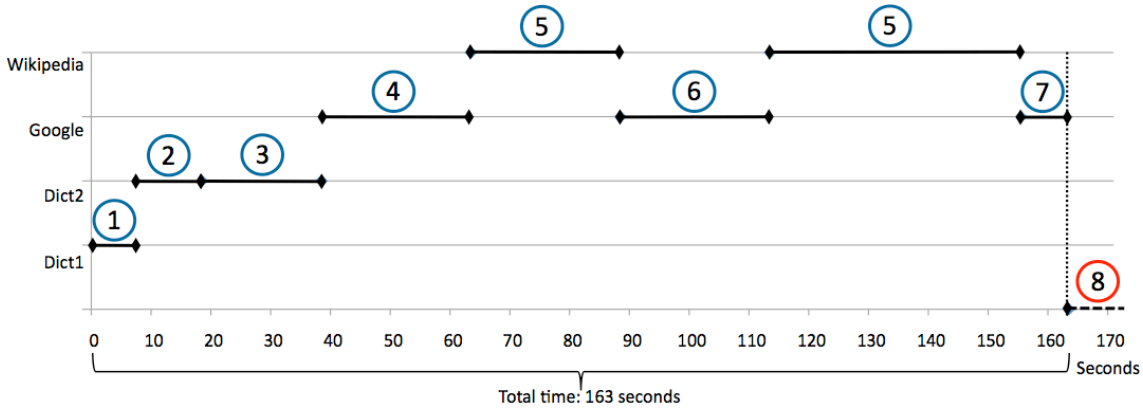


**Figure 4: Temporal pattern of information search for Chinese translation of "*crowdfunding*". (1) Search "*crowdfunding*" in Dictionary 1. (2) Search "*crowdfunding*" in Dictionary 2. (3) Paraphrase search term to "*crowdfund*". (4) Search "*crowdfunding*" in Google. (5) Read English Wikipedia article on "Crowd Funding". (6) Scan Google search results for "*crowdfunding*". (7) Search "*collective cooperation*" copied from "Crowd Funding" article in Google. (8) Read article without resolving the Chinese translation.**

## INFORMATION SEARCH DURING A WIKIPEDIA TRANSLATION TASK

As described in [25], activities can become actions, or actions can become activities when the goals change. In the Activity Analysis of the Wikipedia translation task, information search was denoted as a series of actions belonging to the translation activity (Table 2, (C)). In this section, we analyze information search as a separate activity, where the *object* of the activity is one information item (e.g., word or phrase).

For information search activity the following are defined:

**Activity** – Information search: Search a Chinese translation of a word, phrase, name or a proper noun from online resources. Search additional information about a word or a phrase being translated.

**Object** – Information object: Chinese translation of a word or a phrase; definition of a word or a phrase; transliteration of an English name or a proper noun in Chinese.

**Artifacts** – Online language and search tools: Search engines in Chinese and English, online multilingual dictionaries, machine translators (for word or phrase translation), Wikipedia, Baidu Baike, online publications.

The precondition to engage in an information search activity was that the first tool used would not produce a result valued high enough by the translator (no result, clearly wrong result, too many results). The translator would then choose another tool, modify the search terms, or finally try to find an explanation in the source or target language on the meaning of the original word (i.e., try to find enough additional information to successfully translate the given word when no direct translation could be found).

The example in Figure 4 is extracted from the "Crowdsourcing" article translation, and it represents a typical case of information search activity during a Wikipedia article translation task. The translator begins the information search activity by consulting a dictionary for direct word translation for "*crowdfunding*". When no result is found, the translator searches the same word from another dictionary without a result. The translator then paraphrases the search term by shortening it to "*crowdfund*". When this does not yield a result, the translator searches for additional information on the topic in English through Google and Wikipedia. At the end of the activity, the translator is unable to find a proper translation for the word, or a satisfactory definition on the meaning in the English term.

A simple case of information search *action* found through the Activity Analysis was the search of a single word in a dictionary, or through a machine translator. On average, the translators were engaged 23.32% of the time in single search actions, and 76.67% of the time in information search activities for one information item, such as one word.

Similarly to [14], the transliteration of English names to Chinese proved to be time consuming for the translators of "Akan Volcanic Complex" and "Ying Huang". These two articles included place names, proper nouns and names of people, which could not be directly translated with available online dictionaries or machine translators. When transliterating English names and proper nouns to Chinese, translators need to choose characters with similar phonemic representation [28, 29]. The non-expert translators, however, were unwilling to transliterate the names and proper nouns unless they found a reference through an information search activity. In cases where direct equivalent was not found, the translators gave up on the information search and used the non-translated English word directly in the translated Chinese language article.

