The Cross-Lingual Wiki Engine: Enabling Collaboration Across Language Barriers

Louis-Philippe Huberdeau
École de technologie supérieure
1100 Notre-Dame
Montreal, QC, H3C 1K3
Canada
lphuberdeau@ieee.org

Sébastien Paquet
TELUQ-UQÀM
100, rue Sherbrooke Ouest
Montréal, QC, H2X 3P2
Canada
sebpaq@gmail.com

Alain Désilets
National Research Council of Canada
Bldg M-50, Montreal Road
Ottawa, ON, K1A 0R6 Canada
alain.desilets@nrc-cnrc.gc.ca

ABSTRACT

In this paper, we present the Cross-Lingual Wiki Engine (CLWE), a system designed to support concurrent, collaborative authoring and translation of content in multiple languages. We start by showing how collaborative translation differs from conventional translation environments. In particular, we show how conventional industrial translation processes and tools are based on assumptions that often do not hold in collaborative environments. We then provide a detailed storyboard which shows how the CLWE can be used by groups of users, to collaboratively author and translate content without having to make those assumptions. We then discuss the implementation of the CLWE's change tracking infrastructure, which turns out to be the critical component in enabling this sort of open-ended translation workflow. We show how the problem of tracking changes in multiple languages at once can be greatly simplified using abstract change tokens which are independent of language and textual content. The system has been deployed in several communities, including SUMO (the Firefox documentation site), and preliminary feedback is encouraging.

Categories and Subject Descriptors

E.2 [**Data Storage Representation**]: [Data Storage Representation]; H.4 [**Information Systems Applications**]: [Information Systems Applications]; H.5 [**Information Interfaces and Presentation**]: [Information Interfaces and Presentation]; I.7 [**Document and Text Processing**]: [Document and Text Processing]

Keywords

wiki, collaborative translation, cross-lingual wiki engine, Tiki-Wiki CMS/Groupware, multilingual change tracking, cross-lingual collaboration

1. INTRODUCTION

Communication technology has made our planet smaller. Many of the challenges we tackle today are global in nature, and international collaboration is becoming the norm for many initiatives. Online collaboration generates vast quantities of textual information and increasingly involves people who are separated not only by geography, but also by language.

Fortunately, in this sort of initiative there are usually some participants who master several languages and are able to act as bridges between linguistic communities. In this context, the question arises of how to best enable online collaboration in spite of language barriers. In particular, we need to rethink how content and information is not only produced, but also how it is *translated*[3].

For example, groups of people can now collaboratively author and translate content in several languages concurrently, in an organic, continuous fashion. This new way of organizing translation work is very attractive for many community-built sites. For instance, support.mozilla.com (SUMO), the support site for Mozilla products, recently adopted a wiki approach in order to allow communities of volunteers to author documentation. One of SUMO's goals is to produce up-to-date documentation in at least eight major languages. Collaborative translation will be key to achieving this, and may even allow translation into less mainstream (and often neglected) languages, by providing linguistic minorities with tools they can use to collaboratively "help themselves". Other examples of communities that employ a collaborative translation paradigm have been identified[7].

Collaborative authoring and translation is becoming attractive for corporations as well. For instance, it allows them to crowdsource non-core translation work to communities of volunteers who care deeply about having content translated into a particular language[6, 13] (minority languages, for example). Even in completely conventional corporate translation contexts, teams of professional translators are also finding that this sort of collaborative, organic and agile approach to authoring and translation has definite advantages and may boost productivity[1].

It is worth noting that collaborative authoring environments are diverse and cover a wide range of situations. At one end of the spectrum we have systems used by small, closely-knit circles of collaborators. At the opposite end of the spectrum we find open, loosely structured online com-

munities consisting of large numbers of diverse people with a shared interest, who may come and go and contribute as their time allows.

Unfortunately, very few tools currently support collaborative translation effectively and reliably. Translating content in a collaborative context presents a number of unique technical challenges, compared to more conventional industrial[1] environments[5]. The primary difference is that in a collaborative environment, the process is much less controlled and may be more "chaotic". Figures 1 and 2 make this point visually by contrasting the flow of content in conventional translation processes with the more irregular patterns that may manifest themselves in a multilingual collaborative or community environment.

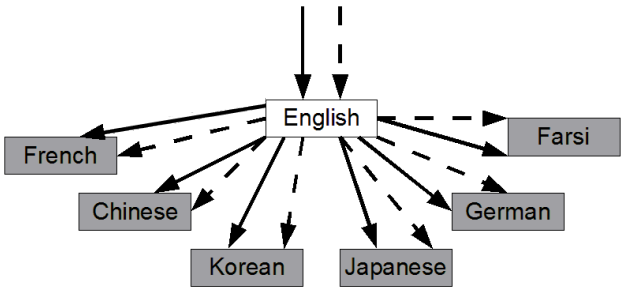


Figure 1: Content flow in a conventional industrial translation setting. Page creation (full arrow) and subsequent edits (dotted arrows) are first done in a master language, and then propagated to other languages.

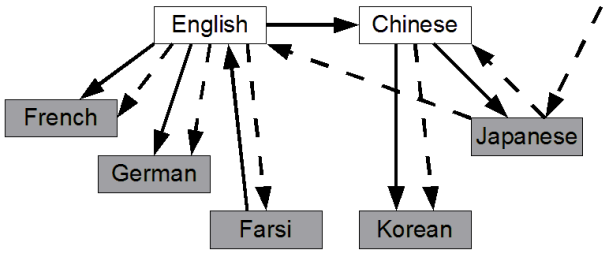


Figure 2: Content flow in a collaborative environment. Page creation (full arrows) and subsequent edits (dotted arrows) may first happen in any language, and may be propagated to other languages following arbitrary paths.

Conventional industrial translation processes and tools have been designed to operate well under a number of assumptions. In collaborative environments, many of those assumptions no longer hold and conventional methods break down.

Assumption 1 - Master language In a conventional environment, original content is often created in a master language, typically English. However, in many multilingual collaborative environments, many volunteer authors are not fluent enough in English to write high-quality content in that language. Collaborative tools thus need to be able to deal with situations where pieces of original content are spread across different linguistic versions of a page and must somehow be consolidated and propagated to all languages.

Assumption 2 - Edit freeze In a conventional environment, there is a strong tendency to refrain from modifying the master language version while translation is underway. In a collaborative environment, content is often in a permanent state of flux, and it is therefore not realistic to freeze it until translation in all languages is complete. Collaborative tools must thus support adaptation to continual changes in the source texts.

Assumption 3 - Enforceable timely translation In a conventional environment, timely translation of content is enforced through contractual or employment obligations. In collaborative environments, translators are often volunteers working on their own time, which may entail long translation delays. Collaborative tools must thus allow the publication of partially translated content, without misleading site visitors who read that content.

Assumption 4 - Controlled language pairs In a conventional environment, there is a tendency to restrict supported languages to a small list of "core" languages, and to limit the set of languages pairs for translation, typically to $P \Leftrightarrow X$, where P is a pivot language – often English – and X may be any other core language. Where the *Master language* assumption is made, unidirectional translation – strictly from the master language to target languages – is also imposed. By contrast, in a multilingual collaborative environment, members of the community may wish, and should be able to, create or translate content in any language, including minority languages; translation may occur between any pair of languages, and in any direction.

