



---

## D6.4: Report on Integration into Community Translation Platforms

---

Philipp Koehn

Distribution: Public

---

CAsMACAT  
Cognitive Analysis and Statistical Methods  
for Advanced Computer Aided Translation

ICT Project 287576 Deliverable D6.4



Project funded by the European Community  
under the Seventh Framework Programme for  
Research and Technological Development.



Project ref no.	ICT-287576
Project acronym	CASMACAT
Project full title	Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation
Instrument	STREP
Thematic Priority	ICT-2011.4.2 Language Technologies
Start date / duration	01 November 2011 / 36 Months

Distribution	Public
Contractual date of delivery	October 31, 2014
Actual date of delivery	January 7, 2015
Date of last update	January 7, 2015
Deliverable number	D6.4
Deliverable title	Report on Integration into Community Translation Platforms
Type	Report
Status & version	Final
Number of pages	13
Contributing WP(s)	WP7
WP / Task responsible	UEDIN, UPVLC, CBS, CS
Other contributors	
Internal reviewer	Barto Mesa
Author(s)	Philipp Koehn
EC project officer	Aleksandra Wesolowska
Keywords	

The partners in CASMACAT are:

University of Edinburgh (UEDIN)  
Copenhagen Business School (CBS)  
Universitat Politècnica de València (UPVLC)  
Celer Soluciones (CS)

For copies of reports, updates on project activities and other CASMACAT related information, contact:

The CASMACAT Project Co-ordinator  
Philipp Koehn, University of Edinburgh  
10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom  
pkoehn@inf.ed.ac.uk  
Phone +44 (131) 650-8287 - Fax +44 (131) 650-6626

Copies of reports and other material can also be accessed via the project's homepage:  
<http://www.casmacat.eu/>

© 2014, The Individual Authors

No part of this document may be reproduced or transmitted in any form, or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission from the copyright owner.

# Contents

<b>1</b>	<b>Overview</b>	<b>4</b>
<b>2</b>	<b>Global Voices</b>	<b>4</b>
2.1	Integration Challenges . . . . .	4
2.2	Existing Translator Support . . . . .	5
2.3	CASMACAT Integration . . . . .	6
<b>3</b>	<b>TED Talks</b>	<b>7</b>
3.1	Integration Challenges . . . . .	8
3.2	Existing Translator Support . . . . .	8
3.3	CASMACAT Integration . . . . .	9
<b>4</b>	<b>Wikipedia</b>	<b>10</b>
4.1	Integration Challenges . . . . .	11
4.2	Existing Translator Support . . . . .	11
4.3	CASMACAT Integration . . . . .	11
<b>5</b>	<b>Usage</b>	<b>13</b>

# 1 Overview

Workpackage 6 evaluates the CASMACAT workbench not only in field trials, but also integrates the workbench into community translation platforms and collect user activity data.

This deliverable reports on the work carried out in Task 6.4, as specified in the description of work:

*Integration into community translation platforms (UEDIN 4PM, MONTH 19-36)  
Customization of the web-based CASMACAT workbench, hosted on servers of consortium members, to the needs of interested community translation platforms.*

We integrated the CASMACAT workbench into the following community translation platforms:

- Global Voices (<http://globalvoicesonline.org/>)
- TED Talks (<http://www.ted.com/>)
- Wikipedia (<http://www.wikipedia.org/>)

This integration is available on the CASMACAT web site at

<http://www.casmacat.eu/community/>

We describe the integration efforts in detail in the next sections.

## 2 Global Voices

Global Voices is a community news web site which describes itself as

*We are a borderless, largely volunteer community of more than 800 writers, analysts, online media experts and translators. Global Voices has been leading the conversation on citizen media reporting since 2005. We curate, verify and translate trending news and stories you might be missing on the Internet, from blogs, independent press and social media in 167 countries.*<sup>1</sup>

Almost all news stories are available in English, and a volunteer translation community translates them into other languages, either from English or the original language of the story. Stories are translated into more than 30 languages.