The Activity Analysis revealed that non-expert translators conduct short sporadic word searches in dictionaries or machine translators, as is apparent especially in the *Translation-Search-Translate* strategy (Figure 2). The key finding here is that non-expert translators tend to engage in the time-consuming information search activities only when

the first initially highly valued tool returns an unsatisfactory result.

## DESIGNING A TRANSLATION SUPPORT SYSTEM FOR WIKIPEDIA

In this section, we focus on the derived design implications specifically for a *Wikipedia translation support system*, and illustrate how Activity Analysis was applied for designing a collaboratively edited multilingual dictionary tool for supporting complex translation tasks in Wikipedia.

### Design for Multiple Activities

Wikipedia is essentially a collective intelligence platform, where users engage in multiple types of activities [9]. In previous approaches to support Wikipedia translation, the focus has been on mechanical translation, often with the help of machine translation (e.g., [23]). However, as we reflected on the actions taken by the participants during a Wikipedia article translation activity (Table 2) and the translation strategies of individual translators (Figures 1, 2 and 3), it became obvious that article translation includes multiple activities in sequences, not just mechanical translation.

Based on this observation, we defined the first requirement for our prototype as "support multiple activities, rather than just mechanical translation". Furthermore, rather than adapting to separate activities, we defined a system that has the same functionality whether the user is reading the source article, translating an article or proofreading the translated text.

### Design for Continuous Action

As can be seen in Figures 1, 2 and 3, all translation strategies require the translators to hop between multiple *types* of tools when searching for information (dictionaries, machine translators, search engines and information resources). The second requirement for our prototype was that it is "usable in any Wikipedia page, and only in any Wikipedia page". Furthermore, we observed that users tend to spend a lot of time changing between actions, such as reading an article to viewing a dictionary. Hence, we defined a requirement stating "in order to view a dictionary entry, the current action/activity should not be interrupted".

A solution derived from the Activity Analysis is a pop-up page element reacting to mouse-over operations, displaying existing multilingual dictionary entries for the words in a Wikipedia article body text. The designed page element was defined to include an added feature to support information search by displaying search engine results using the highlighted word and it's existing translations as search terms. The first two design implications are aimed to lower the number of interactions needed to access the multilingual information online during different translation activities in Wikipedia, such as in information search activities (Figure 4).

### Design for Context Awareness

The Activity Analysis revealed, that even when using the *Paraphrase Machine Translation* strategy (Figure 3), translators had to search word translations from other information resources (Figure 4). This is partly due to the fact that, currently, machine translators do not handle the context of the translated sentence very well, and often give inappropriate word translations as a result (word polysemy). For example, the word "*crowdfunding*" in Figure 4 can be translated in at least three different ways: 众筹 (zhong chou), 集体融资 (ji ti rong zi), and 大众集资 (da zhong ji zi). In this case, the first option is the correct one given the domain of the article, while the last two are technically not incorrect, but less appropriate.

Based on this observation and the frequency of information search activities, we defined a requirement for our prototype to "include only one translation for each word given the context of the article". We define the context of a dictionary entry by using category-tags associated with an article being translated and the tags added to each domain specific dictionary. Furthermore, the inter-wiki links associated with category pages are used to preserve the context in the multilingual articles. Each entry is displayed with additional information on the source dictionary (name) and domain (categories) in order for users to determine whether an entry is appropriate given the context of the article. We defined that conflicts in the context specific dictionary entries are to be resolved by the community to keep the human in the loop, as mentioned in [17].

### Design for Knowledge Sharing

It should be taken in to consideration how human actions create value in multilingual collective intelligence platforms, such as Wikipedia. Even though direct collaboration is possible through the Wikipedia discussion pages [14], indirect collaboration is much more common [5]. The changes in the source and target articles, as well as any resources should be reflected to all collaborators in the multilingual and monolingual dimensions.

"Akan Volcanic Complex" and "Ying Huang" articles both included over 40 specialized words, whereas the "Crowdsourcing" article included 17 specialized words. Non-expert translators, as opposed to bilingual domain experts, need to put considerable effort on resolving all the correct word translations (e.g., Figure 4). Based on this observation, we defined a requirement for our prototype to "support reusability of the resolved word translations". In other words, support knowledge sharing by enabling article translators as well as domain experts to contribute translations of words to domain specific dictionaries.

Our prototype was designed so that any new unique multilingual dictionaries would be visible in any Wikipedia article in any language, thus making them available to translators as well as "Site Visitors" [9]. Furthermore, we included full history and versioning function to each

dictionary to foster collaboration in the dictionary creation process. In the nature of Wikipedia, we included a discussion function to each dictionary page, with a possibility to translate messages to other languages with machine translation.