Assumption 5 - Strong coordination In a conventional environment, the community of authors and translators is a "closed" world, where some central authority can coordinate activities. In contrast, collaborative environments usually operate without central coordination. Therefore, tools must provide light coordination in the form of subtle cues that signal what translation work needs to be done, without necessarily mandating it.

Assumption 6 - Separation of Authoring and Translation In a conventional environment, authoring and translation are clearly segregated, and the two rarely interfere with each other. Authors do not have to worry about the translation process and translators need not be concerned with the authoring process. In a collaborative environment, it is usually more difficult to separate those two processes, and the same people are often involved in both. As a consequence, collaborative translation tools must integrate translation and

[1]Note that in this paper, we do not consider literary translation, which is an altogether different sort of activity, closer to artistic work, and usually performed by a "solitary" translator.

authoring without sacrificing simplicity in the authoring functionality.

Assumption 7 - Trained translators In a conventional context, translators are professionally trained, and can be socialized into the organization's tools, processes and linguistic norms. In a collaborative environment, translators are often amateurs, and the amount of tool, process and linguistic training that can be imposed on them is limited. Therefore, collaborative tools must be very simple to use, and must cater to the needs of amateur translators (for example, by including linguistic and terminology resources designed specifically to help amateurs avoid common translation mistakes).

In short, the main technological challenge of collaborative translation is to come up with tools and processes whose operation does not depend on the above assumptions. By lifting them even partially, we can move away from trying to control change, and move towards embracing it instead.

While lifting assumptions and constraints is helpful, one must also make sure that the process retains sufficient structure to allow tools to assist authors and translators in their work. Indeed, all of the above assumptions could be lifted trivially by creating a completely freeform tool where authors and translators are required to do everything manually (this, in a sense, is the approach that Wikipedia has taken for supporting cross-lingual content).

In this paper, we describe a tool called the Cross-Lingual Wiki Engine (CLWE), which lifts all of the above assumptions, while still offering sufficient structure to support effective collaboration. This system is based on TikiWiki CMS/Groupware[2], a fully-featured, open source content management system.

Although the system can support completely open-ended collaborative translation workflows, it may also be configured to support hybrid workflows that sit somewhere between conventional and completely open collaborative workflows. For example, the system may be configured to enforce a master or pivot language structure, or to provide a staging and approval process for ensuring the quality of contributions before their actual publication.

In this paper, however, we focus most of our attention on the completely open collaborative situation, because it is the most challenging and difficult case, and because no existing tool supports it efficiently. We also devote particular attention to a critical technical component of the system, namely, its simple but highly flexible model for tracking edits and translations in completely unconstrained collaborative workflows. Note that while we believe Machine Translation (MT) can play an important role in relaxing conventional workflow assumptions, this first version of CLWE does not include any MT features, and focuses only on the coordination of distributed *human translation* activities. However, plans for MT integration are described in section 6.

To our knowledge, CLWE is the first system to go this far in supporting collaborative authoring and translation of content, and to be usable in actual production settings.

The remainder of this paper is divided as follows. Section 2 presents the context and scope of the work. Section 3 surveys related efforts. Section 4, the heart of the paper, describes the tools we developed. Section 5 reports on actual use of our system. Finally, section 6 signals directions for future work and is followed by a conclusion.

[2]Website: http://tikiwiki.org

2. CONTEXT AND BACKGROUND

The work described in this paper is part of an open source project called the Cross-Lingual Wiki Engine (CLWE), which was started in the Fall of 2007.

This project aims to design, develop and test lightweight wiki tools that can be used to translate content in a collaborative, organic, wiki way. Our aim is to develop and evaluate processes and tools that may be applied in any wiki engine. However, we selected TikiWiki CMS/Groupware as our initial development platform, owing to the openness of its developer community to external contributions and its manifest commitment to multilingual support.

Around the same time, the Mozilla support community (SUMO) selected TikiWiki amongst a number of content management systems, to run the new support site for the Firefox browser. The knowledge base contained in the support site needed to be made available in multiple languages in order to reach a user base that is as large as possible.

In the context of the Cross-Lingual Wiki Engine Project, the SUMO knowledge base appeared to be an excellent primary test case, because of the large number of languages to be supported and the significant potential community of content and translation contributors.

3. RELATED WORK

Collaborative, wiki-style translation has raised a lot of interest in recent years. This has led to some academically published work, as well as relevant work by practitioners and wiki communities.

The system described in this paper builds heavily on prior work by Désilets et al.[5] and the ideas proposed by Huberdeau[8].

The LizzyWiki system presented in [5] removes dependence on many of the constraints and assumptions described in the introduction, but still relies on the following conventional assumptions: *Trained Translators*, *Separation of Authoring and Translation* and *Controlled language pairs*. The paper has a very strong focus on the needs of end users, and deals mostly with front-end and workflow design.

In his blog[3] , Huberdeau describes design principles for a backend that could support a relatively unconstrained translation workflow. The article introduces data management principles to allow original content modifications on any linguistic version and their orderly propagation to other languages, but does not discuss implementation nor front-end and workflow details.

One can think of the CLWE as an implementation of the backend design proposed by Huberdeau[8] combined with a generalization of the frontend and workflow design as per Désilets et al.[5]

Müldner et al.[14] describe a system called Cooperative Development of Internationalized Documents (CDIC), which looks at similar issues in the context of structured XML documents. Although it is not clear from the exposition in the paper, it seems that the system still assumes *Master languages*, *Edit freeze* of original content and clear *Separation of Authoring and Translation*.

[3]Website: http://blog.lphuberdeau.com

Other researchers have turned their attention to the collaborative localization of the *User Interface* of wiki engines[11]. This sort of work has also been done by practitioners in the TikiWiki community[12]. Our work has a different focus in that it deals with collaborative translation of the *actual content* of wiki sites.

Some researchers have investigated the collaborative creation of *linguistic resources* that can be used to help communities of translators[2, 3, 4]. Although this work has a very different focus from ours, it does contribute to lifting the *Trained translators* assumption.

Wikipedia publishes content in several languages, and has an active community of translators. The tools and workflows used by this community do not depend on any of the conventional assumptions. Instead, they provide different guidelines and various indicators that can be manually added by contributors[17]. However, they are so unstructured that they provide little in the way of automated or semi-automated support to help the community work efficiently. Our work differs in that it also lifts each of the conventional assumptions (at least partially), while still offering a good level of automation and support to assist authors and translators in their tasks.

Other sites have tackled collaborative translation using a more structured workflow and explicit system support for the task. For example, TraduWiki[4] supports collaborative, sentence-by-sentence translation of content available under Creative Commons. World Wide Lexicon[5] offers libraries to collaboratively edit and translate the content of websites in-place. DotSub[6] supports collaborative translation of text subtitles for movies. All three of those technologies assume that the content being translated has reached a final stage and will not change once translation has started. In other words they still rely heavily on the *Master language*, *Edit Freeze* and *Separation of Authoring and Translation* assumptions.

4. SUPPORTING COLLABORATIVE TRANSLATION

Our CLWE system allows communities to break out of the constrained mold imposed by conventional translation processes, and allows contributors to follow an open-ended workflow that is more consistent with modern collaborative environments. As pointed out earlier, the CLWE is very flexible and can support workflows that sit anywhere on the continuum between conventional workflows and completely open, collaborative ones. However in this paper, we focus our attention on supporting the completely open workflow, because that is the more challenging case, and it is a situation that existing tools do not support well.