### 2.1 Integration Challenges

The news stories on the Global Voices web site (see Figure 1 for an example) consist of a title and a main body broken up into paragraphs.

Stories include hyperlinks and embed tweets, images, and videos. This requires the proper handling of tags. Since the tags will be identical for the original and the translated version, their content has to be preserved and can be hidden from the translator — except for some tag elements such as “title” in image tags.

Quotations are handled in a special blockquote format, which preserves the quote in the original language alongside the translation.

---

<sup>1</sup><http://globalvoicesonline.org/about/>

## Vladimir Putin and Russian Nationalists Don't Get Along. Here's Why.

Posted 14 October 2014 18:20 GMT



Like “Pussy Riot,” the Russian nationalist website “Sputnik & Pogrom” has a name that gives many pause, when they first hear it. The site’s founder and chief editor, Egor Prosvirnin, wanted to produce Russia’s first “truly nationalist journal,” and the name is meant to capture the heights of Russian intellect (which launched the world’s first artificial satellite) and the [mayhem](#) of Russian popular will (which periodically explodes in ethnic riots). This “synthesis of modernism,” Prosvirnin [says](#), shapes the post-Soviet Russian identity.

Prosvirnin’s animosity toward the Kremlin might strike many outside Russia as mysterious. Wouldn’t a self-avowed Russian nationalist revere Putin for sacrificing Moscow’s reputation with the West to deliver Crimea and rescue the [Donbas](#)? Isn’t Putin a nationalist?

“Putin is no nationalist—he’s just a spectator,” Prosvirnin told RuNet Echo. “He was put there [in the Kremlin] by the ruling corporation to manage the political process, while the noble members of the secret police buy villas and mansions in Cote d’-Azur.”

Prosvirnin believes that Russia’s current state—what he’s taken to calling “[Great Rotenbergia](#),” after the Rotenberg oligarch brothers—is antithetical to Russian nationalism. Indeed, the Duma is poised to approve new legislation, commonly known as “the Rotenberg Law,” that would compensate the individuals targeted by Western sanctions—some of Russia’s wealthiest people—using taxpayers’ money.

Translation

Original Quote

“ Have you ever thought about how it’s impossible to say what Putin has accomplished these 14 years, if you forget a moment about Crimea? “Stability,” “doubling GDP,” “the Olympics”—it’s gotten to the point that Putin is fundamentally incapable of explaining to the population why he’s still in the Kremlin. He’s a hired manager—only it wasn’t the people who hired him, but the KGB generals and Yeltsin’s oligarchs. Have a look at the Rotenberg Law. Is this something a nationalist, or even a populist, could even consider?

Earlier this month, Prosvirnin [revealed](#) that federal police are building a criminal case against him, on

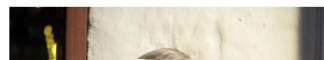


Figure 1: Example of a Global Voices news story, which includes images and quotations

## 2.2 Existing Translator Support

Global Voices has a story editor, which is identical for original authoring of stories and translation of stories. The editor is a HTML textarea which uses a format similar to many Wikis. Alternatively, a WYSIWIG environment can be used.

## 2.3 CASMACAT Integration

Volunteer translators are most likely interested in translating the most recent news stories that have been published. Hence, we offer on the opening menu<sup>2</sup> of our support site a choice of recent articles in reverse chronological order (newest first). This list is based on a regular crawl of each of the language editions of the Global Voices web site. Figure 2 shows a partial screenshot of this menu in the CASMACAT support site.

**Community Translation Platform: Global Voices**

---

Select which language pair you want to translate.  
To associate your translations with you,  
please also provide an email address.