By combining the previous design implications, we created a prototype of a user edited multilingual domain specific dictionary to support multiple translation activities in Wikipedia directly in the standard MediaWiki interface, which we shall report in a future paper.

## CONCLUSION

Wikipedia translation activities aim to increase the quality of the multilingual Wikipedia via article translation, especially in languages where the active contributors do not include domain experts in certain topics. However, previous studies have not answered one of the central questions in system design for Wikipedia translation support: *How* do individual Wikipedia contributors translate articles?

In this study, we identified three translation strategies of bilingual non-expert English to Chinese translators. We also found indications that paraphrasing machine translation results can be a more desirable strategy for translating conceptual articles than biographical articles for non-experts. This suggests that the current support systems may be effective only for certain type of articles or translators.

Translators engage in time-consuming information search activities when a translation of a word, especially proper nouns and names, cannot be resolved through the initial dictionary or machine translation service. The information needs of translators are met through multiple separate online resources, and translators need to put considerable effort into finding the relevant information online.

The results of the Activity Analysis were used to define design implications for a Wikipedia translation support system. The implications discussed in this paper included; i) design for multiple activities, not just mechanical translation, ii) design for continuous action in Wikipedia, rather than external tools iii) design for context awareness in the monolingual and multilingual dimensions of the Wikipedia article space, and iv) design for knowledge sharing to include domain experts in the collaborative translation. These guidelines were applied to a design of a support system for human translation activities in Wikipedia.

In order to support the needs of non-expert Wikipedia translators, we enable domain experts to work as resource providers to support non-expert translators through knowledge sharing via multilingual domain specific dictionaries directly in Wikipedia. We also included assisted information search about multilingual dictionary entries as one element of the supporting tool to reduce the time spent on information search activities, and to reduce interruptions during translation tasks.

### Limitations and Future Work

As this was an exploratory study on the actions taken by individual non-expert translators in Wikipedia, a number of limitations were identified. Although we collected a large amount of data, the number of participants could have been higher to better address individual differences. We found that the three articles chosen for this experiment were a good fit for a laboratory experiment, considering time and resource limitations, but were not representative of very long and elaborative Wikipedia articles. Furthermore, in this study we focused only on English to Chinese translation by bilingual participants. With different language pairs, and with translators from different cultural and linguistic backgrounds, the results might be different. Most notably, in certain language pairs, problems with transliteration of proper nouns and names are unlikely to occur.

This study is one part of a multidisciplinary approach to better understand the framework of Wikipedia translation in the multilingual and multicultural dimensions. In the future work, the developed support system will be introduced and evaluated in more detail. We will also observe how the activities identified in this study reflect to Wikipedia translation “in the wild”, and how our approach affects the work practices of real-life Wikipedia translators.

## ACKNOWLEDGMENTS

This research was partially supported by a Grant-in-Aid for Scientific Research (S) (24220002, 2012-2016) from Japan Society for the Promotion of Science (JSPS), and Kyoto University Global COE Program: Informatics Education and Research Center for Knowledge-Circulating Society, and Strategic Information, and Communications R&D. A special thanks to Linsi Xia, NTT Communications, for the help with Chinese translations.

## REFERENCES

1. Al-Dabbagh, U. The 'Explicitation Hypothesis' and nonprofessional translators: A case study. *IJAES, 9* (2008), 85-106.
2. Ballesteros, L. and Croft, W.B. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proc. SIGIR 1997*, ACM Press (1997), 84-91.
3. Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M. and Gergle, D. Omnipedia: Bridging the Wikipedia language gap. In *Proc. CHI 2012,* ACM (2012), 1075-1084.
4. Bardram, J. and Doryab, A. Activity Analysis - Applying activity theory to analyze complex work in hospitals. In *Proc. CSCW 2011*, ACM (2011), 455-464.
5. Bryant, S.L., Forte, A. and Bruckman, A. Becoming Wikipedian: Transformation of participation in a collaborative online encyclopedia. In *Proc. GROUP 2005*, ACM (2005), 1-10.