A critical element of the system is its highly flexible, yet simple model for tracking edits and translations in completely unconstrained collaborative workflows. It enables the system to show contributors what translation work needs to be done, no matter how convoluted the prior sequence of edit and translation operations.

The fundamental technical insight behind this tracking model is that it treats edits as *abstract entities which are independent of language and actual textual elements.* This approach greatly simplifies the apparently intractable problem of tracking concurrent edits and translations in multiple languages. More details on this model will be provided in Section 4.4.

Besides being critical to support authors and translators in their collaborative work, this tracking mechanism also has the advantage of collecting information about collaborative translation behaviors. Such data might allow researchers to study the dynamics of translation communities in the future.

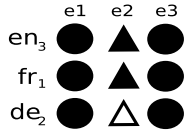
The present section has four parts. Section 4.1 establishes the vocabulary and notation which will be used to explain the tracking mechanism. Section 4.2 describes the system's functionality through a detailed usage storyboard. Section 4.3 describes how the system allows all assumptions of the conventional translation model to be partially or completely lifted. Section 4.4 explains some of the implementation details and their impact on the whole solution, including certain limitations.

4.1 Concepts and notation

Here we describe the fundamental building blocks at work in our change tracking model. As we have already mentioned, the basic concept is that of an *edit*, which is understood to mean a change that has been effected to the text of a page. An edit may correspond to several insertions, deletions, or modifications made to the page's content when going from one revision to the next. For our tracking purposes, the exact positions in the page where these changes occurred do not matter. Because wikis preserve the complete page history, only the version number is required to obtain the content and find out the actual textual changes that were made in a particular edit.

Edits always initially occur in a single language, but they may *propagate* to other languages through translation activity.

Throughout this section, we will use simple diagrams to describe the translation state of all the linguistic versions of a given page. For example, given a particular page and four distinct languages (English, French, Spanish and German), the diagram to the right can be used to represent their overall state. Each line represents a distinct linguistic version of the page, and each column corresponds to a unique edit. This diagram indicates that:



- The English page is currently at version 3, and includes three edits: e1, e2 and e3.
- The French version is at version 1, and incorporates the "same" three edits, albeit in the French language.
- The German version is at version 2 and only incorporates edits e1 and e3.
- The Spanish version does not yet exist.
- Edit e2, above the column of triangles, is a *"critical"* edit, meaning it should be translated into all other languages as soon as possible.

In this diagram notation, adding an original edit means adding a new column in the diagram, initially with all shapes hollow except for the language of the original edit. Propagating an edit through translation has the effect of filling

[4]Website: www.traduwiki.org
[5]Website: www.worldwidelexicon.org
[6]Website: www.dotsub.com

one or more shapes. In the ideal, "fully translated" state, all original edits created in any of the linguistic versions of the page have been propagated to all languages. Visually, this corresponds to a diagram that does not contain any hollow shape.

4.2 Storyboard

In order to illustrate how the CLWE supports unconstrained collaborative authoring and translation, we now provide a detailed usage storyboard. Our story involves three users (John Doe, Marie Quidam and Juan del Pueblo) who are collaboratively writing a workshop Call for Participation (CFP) in three languages – that is, on three distinct pages – simultaneously. John speaks English and French, Marie speaks French and English, and Juan speaks Spanish and French.

In our story, each of the participants has naturally assumed responsibility for the page in his or her first language, though this is not a constraint imposed by the system. John, Marie and Juan could just as easily be trilingual and share the responsibility for all three linguistic versions. For the sake of readability, all of them will use the English interface of the software, but again, this is not imposed by the system.

Our story will follow John, Marie and Juan as they concurrently edit and translate parts of the document in those three languages, and end up in a situation that looks like a hopeless mess. We will then show how the system allows them to easily and naturally recover from this situation – so easily, in fact, that one is left wondering whether there was even a "mess" to recover from in the first place.

Throughout the narration, we include inset diagrams which each indicate the state of translation *after* the scene where they appear.

Scene 1 Our story starts when John Doe writes basic information about the workshop on an English page. He does so over the course of two consecutive edits (see Fig. 3).

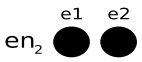


Figure 3: Original English version.

Scene 2 Responding to a page creation notification, Marie Quidam decides to start translating the English CFP to French, even though she can see that this is only a preliminary draft. She clicks on the *Translate* button (shown on Fig. 3), selects French as the target language, specifies a French name for the page, and hits the *Create Translation* button (Fig. 4). The system then displays the page in Fig. 5. Notice how the system automatically pasted the English page's contents into the edit box for this new French page, and also inserted a "Translation in progress" notice at the top. Marie gradually overwrites the English text with her French translation, then deletes the "Translation in progress" notice, and saves the page by hitting the *Complete Translation* button. This brings her to the page on Fig. 6.

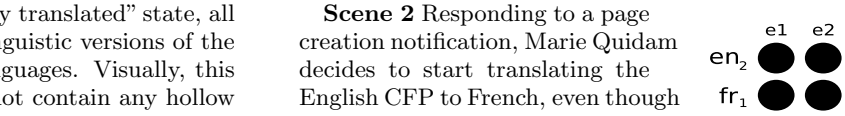


Figure 4: Creating a French translation.

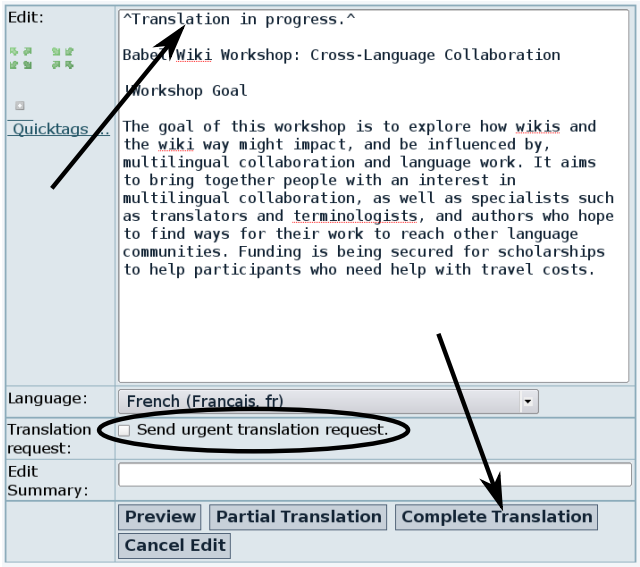


Figure 5: Translating initial English content to French.

A French page now exists, which is deemed to incorporate the same edits as the English page. This is reflected by the presence of a link to the English page in the "Page translation" box, and the fact that this English page is listed as being *equivalent*. Also, the French page is listed as being *100% up-to-date*. If an English reader were to go to the English page, he would see the reciprocal view, with a link to the French page instead.