Language Pair

Email

Translate URL

**Recent English Articles**

Date	Title	Words
10/14	<a href="#">Some Gambians Don't Feel Like Celebrating President Jammeh's 20 Years in Power</a>	815 <a href="#">translate</a>
10/14	<a href="#">Vladimir Putin and Russian Nationalists Don't Get Along</a>	779 <a href="#">translate</a>
10/14	<a href="#">Macedonian Civic Sector Starts Fundraising to Aid Independent Fokus Magazine</a>	313 <a href="#">translate</a>
10/14	<a href="#">Following Political Pressure, Citizen-Led Rural Libraries Shut Down in China</a>	808 <a href="#">translate</a>
10/14	<a href="#">Child's Murder Unveils Lack of Sympathy for Japan's Single Mothers</a>	837 <a href="#">translate</a>
10/14	<a href="#">Artists Create Climate Change Mural in Grenada to Warn of Modern-Day 'Paradise Lost'</a>	1165 <a href="#">translate</a>
10/14	<a href="#">Animated Video Dispels Ebola Myths</a>	100 <a href="#">translate</a>

Figure 2: Main menu in CASMACAT support website: recent stories are listed as candidates for translation.

The volunteer translator may click on the story title to see the original page, or click on “translate” to enter the CASMACAT workbench and work on translation of a story. In order to translate an article, the translator also has to enter an email address, which establishes a user account. There are no passwords or any further validation of the email address.

Once a translator has entered an email address and started translating news stories, the main menu also displays a personalized list of news stories — see Figure 3. In this list, in addition to links to the original story and the CASMACAT workbench, a third link allows the conversion of the translation entered into the CASMACAT workbench into the format used by the Global Voices content management system.

In order to support the different formats involved in this platform (XLIFF format of CASMACAT and Wiki/HTML of the Global Voices web site), we built the following converters:

- HTML → Global Voices editor format
- Global Voices editor format → XLIFF
- XLIFF → Global Voices editor format

<sup>2</sup><http://www.casmacat.eu/community/?action=globalVoices>

## Translations by phi@jhu.edu

Date	Title	Language	
09/27	<a href="#">Video of Japan's Mount Ontake Eruption as it Happened</a>	en-de	<a href="#">translate</a> <a href="#">convert</a>
09/21	<a href="#">Egyptian Leftist Activist Mahinour El-Masry Freed after Spending 125 Days in Jail</a>	en-da	<a href="#">translate</a> <a href="#">convert</a>
09/21	<a href="#">Egyptian Leftist Activist Mahinour El-Masry Freed after Spending 125 Days in Jail</a>	en-es	<a href="#">translate</a> <a href="#">convert</a>

Figure 3: List of stories selected by a translator. To the right: “translate” links to the CASMACAT workbench and “convert” links to the Global Voices editor.

The XLIFF format also encodes meta information about tag contents, paragraph breaks, blockquotes, tweets, etc. This allows us to present in the CASMACAT workbench the text of the news stories, while ignoring these formatting issues as much as possible. Tags are just presented by their tag names, so a complex hyperlink becomes a simple {a} and {\a} markup.

At this point, the CASMACAT integration in Global Voices supports 10 language pairs:

- English to German, Spanish, French, Greek, Danish and Portuguese.
- French and Spanish to English.
- Spanish and Portuguese to German.

The most active language pair so far for this integration has been English to German. These German translators also requested additional language pairs (such as Spanish and Portuguese to German), which have also been provided.

Cut and paste the following into Global Voices' content management system.

### Headline

Pazifikinselnbewohner planen Blockade des weltgrößten Kohlehafens Kanus, um gegen den Klimawandel zu protestieren

### Content

[caption id="attachment\_488814" class="aligncenter" width="600"]Pacific Climate Warriors Vanuatu canoe launch. Photo credit: 350.org[/caption]Start der Kanus der pazifischen Klimakrieger Foto: 350.org[/caption]

<em><span style="color: #888888;"> Diesen Artikel schrieb Aaron Packard für </span><a href="http://350.org/pacific-climate-warriors-will-block-the-worlds-largest-coal-port/">350.org</a></em><span style="color: #888888;">,</span> <span style="color: #888888;">eine Organisation zum Aufbau einer globalen Klimaschutzbewegung, und wird von Global Voices im Rahmen einer Inhaltsaustauschvereinbarung veröffentlicht.</em></span>

Figure 4: CASMACAT conversion to the format expected by the Global Voices editor.