6. Callahan, E.S. and Herring, S.C. Cultural bias in Wikipedia content on famous persons. *J. Am. Soc. Inf. Sci. Technol*., 62, 10 (2011), 1899-1915.
7. Carl, M., Dragsted. B. and Jakobsen, A.L. A taxonomy of human translation styles. *Translation Theory 16, 2* (2011). http://www.translationjournal.net/journal/56taxonomy.htm.
8. Carl, M. and Buch-Kromann, M. Correlating translation product and translation process data of professional and student translators. In *Proc. EAMT 2010,* European Association for Machine Translation (2010). http://mt-archive.info/EAMT-2010-Carl.pdf.
9. Désilets, A., Gonzalez, L., Paquet, S. and Stojanovic, M. Translation the Wiki way. In *Proc. The 2006 International Symposium on Wikis 21, 23* (2006), 19-32.
10. Domas White, M., Matteson, M. and Abels, E.G. Beyond dictionaries: Understanding information behavior of professional translators. *Journal of Documentation 64, 4* (2008), 576 – 601.
11. Engeström, Y. Activity theory and individual and social transformation. In *(eds.) Engeström, Y., Miettinen, R. and Punamäki, R. L. Perspectives on Activity Theory,* Cambridge University Press (1999), 19-38.
12. Gordon, R.G. Jr. Ethnologue: Languages of the World, 15th ed., SIL International (2005).
13. Grudin, J. Why CSCW applications fail: Problems in the design and evaluation of organizational interfaces. In *Proc. CSCW 1988*, ACM (1988), 85-93.
14. Hautasaari, A. and Ishida, T. Analysis of discussion contributions in translated Wikipedia articles. In *Proc. ICIC 2012*, ACM (2012), 57-66.
15. Hautasaari, A., Takasaki, T., Nakaguchi, T., Koyama, J., Murakami, Y. and Ishida, T. Multi-language discussion platform for Wikipedia translation. In *(ed.) Ishida, T. The Language Grid - Service-Oriented Collective Intelligence for Language Resource Interoperability*, Springer (2011), 231-244.
16. Hecht, B. and Gergle, D. The tower of Babel meets web 2.0: User-generated content and its applications in a multilingual context. In *Proc. CHI 2010,* ACM (2010), 291-300.
17. Hutchins, J. Computer-based translation in Europe and North America, and its future prospects. *JAPIO 20th anniversary,* Japan Patent Information Organization (2005), 156-60.
18. Jakobsen, A. L. and Schou, L. Logging target text production with Translog. *Copenhagen Studies in Language, 24* (1999), 9-20.
19. Jääskeläinen, R. Focus on methodology in think-aloud studies on translating. In *(eds.) Tirkkonen-Condit, S. and Jääskeläinen, R. Tapping and Mapping the Processes of Translation and Interpreting: Outlooks on empirical research*, John Benjamins Publishing Company (2000), 71-82.
20. Jääskeläinen, R.H. Translation assignment in professional versus non-professional translation: Think-aloud protocol study. In *(Ed.)* S*e'quinot, C. The Translation Process*, H.G. Publications (1989), 87-98.
21. Jääskeläinen, R. Automatised processes in professional vs. non-professional translation: A think-aloud protocol study. In *(ed.) Künzli, A. Empirical Research into Translation and Interpreting: Processes and Products,* Neuchâtel: Université de Neuchâtel (1988), 89-109.
22. Kaptelinin, V. and Nardi, B.A. Acting with technology: Activity theory and interaction design. The MIT Press (2006).
23. Kumaran, A., Datha, N., Ashok, B., Saravanan, K., Ande, A., Sharma, A., Vedantham, S., Natampally, V., Dendi, V. and Maurice, S. WikiBABEL: A system for multilingual Wikipedia content. *AMTA Workshop* (2010). http://amta2010.amtaweb.org/AMTA/papers/7-01-04-KumaranEtal.pdf.
24. Kuutti, K. Activity theory as a potential framework for human-computer interaction research. In *(ed.) Nardi, B.A. Context and Consciousness: Activity Theory and Human-Computer Interaction*, MIT Press (1996), 17-44.
25. Leont'ev, A. The problem of activity in psychology. *Soviet Psychology 13, 2* (1974), 4-33.
26. Lörscher, W. The translation process: Methods and problems of its investigation. *Meta: Translators' Journal 50, 2* (2005), 597-608.
27. Lörscher, W. A. Psycholinguistic analysis of translation processes. *Meta: Translators' Journal 41, 1* (1996), 26-32.
28. Virga, P. and Khudanpur, S. Transliteration of proper names in cross-lingual information retrieval. In *Proc. MultiNER 2003,* ACL (2003), 57-64.
29. Wan, S. and Verspoor, C.M. Automatic English-Chinese name transliteration for development of multilingual resources. In *Proc. COLING 1998,* ACL (1998), 1352-1356.
30. Yamashita, N. and Ishida, T. Effects of machine translation on collaborative work. In *Proc. CSCW 2006,* ACM (2006), 515-524.
31. Yasseri, T., Kornai, A. and Kertesz, J. A practical approach to language complexity: A Wikipedia case study. *arXiv:1204.2765v1 [cs.CL]* (2012), 1-14. http://arxiv.org/pdf/1204.2765v2.pdf.
32. Zaidan, O.F. and Callison-Burc, C. Crowdsourcing translation: Professional quality from non-professionals. In *Proc. ACL HLT 2011*, ACL (2011), 1220–1229.