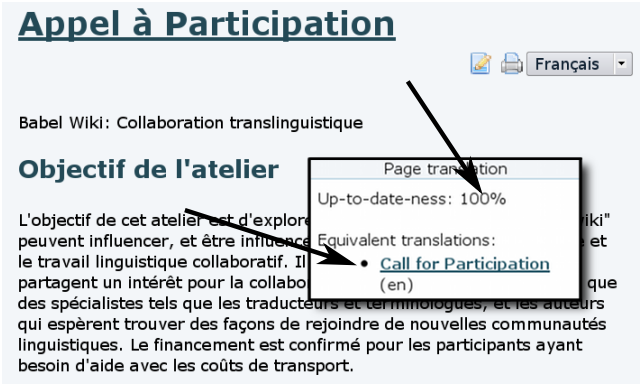
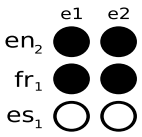


Figure 6: French page after initial translation from English. Note that in the actual system, the "Page translation" box is displayed to the right of page content.

Scene 3 Later on, Juan del Pueblo starts translating from the French page. Juan follows a process similar to Marie in *Scene 2*, except that the phone rings just as he finishes translating the second sentence. To avoid losing this work, he saves, but clicks the *Partial Translation* button since his translation is not complete. This brings him to the page on Fig. 7.

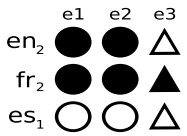
Figure 7: Spanish page after partial translation from French.

Notice how the *Controlled language pairs* assumption is lifted. Juan does not have to translate from the more mainstream English language, and can use whichever language he feels most comfortable with as a starting point. Also note that this initial decision does not constrain future translators of this page. If another Spanish-speaking translator decided to work on that page, he would be free to do so using any language as the source.



After Juan's effort, a visitor coming to the Spanish page would see a very conspicuous *"Translation in progress"* notice and would know to treat its content with caution (Fig. 7). He would also see a note in the "Page translation" box on the right telling him that, in contrast, the French and English pages are up-to-date, so that he might go read them instead, if he is fluent in one of those languages. Notice how those simple features contribute to lifting the *Enforceable timely translation* assumption, by providing a means to publish partially translated content without fear of misleading readers.

Scene 4 A few hours later, Marie Quidam decides to modify the budding CFP, and she does so in the French page. Marie adds a list of themes for the workshop, and deletes the last sentence about scholarships, because she feels it's premature to announce this until funding has actually been secured. As she feels the last point is really important, she checks the *Send urgent translation request* box (shown in Fig. 5) before saving her edit.



Note how both the *Master language* and the *Edit freeze* assumptions have been lifted. Although the CFP originally started life in the English language, Marie is allowed to make original new contributions to it in any language (French in this case). Also, even though the current version of the French page is still being translated to Spanish, Marie is allowed to modify the French page without having to wait for Spanish translation to be completed.

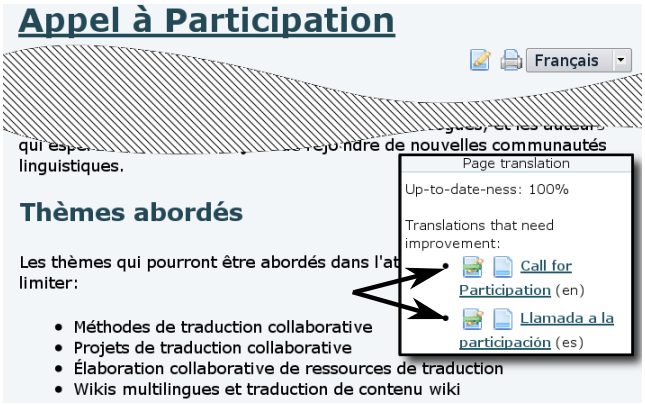
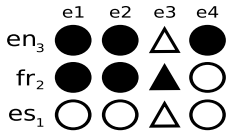


Figure 8: French page after creation of an urgent edit.

The French page is now the only one to be considered completely up-to-date, and this is reflected by the fact that both the English and Spanish pages are now listed as *needing improvement*, while the French page is considered to be *100% up-to-date* (see Fig. 8).

Scene 5 Moments later, John decides to enter the deadline information for paper submissions. When he goes to the English page, he sees that there is an urgent translation request (see Fig. 9). John decides not to translate the urgent request just yet, because his own edit will only take a few seconds and he



doesn't want to forget it. So he edits the English page, adds the deadline information and saves. This brings him to the page depicted on Fig. 10.

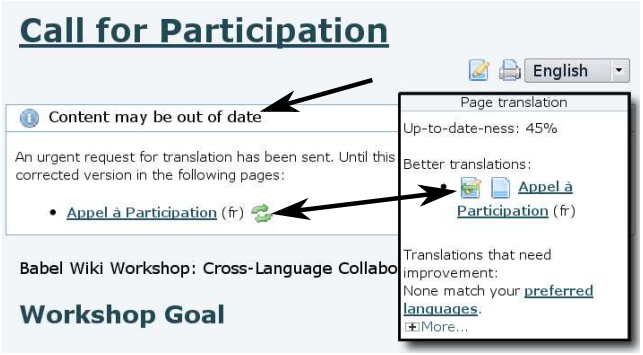


Figure 9: English page in need of urgent translation.

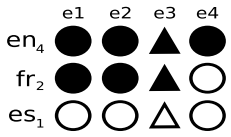
At this point, none of the three linguistic versions can be considered to be 100% up to date. The English and French pages each include edits that the other page has not yet incorporated, and in particular, English still hasn't incorporated the urgent edit from French. As for Spanish, it is still in the midst of incorporating the very first two edits based on the French page.

In short, it looks like John, Marie and Juan have buried themselves waist-deep in a mud pit. There does not seem to be a stable point from which all edits could easily and safely be propagated to all languages. None of the current linguistic versions of the page is fully up-to-date and can be used as a single source of information.

However, this is only an apparent mess, and as we shall see, the system allows them to fall back on their feet easily and effortlessly.

Figure 10: A small English edit is carried out before the urgent translation request.

Scene 6 First, John updates the English page by translating changes from the French version. He does so by clicking on the update icon next to the French page title (document icon with a left-pointing arrow shown in Fig. 10). The CLWE presents him with changes that were made to the French page (see Fig. 11), and John promptly reproduces them in the English page. After saving, the English page shows up without a critical translation warning and is listed as being *100% up-to-date* (see Fig. 12).



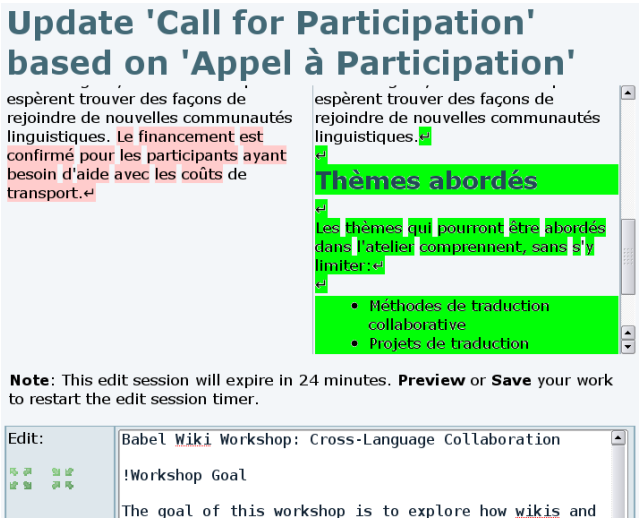


Figure 11: Translating edits (including an urgent one) from French to English.