## 3 TED Talks

TED is an audiovisual platform which describes itself as:

*TED is a nonprofit devoted to spreading ideas, usually in the form of short, powerful talks (18 minutes or less). TED began in 1984 as a conference where Technology, Entertainment and Design converged, and today covers almost all topics from science*

*to business to global issues in more than 100 languages. Meanwhile, independently run TEDx events help share ideas in communities around the world.*<sup>3</sup>

See Figure 5 for a screenshot of the video for one of the talks featured on the TED web site.

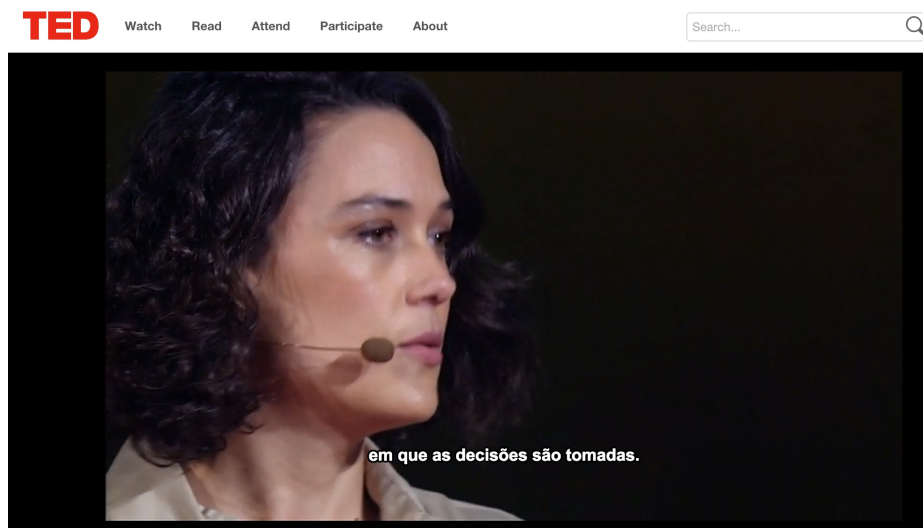


Figure 5: TED talk with Portuguese subtitles.

TED is supported by a thriving volunteer community that creates and translate subtitles for the talks. According to statistics on the web site, more than 63,000 translations have been published in 105 languages, created by almost 18,000 volunteers by October 2014.

### 3.1 Integration Challenges

TED translators work from the TED-Amara platform with transcribed speech, so there is no markup or annotation to deal with. The main challenge is the space and time constraints for the subtitles in order to make them readable in real time. For this end, sentences have to be broken up into shorter fragments, typically along syntactic lines. With fast speakers, it is difficult for the listener to keep up with reading the subtitles. This means that translators cannot expand the text (even though they may be inclined to do so to provide helpful background).

In the TED-Amara platform, translators are able to change the timing of the subtitles for each language — for instance allowing more time for lengthened subtitles by shrinking the time for adjacent shortened subtitles.

### 3.2 Existing Translator Support

The Amara platform is an editor for segment by segment translation of the platform, that also displays the video (see Figure 7). Amara is a service of Participatory Culture Foundation, a non-profit organization building free and open tools for more democratic and decentralized media<sup>4</sup>. Amara is used not only by TED, but also by companies such as Netflix, the MOOC provider Udacity, or the American public television broadcaster PBS.

<sup>3</sup><http://www.ted.com/about/our-organization>

<sup>4</sup><http://amara.org/en/about>



### 3.3 CASMACAT Integration

The integration of the CASMACAT workbench with the existing translation tools for TED-Amara platform follows a similar pattern as the previously described integration of Global Voices (see section 2.1).

Talks for translation are associated with a translator via email address. Entering the email address gives a list of translation tasks in progress or completed, as well as a form to start a new translation task.