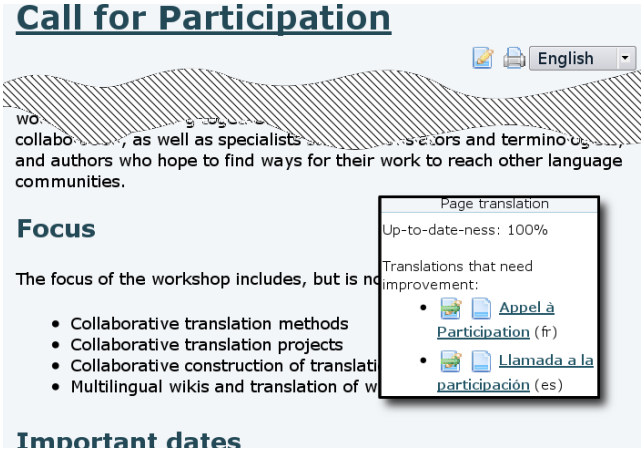


Figure 12: English is now up-to-date (including urgent edit).

Scene 7 Next, Marie updates French to incorporate the English edit about submission deadlines (see Fig. 13), and saves using the *Complete Translation* button. The French page is now also listed as being *100% up-to-date*.

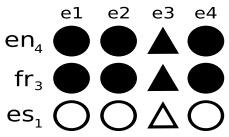
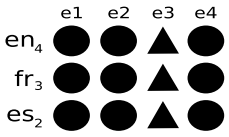


Figure 13: Translating remaining edits from English into French.

Here, we notice something odd. Indeed, in showing parts of the English text that need to be incorporated into French, the system highlights not only the deadline information which was originally created in English, but also the list of themes. In other words, Marie sees an English translation ofher *own original French edits* as needing to be translated to French. This limitation will be explained in section 4.4. For now, suffice it to say that this does not bother Marie. She can quickly tell that this is a part that does not need to be translated, and just focuses on translating the deadline information.

Scene 8 Finally, Juan comes back from his phone conversation and picks up his translation from French. He sees highlighted in green, not only the edits that he had seen when he first started translating, but also all other edits which have been done or replicated in French since then. Juan finishes translating all of the edits, deletes the "Translation in progress" notice, and saves as a *Complete Translation*.



Et voilà! All three pages now incorporate the exact same edits, and all are displayed as being *100% up-to-date*.

Note that the particular order in which John, Marie and Juan pulled themselves from the "mud" was not prescribed by the tool, and that they could have done it in a number of other ways. To be sure, the sequence of actions was conditioned by the collective linguistic competencies of the three actors. For example, because Juan was the only Spanish translator involved and he does not read English, content in that language could not be translated to Spanish directly, without first going through French. But this constraint was not actually imposed by the system and, if an additional English-Spanish translator had been involved, translation could indeed have proceeded directly from English to Spanish.

4.3 Lifting conventional assumptions

We now discuss how the CLWE helps lift conventional assumptions about authoring and translation workflow. The storyboard already illustrated how the system helps lift the following assumptions: *Master language*, *Edit freeze*, *Enforceable timely translation* and *Controlled language pairs*. We discuss these in more detail, as well as other assumptions which were not explicitly mentioned in the storyboard.

The architecture behind the CLWE completely removes the assumptions of *Master language* and *Edit freeze* by allowing the contributors to create original content without having to worry about the state of the various linguistic versions of a page. While they are made aware of that state through various indications in the user interface, no additional constraint is imposed.

The *Enforceable timely translation* assumption is also mostly lifted by providing readers of the site with information that they can use to make intelligent decisions in situations where a particular page is out of date compared with other linguistic versions. This effectively means that information can be published on the site even if it may take a while before it is completely translated into all the languages supported by the site. However we don't consider this constraint to be completely lifted because visitors cannot get the most up to date information unless they can read one of the languages whose version of the page is completely up to date.

The *Controlled language pairs* assumption is also lifted, at least partially. Members of any linguistic community are free to translate pages into their own language, and the system provides them with the indications they need to maintain the content in that language in sync with other languages. However, this can only work if the community of translators for that language is large enough to keep up with the pace of changes in other languages. The CLWE does not currently address this issue directly, but Section 6 describes how integration of Machine Translation features could help alleviate it.

Although this was not mentioned in the storyboard, we can see how the *Strong coordination* assumption is completely lifted, because at no point in the process did our three actors need to communicate with each other or with any central "supervising" authority. Coordination is achieved implicitly through intuitive notices that act as invitations (as opposed to commands) to take a particular action.

Regarding the *Trained translators* assumption, we can see that users need not be trained to follow a rigid authoring or translation process, and that the tool is intuitive enough to be used without extensive training. However, even if the system adds as little complexity as possible, translation remains in itself a complex activity. The CLWE does not lift this assumption completely, since it does not provide any special linguistic resources (glossaries, translation memories) that might be needed by non-professional translators in their work.

Finally, regarding *Separation of Authoring and Translation*, we can see how the system supports easy transition from authoring to translating, and vice versa. Moreover, the editing operations are not complicated by the integration of the translation process and tools: authors may still edit content largely without worrying about where and when translation will occur. However, this particular assumption is also not completely lifted. Although this was not apparent in the storyboard, the system assumes that users will never

carry out original edits while in the course of doing translation work. (This limitation is discussed in more depth in section 4.4.)

It is worth noting that although our three storyboard users organized themselves into a structure where French acted as a pivot language (Fig. 14a), this was never imposed by the system. In practice, each community of users is free to adopt any translation structure it sees fit. For example, had Juan been able to read English, or had a fourth actor been able to translate directly between English and Spanish, the community might have naturally gravitated towards a clique structure where translation can take place between any of the three language pairs (Fig 14b). Even if the pathways change over time, the system does not require any adjustment. This flexibility is a major advantage of the CLWE approach.

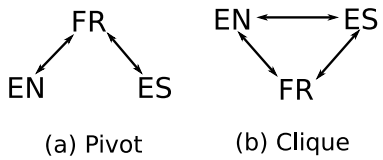


Figure 14: Different organizational structures for collaborative translation

It is also worth noting that the CLWE does not impose a strict (e.g., sentence-level) correspondence on the content of linguistic versions of a page. For example, if John adds a sentence to the English page, and Marie decides that it is not relevant for a French audience, she can simply choose to not reproduce that sentence in French, and still press the *Complete Translation* button. The system will then consider John's sentence to have been "dealt with" in French, even though Marie decided not to translate it. Note, however, that in this kind of "cultural adaptation" situation, a distortion of the content might occur when Juan translates from French. Indeed, in this scenario, Juan would never see John's original sentence, even though it might be appropriate for a Spanish audience.

4.4 Implementation

A key element for supporting the open-ended workflow that we describe is a backend capable of tracking edits and translations made in the different pages, in such a way that it can help users efficiently propagate them into other linguistic versions of those pages. Here we present an overview of the important concepts we combined to address this challenge; a complete description of the theory behind the tracking engine and the implementation details are available in the architecture document[9]. We will also discuss certain limitations of this implementation, some of which have already been alluded to in Section 4.2.

Tracking changes. As mentioned before, the main technical insight behind this model is that edits can be treated as *abstract entities which are independent of language and actual textual elements.* Whenever an original edit is made by a user on a page, a unique token is generated to stand for that edit, and added to an *edit set* which represents the state of that page. The particular revision where the edit occurred is also linked to the token. When a target page in one language is updated based on a source page in an other language, all missing edit tokens from the source page's edit set are added (*propagated*) to the target page's edit set.

This simple tracking model allows the system to easily identify which linguistic versions need updating, without having to actually analyze their actual textual content: a page needs updating whenever it is missing some edit tokens. By simply comparing their edit sets, different linguistic versions can be compared, irrespective of the order in which edits actually occurred. For any two edit sets α and τ, one the following holds:

- $\alpha = \tau$: means the pages are equivalent
- $\alpha \subset \tau$: means α can be updated from τ (τ is more complete)
- $\tau \subset \alpha$: means τ can be updated from α ($\alpha$is more complete)
- Otherwise: means the pages need updating from each other

This last case means that a page can both be updated and be used to update the other page; it occurs when each page includes edits that its counterpart hasn't yet incorporated.

The simplicity and tractability of this model is what allows CLWE to support collaborative translation without having to impose constraints like *Master language* and *Edit freeze*. All the same, it helps users make sure that no change is lost in the translation process, no matter how convoluted the chain of edits and translations. It can also be used to assist readers of the site by telling them when a page may be missing important information, and where they might find this information in more up-to-date linguistic versions of that page.

Presenting differences. Although the above formulae allow CLWE to know which page need translation work done, and which linguistic versions they could be updated from, they do not say anything about *which parts of the text* need to be translated. Of course, if we are to help users translate an edit between languages, we need to be able to show this text. This information can be retrieved using the standard page revision history provided by most wikis, and for the purpose of tracking the translation state, we can ignore these details until a user actually gets down to translating a particular edit.

In presenting textual elements that need to be translated, the system tries to select the "best" possible text difference from the stored edit and translation history. Fig. 15a illustrates how this is done at the beginning of *Scene 7* where edit e_4 needs to be translated from English to French. Basically, the system computes the textual difference between the current English version (here, version en_4) against the most recent English version that is missing e_4 (i.e. version en_2). Note that another reasonable strategy might be to compare the first English version where e_4 appeared (i.e. version en_3), against the English version that immediately preceded it (i.e. version en_2). This alternative strategy is illustrated in Fig. 15b. Reasons for choosing the first strategy over the second will be discussed later in this section, when we discuss limitations of the system.

Note how our diff strategy does not depend on the existence of an earlier synchronization point between source and target pages. For example, in Fig. 15a, we can see that the

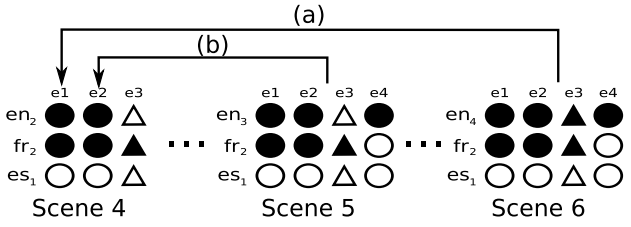


Figure 15: Two possible diff strategies for identifying textual changes in edit e4. (a) starting from most recent version, and (b) starting from first version where e4 appeared.

English version we diff against (version en_2) was not completely in sync with its French counterpart at the time. All that really matters is that version en_2 is the latest version of the source page that does not include edit e_4 in English. As a result, this approach can support a very large number of languages and translation paths.

Computing "up-to-dateness". The actual text content difference is also used to compute the percentage of "up-to-dateness" which is displayed on each page. These percentages have a dual purpose. On the one hand, they are meant to give the *reader* an idea of how much untranslated information is accessible in other linguistic versions of the page. On the other hand, they are also meant to give *translators* an idea of how much translation work needs to be done to bring the page in sync with its linguistic counterparts.

Finding an accurate measure for this is no easy task. In natural languages, a single character (e.g. adding an "s" to pluralize a word) can completely change the meaning of a sentence or even whole paragraphs. For this reason, any measure is bound to be imprecise and can only be used as an estimate. But the important thing is that it provide readers and translators with a reasonable idea of up-to-dateness and translation effort required.

After some trial and error, we adopted the following formula to compute up-to-dateness:

$$\begin{aligned} ST &= \#\,of\ sentences\ in\ target\ page \\ IS &= \#\,of\ original\ sentences\ inserted\ in\ any\ language \\ DS &= \#\,of\ original\ sentences\ deleted\ from\ any\ language \\ ratio &= \frac{ST}{ST + IS + \frac{1}{10}DS} \end{aligned}$$

This formula accounts for modifications in terms of insertions and deletions. When a sentence is *modified* (i.e., a few of its words are changed), it will be counted as a deletion of the original sentence, followed by an insertion of the modified sentence. Also note that insertions are weighted more heavily than deletions, because deleting a sentence represents less effort for translators than translating a new one. Finally, we opted to count the number of *sentences* that have changed, as opposed to say, words, characters, or paragraphs. This seems reasonable given that we are trying to measure up-to-dateness of the *informational content* and that "one idea per sentence" is a commonly used authoring guideline. This choice was confirmed by initial tests which demonstrated that other techniques would make up-to-dateness percentages too large or too small. More details about the measure are available on the project website[15] and in the complete report[10].

Limitations. The simple tracking and text differencing techniques that we described here turn out to work well in most circumstances, but they do present limitations which we now discuss. Preliminary user feedback on the system suggests that those limitations may not be problematic in actual practice. Also, Section 6 discusses possible solutions to those limitations.

Limitation 1 Erroneous artifacts may be displayed as part of the text differencing process.

As could be seen in *Scene 7*, where changes were shown to the translator that were already incorporated in the target page, our differencing strategy may lead to erroneous artifacts in the highlighted text.

While the CLWE is able to easily track when a page in one language needs updating from another language, identifying the exact textual elements that need to be reproduced is much trickier. This is a direct consequence of lifting the *Edit freeze* assumption. Indeed, by the time a user decides to translate a particular edit from say, English to French, the English page may have incorporated several other edits or translations. Some of those may already exist in the French page, either having originated there or propagated into French from other linguistic versions.

It would be easy enough for the CLWE to isolate an original English edit by comparing the English revision where the edit first appeared to the previous English version. This corresponds to the strategy depicted in Fig. 15b. In the specific case of *Scene 7*, this would have indeed isolated English edit e4 which Marie was trying to reproduce in French. The problem with this strategy, however, is that it compares two versions which might be very old. In our experience, users have difficulty making sense of such "old" differences, in the context of the most recent versions of the English and French pages, because by then, English and French may have integrated many other changes. Our preliminary experience with these two differencing strategies leads us to believe that differencing against the most recent version of the source language creates less confusion for end users, but more work needs to be done to confirm this in real use situations.

Limitation 2 The CLWE relies on human translators to correctly signal whether or not they have performed an accurate and complete translation.

It should be apparent by now that when the user saves a translation, the system has no way of knowing which edits were *actually reproduced* by the user in the target language. In particular, if the user did not complete the translation, the system relies on him to click on the *Partial Translation* button instead of *Complete Translation*, and assumes that no edit was propagated.