Data between the TED-Amara platform and the CASMACAT workbench is exchanged in a text format. Specifically, once the translator decides to work on the translation of a talk, she proceeds with the following steps:

1. Download the original language subtitles in text format.
2. Upload the text file to the CASMACAT community translation platform.  
(involves automatic conversion of the text into XLIFF format)
3. Translate the subtitles with the CASMACAT workbench.
4. Download the translations from the CASMACAT workbench as a text file.  
(involves automatic conversion of the XLIFF into text format)
5. Upload the text file with the translations to the Amara platform.

Figures 6 and 7 show how subtitles can be downloaded from the Amara platform and then their translation uploaded back to the Amara platform.

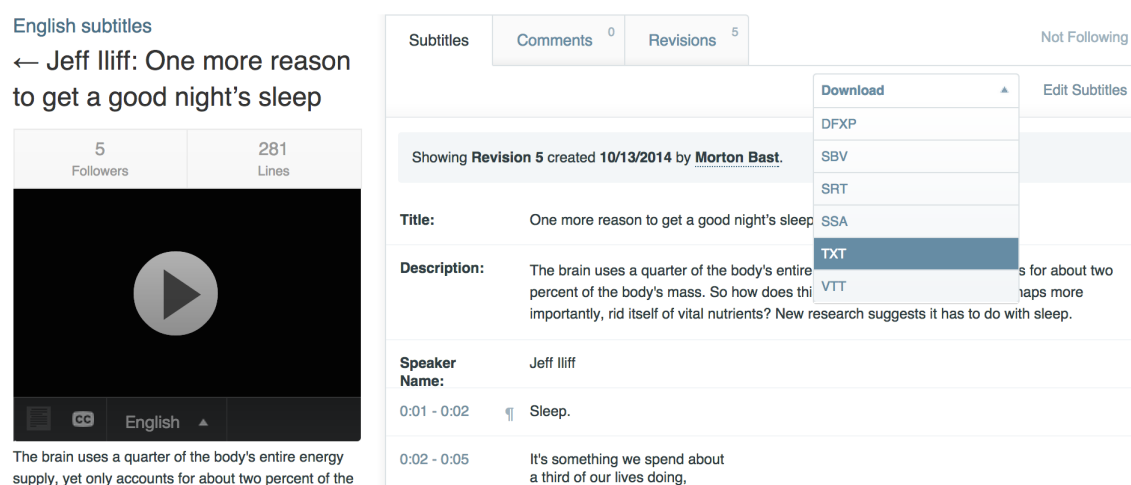


Figure 6: The TED-Amara platform allows the export of the original subtitles in text format.

On the CASMACAT web site<sup>5</sup> for the TED community translation platform, translators can start new translation projects with the original subtitles file in text format and convert the completed translation into the text format required by translators. See Figure 8 for a screenshot of the webpage offering these options.