This is a limitation of the approach which could potentially result in loss of edits or mislabeling of pages as being out of date. As an example, imagine that Marie is translating an edit from English to French and she inadvertently forgets to translate a sentence, but still hits the *Complete Translation* button. The net result is that this particular sentence will be lost to the French alternative, as well as to any linguistic version that subsequently translates from French.

Conversely, if Marie inadvertently hits the *Partial Translation* button when she did in fact translate all the sentences that required translation, then the French page will be labeled as needing translation. If another user attempts to update it later on, he may be confused and spend a fair amount of time inspecting the highlighted sentences before concluding that the page was, in fact, already up-to-date.

Finally, as we pointed out at the end of Section 4.2, the "freedom in translation" offered by the system may in some circumstances let content loss occur in "cultural adaptation" scenarios.

Limitation 3 Translators should not make original edits while translating, but the system is unable to prevent this.

Imagine a situation where Juan is carrying out a translation task and he suddenly notices an important factual error in the text. In a situation like this, chances are that he will want to correct this mistake right then, while still in the middle of the translation dialog. But with our current implementation, Juan must abstain from doing this. Instead, he must first complete and save the translation, and only then can he make an original edit to fix the mistake.

The reason for this is that the system has no way of telling for sure if a particular textual change corresponds to an original edit or a translation. Consequently, our CLWE system simply assumes that changes made from inside an edit dialog correspond to original edits, while changes made from inside a translation dialog correspond to translation of original edits. Therefore, if Juan does an original edit while in a translation dialog, his change will be taken as part of a translation rather than generating an edit token, and the system will never notify Juan or other users that this change needs to be propagated to other languages.

Limitation 4 The up-to-dateness measure is not fully accurate.

While our up-to-dateness measure provides reasonable values in most circumstances, it is admittedly not sophisticated. For example, on a short page, the replacement of a single word by a synonym will cause the page to be labeled as seriously out of date, even though its information content is in fact mostly up-to-date.

5. EVALUATION

The storyboard presented in Section 4.2 serves as a kind of cognitive walkthrough[7] for the system. Cognitive walkthrough is a simple usability technique where one thinks carefully and systematically through the steps that users must take to accomplish relevant tasks with the system. Our storyboard indicates that the important user tasks can be carried out with the system, without reliance on conventional assumptions about authoring and translation processes. The storyboard also allowed us to identify certain minor problems with our approach.

Of course, in the wiki world, the proof of the pudding is in the eating, and no amount of cognitive walking through will prove that the system is actually usable in practice. At the moment of writing, CLWE had just been deployed in a number of communities:

- TikiDoc (doc.tikiwiki.org): the community that writes user documentation for the TikiWiki system.
- Tiki for Smarties (twbasics.keycontent.org): a site providing tutorials on TikiWiki.
- JIAMCATT demo site (jiamcatt.ourwiki.net): a demo site presented at the JIAMCATT conference, where attendees could collaboratively create multilingual content.
- SUMO (support.mozilla.com): the Firefox documentation site.
- Global Voices (www.globalvoices.org): a site that aggregates and translates blog postings worldwide.

Moreover, the CLWE has been continuously put to the test through its various iterations for several months on the site wiki-translation.com, where it has, among other things, served to create and update a workshop call for participation in three languages[8]. Although preliminary feedback from various user bases has been overwhelmingly positive, we do not feel at this point that we have sufficient data on use of the system to make strong claims about its usefulness in actual operational situations; this is left for future work.

6. FUTURE WORK

The work described in this paper constitutes a very significant advance in support for collaborative authoring and translation, and preliminary user feedback indicates that it is already usable as is.

However, one can easily think of additional work that could be done to improve it. In order to avoid implementing features that turn out to not be really useful in the end, our plan is to follow an incremental approach and implement only those improvements which we find are needed, based on feedback from our pilot users. Below is a discussion of some possible directions in which this might take us.

6.1 Evaluate use on pilot sites

The first step is, of course, to deploy the system on a number of pilot sites like the ones mentioned in the previous section, then gather and analyze feedback and usage data to evaluate the system.

An interesting question is how communities will organize themselves in a context where the tool imposes virtually no limits on the translation workflow. For example, will they tend to naturally evolve towards the use of a single pivot language, even though the system does not impose such a structure? Similarly, will communities tend to write original content in English first, even though the system does not impose a master language? While these are plausible outcomes, one cannot predict for sure that this is what will happen. Another very likely scenario is that communities will evolve towards concurrently supporting more than one pivot language (say, English and Chinese), in order to better reach different geographical areas. Another possibility is that non-standard translation paths might coexist with pivotal ones within the same site, and that such non-standard paths turn out to be critical for reaching certain minority languages.

[7] http://en.wikipedia.org/wiki/Cognitive_walkthrough

[8] http://wiki-translation.com/BabelWiki

The structure adopted by communities will in turn impact the length of translation chains, and it will be interesting to see how long they will tend to be, and whether longer chains result in significant distortion of the original message.

The tracking data collected by our CLWE system could easily be used to answer those questions, with custom-built analysis and visualization tools. This usage data may in turn help us identify key improvements to enhance the CLWE's support of collaborative translation and increase its adoption rate.

6.2 Better isolate textual changes

As pointed out in *Limitation 1*, our current text differencing strategy sometimes causes the system to show certain textual changes as needing to be translated, when they have, in fact, already been translated.

We plan to investigate alternative diff strategies (as per Fig. 15b) to isolate only untranslated changes, and combine this with patching strategies to display those changes in the context of the most recent version of the source text.

6.3 Decrease reliance on users for assessing translation completion

As pointed out in *Limitation 2*, the system currently relies heavily on the user to tell it when a particular translation task is complete. If the user mistakenly pushes the wrong button, this may result in changes not being propagated to other languages, or in substantial confusion for subsequent translators of the same page.

One way to alleviate this problem would be to use automatic bilingual sentence alignment technologies[16] to perform a basic sanity check on the alignment of the saved target page with the source page. The system could then notify the user when the alignment does not seem to correspond to his choice of *Complete Translation* versus *Partial Translation* button.

6.4 Prevent content loss in cultural adaptation situations

As pointed out also in *Limitation 2*, content loss may occur in cultural adaptation situations where a translator decides to not translate a particular part of an edit, because he sees it as being irrelevant to his particular linguistic audience. We could deal with this by providing the user with a *Cultural adaptation* button. For example, if Marie clicks on this button in the course of translating edits e_4 and e_5, the system would deem the translation to be complete and the French page would be labelled as being up-to-date as far as those particular edits are concerned. However, in the edit set for the French page, e_4 and e_5 would be labeled as being "non-propagatable", meaning that the system would never allow them to be propagated to another language from French. Instead, users would have to propagate those changes starting from other languages.

6.5 Prevent original contributions in the context of a translation transaction

As pointed out by *Limitation 3*, the system requires that users not mix translation and original contributions within the same transaction. In our limited experience using the system, this can be hard to do, especially when one notices an important mistake in the source text, while in the midst of translating it. Unfortunately, if a user makes an original edit while in the midst of a translation dialog, that original edit may never be propagated to other languages.

There does not seem to be an easy way to allow users to mix original edits and translations in the same transaction. However, we can constrain the translation user interface in such a way as to prevent the temptation. For example, instead of displaying the full text of the source page in an edit box, we could display most of it in a read-only text box, and only display those parts that need to be translated in editable text boxes.

This constrained user interface may also help track translations at a sentence-by-sentence level, which in turn may help perform sanity checks on translation alignments (as per the previous section). Or, it could be that conversely, automatic bilingual alignment technology is needed in order to identify which sentences the user should be able to edit in the target text (that is, which sentences in the target text correspond to changed sentences in the source text).

6.6 Experiment with alternative up-to-dateness measures

As pointed out in *Limitation 4*, the current measure for up-to-dateness is acceptable and provides useful information as-is. However, it is imprecise and could certainly be improved. Potential solutions include:

- Changing the unit used for counting changes and employ words or characters instead of sentences.
- Changing the insertion/deletion weights.
- Dynamically adapting the change counting unit as well as the insertion/deletion weights, based on the length of the page.
- Performing deeper content analysis to determine if an edit actually modified the meaning of a sentence.
- Presenting the measure graphically instead of numerically (ex: an up-to-dateness gauge) to better convey the imprecise nature of the value to the end user.

6.7 Integrate Machine Translation

The current implementation of the CLWE assumes that a particular community will have a sufficient critical mass of translators, to ensure timely translation of fast changing content to all the languages supported by the wiki site. But this may not always be the case, given that a site may choose to allow the creation and translation of content into any language (including some small, minority languages), and the fact that translation on such sites is typically done by volunteers, who might be in short supply. Also, although the system helps polyglot users find the most up to date version of a page among those languages he can read, it is of little assistance for unilingual users, or in situations where none of the up to date versions are in a language that the user can read.

To address these problems, we are thinking about integrating machine translation tools into the system, along the lines of what was proposed in [5], to allow translators to get the gist of original contributions written in languages that they cannot read. Machine translation could also be used to help unilingual site readers, by providing them with temporary automatic translations of those edits that have not yet been translated to their native language.

6.8 Incorporate translation management tools

Although the features described in this paper allow users to find out what translation work needs to be done for any given page, users have no way of easily assessing which pages, among all those on a given site, are in most need of translation work. To deal with this issue, we could implement simple reporting and visualization tools to help users answer questions such as:

- What urgent translation requests need to be fulfilled in my native language?
- What highly-visited pages in my native language are currently severely out of date?
- What's the average state of up-to-dateness for pages in my native language?

6.9 Replicate the method in other wiki engines

Finally, since the objective of the CLWE is to make collaborative translation widely available, it would make sense to try and reuse the concepts in other wiki engines, using the designs and lessons learned from our TikiWiki implementation.

7. CONCLUSION

We have presented the Cross-Lingual Wiki Engine (CLWE), a system designed to support concurrent, collaborative authoring and translation of content in multiple languages. The CLWE lifts (at least partially) all of the assumptions made by conventional translation tools. While still largely untested in practice, we believe it can efficiently support true collaborative translation, including in completely open environments and workflows. Using simple change tracking mechanisms, it provides the flexibility required for translators to choose their favorite source language, while letting content authors contribute in their native language independently of the translation process. Site visitors get an improved navigation experience by being able to read in their favorite language while staying aware of the evolution occuring in other linguistic versions. The system has already been deployed in several communities. Further work should assess more deeply and improve the system's usefulness across a range of multilingual collaboration scenarios.

8. REFERENCES

[1] Beninatto, R. S., and DePalma, D. A. Collaborative translation. In *Multilingual, 2008 Resource Directory & Index 2007* (2006).

[2] Bey, Y., Kageura, K., and Boitet, C. Beytrans: A free online collaborative wiki-based CAT environment dedicated for online translation. In *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation (PACLIC 21)* (2007).

[3] Désilets, A. Translation wikified: How massive online collaboration impact the world of translation. In *Translating and Computer Conference Proceedings 29* (Nov. 2007).

[4] Désilets, A., Barriere, C., and Quirion, J. Making wikimedia resources more useful for translators. In *Proc. WikiMania'07: The International WikiMedia Conference* (2007).

[5] Désilets, A., Gonzalez, L., Paquet, S., and Stojanovic, M. Translation the wiki way. In *Int. Sym. Wikis* (2006), D. Riehle and J. Noble, Eds., ACM, pp. 19–32.

[6] Facebook. Facebook releases site in German, 2008. www.facebook.com/press/releases.php?p=20727.

[7] Healey, J. Changing the world one word at a time. *MultiLingual* (February 2008).

[8] Huberdeau, L.-P. I18n revisited. blog.lphuberdeau.com/wordpress/2006/08/20/i18n-revisited/, 2006.

[9] Huberdeau, L.-P. Architecture for collaborative translation synchronization. blog.lphuberdeau.com/wordpress/wp-content/uploads/2008/02/architecture.pdf, 2008.

[10] Huberdeau, L.-P. Moteur de wiki multilingue. Tech. rep., École de Technologie Supérieure de Montréal, Apr. 2008.

[11] Jones, M. C., Rathi, D., and Twidale, M. B. Wikifying your interface: facilitating community-based interface translation. In *Proceedings of the 6th conference on Designing Interactive systems* (2006).

[12] Laporte, M., and de Pedro Puente, X. Interactive translation. doc.tikiwiki.org/Interactive+Translation. Archive: www.webcitation.org/5XgWgJDtv, 2007-2008.

[13] Luetke, T. Crowdsourcing internationalization. jadedpixel.com/2006/9/4/crowdsourcing-internationalization. Archive: www.webcitation.org/5XgWk7tft, Sept. 2007.

[14] Müldner, T., and Shen, Z. Cooperative development of internationalized documents. In *COOP'06, 7th International Conference on the Design of Cooperative Systems* (2007).

[15] Paquet, S. Measuring the translation progress in the CLWE project. www.wiki-translation.com/Measuring translation progress in the CLWE Project. Archive: www.webcitation.org/5XgVUGZJj, 2008.

[16] Simard, M., and Plamondon, P. Bilingual sentence alignment: Balancing robustness and accuracy. In *Proc. Conference of the Association for Machine Translation in the Americas (AMTA)* (1996).

[17] Wikipedia collective. Wikipedia translation project, 2006-2008. meta.wikimedia.org/wiki/Translation. Archive: http://www.webcitation.org/5XgYId3Pb.

Acknowledgements

The Cross-Lingual Wiki Engine project would have been impossible without all the support we had. While we cannot thank all those with whom we had discussions during various events, we can at least name a few.

First of all, Marc Laporte recognized our common interests and brought us all together. Without him, we would still be playing with our ideas individually. Nelson Ko brought us a real use case and dealt with the Mozilla foundation all along. Xavier de Pedro Puente was with us from the begining, reviewed every steps we made and proposed many improvements. Olaf Michael Stefanov came in late in the

project, but his insights greatly inspired this paper and we have no doubt he will play a significant role in the future of this project.

The TikiWiki community is also connected to the success of the CLWE. Not only did they allow us to implement our ideas in the core version *during the release process*, they also supported us in doing so and provided invaluable feedback. Among them, special mention must be made of Sylvie Greverend, who made the first inroads into multilingual support in TikiWiki a long time ago, and Rick Sapir, who spent countless hours producing a great screencast[9] to demonstrate the CLWE.

[9] http://clwe-demo.notlong.com