<sup>5</sup><http://www.casmacat.eu/community/?action=tet>

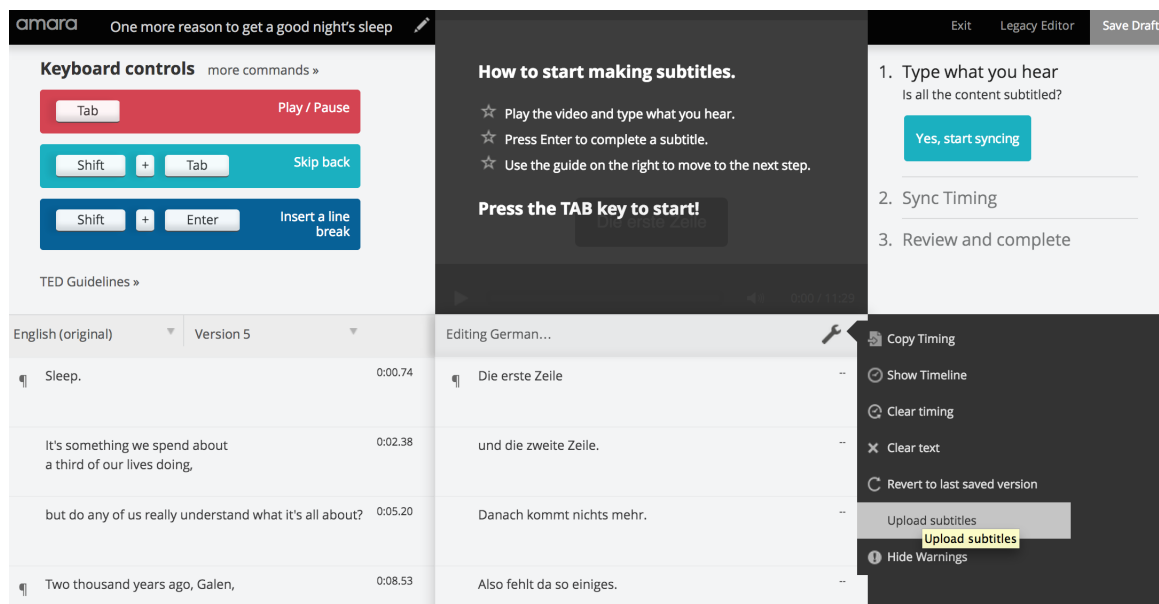


Figure 7: Translated subtitles from the CASMACAT workbench can be uploaded the Amara platform for TED translators.

Select which language pair you want to translate.  
To associate your translations with you, please also provide an email address.

Language Pair: English-German

Email: phi@jhu.edu

**SIGN IN**

**Start new project**

Download the English subtitles in TXT format for upload into the CASMACAT workbench ([how?](#)). Upload the file below and click "TRANSLATE".

Translate TXT: Choose File No file chosen

**TRANSLATE**

**Translations by phi@jhu.edu**

Title	Language
Jeff Iliff One more reason to get a good nights sleep.en (1).txt	en-de <a href="#">translate</a> <a href="#">convert</a>

Figure 8: Translators may start new projects with downloaded source subtitles files in text format, and then manage their translation tasks by continuing the translation process (translate) or conversion of the completed translation into the format required by Amara (convert).

## 4 Wikipedia

Wikipedia defines itself as

*[...] a multilingual, web-based, free-content encyclopedia project supported by the Wikimedia Foundation and based on an openly editable model. [...] Wikipedia is written collaboratively by largely anonymous Internet volunteers who write without pay. [...] Anyone with Internet access can write and make changes to Wikipedia articles.*<sup>6</sup>

<sup>6</sup><http://en.wikipedia.org/wiki/Wikipedia:About>

Wikipedia has become the de facto standard online encyclopedia, and its content has been integrated into many other platforms and applications (such as Google search). Wikipedia contains almost 5 million articles in 287 languages<sup>7</sup>.

## 4.1 Integration Challenges

Wikipedia uses a wiki format, similar to Global Voices, but much more complex since it also allows for hierarchical document structure, tables, and special information boxes. Translation of content for this platform requires clear distinction between functional elements and content, and the proper preservation and adaptation of the former.

A particular issue is the translation of links since they link to other Wikipedia articles. The translation of a Wikipedia article should then point to a Wikipedia article in the translated language instead of the original language.

## 4.2 Existing Translator Support

Wikipedia does not expect that articles are created through translation, but rather authored independently. Hence, there is no integrated support for translators. It does encourage translation of English articles into other languages<sup>8</sup>, but does not prescribe any specific tool.

## 4.3 CASMACAT Integration

The integration of Wikipedia as a CASMACAT community translation platform<sup>9</sup> is modeled very closely to Global Voices due to similarities in document format (marked up wiki) and edit interface (HTML textbox). Since English coverage is very good, we limit ourselves to translation from English to other languages (German, Danish, Spanish, French, Greek, and Portuguese).

As a back-end, we currently use the machine translation engines as for Global Voices and, instead of providing a list of recent articles, we show the currently most popular articles in English that have not yet been translated into other languages (see Figure 9). The user may also enter the name of any other Wikipedia article to be translated.

The more complex format of the wiki markup requires more dedicated methods for conversion of the wiki format into XLIFF and back. We directly convert from the wiki edit format (see Figure 10), which is available by crawling the edit links. Notable special handling include:

- Citations may include text to be translated and references (article titles, book names, etc.) that remain unchanged.
- Tables structure must be preserved and only the content of table cells passed on to the translator.
- Markup can be nested - Tables may include references, references may include links, etc.

As mentioned above, a especially interesting problem are links to other Wikipedia articles. For instance, when translating an English document about *cats* into German, and it links to *dogs*, then we also have translated the link to *Hunde*. These links have to match up with the existing Wikipedia link structure. We automate this process as follows:

---

<sup>7</sup>[https://meta.wikimedia.org/wiki/List\\_of\\_Wikipedias](https://meta.wikimedia.org/wiki/List_of_Wikipedias)

<sup>8</sup><http://en.wikipedia.org/wiki/Wikipedia:Translation>


<sup>9</sup><http://www.casmacat.eu/community/?action=wikipedia>

## Popular Untranslated English Articles







Views	Title	Words	
1,263,228	<a href="#">Ultron</a>	3	<a href="#">translate</a>
741,451	<a href="#">Avengers: Age of Ultron</a>	4778	<a href="#">translate</a>
733,860	<a href="#">Happy New Year (2014 film)</a>	1941	<a href="#">translate</a>
636,016	<a href="#">American Horror Story: Freak Show</a>	370	<a href="#">translate</a>
605,331	<a href="#">June and Jennifer Gibbons</a>	1125	<a href="#">translate</a>
445,654	<a href="#">Gamergate controversy</a>	8957	<a href="#">translate</a>
408,117	<a href="#">Joyce Vincent</a>	1051	<a href="#">translate</a>
380,481	<a href="#">Edward Mordake</a>	670	<a href="#">translate</a>
379,614	<a href="#">Teenage Mutant Ninja Turtles (Mirage Studios)</a>	2390	<a href="#">translate</a>
375,561	<a href="#">Kaththi</a>	3051	<a href="#">translate</a>
369,216	<a href="#">Annabelle (film)</a>	2001	<a href="#">translate</a>
351,138	<a href="#">The Walking Dead (season 5)</a>	958	<a href="#">translate</a>
349,045	<a href="#">Benjamin Kyle</a>	1701	<a href="#">translate</a>
342,145	<a href="#">List of Bollywood films of 2014</a>	1336	<a href="#">translate</a>
287,308	<a href="#">Hank Pym</a>	5668	<a href="#">translate</a>

Figure 9: Popular Wikipedia articles (end of October 2014) that have not yet been translated into German.

## Editing Edward Mordake

 You are not logged in. Your IP address will be publicly visible if you make any edits. If you [log in](#) or [create an account](#), your edits will be attributed to a user name, among [other benefits](#).

Content that [violates any copyrights](#) will be deleted. Encyclopedic content must be [verifiable](#). Work submitted to Wikipedia can be edited, used, and redistributed—by anyone—subject to [certain terms and conditions](#).







▶ [Advanced](#)
▶ [Special characters](#)
▶ [Help](#)
▶ [Cite](#)

```

{{For|the "American Horror Story: Freak Show" episode|Edward Mordake (American Horror Story)}}
'''Edward Mordake''' (sometimes spelled '''Edward Mordrake''') was the name given to an [[apocryphal]] 19th-century heir to an unspecified [[English people|English]] [[peerage]] who was said to have suffered from a form of [[diprosopus]]. According to sources, he had, on the back of his head, an extra face which could neither eat nor speak out loud, although it was described as being able to laugh and cry. Mordake reportedly begged doctors to have his "Demon face" removed, claiming that it whispered to him at night, but no doctor would attempt it. He committed suicide when he was 23 years old.<ref name="Gould1956">{{cite book|last=Gould|first=George M.|authorlink=George M. Gould|title=Anomalies and Curiosities of Medicine|url=http://books.google.com/books?id=TDZvCfUVO0QC&pg=PA124|accessdate=October 29, 2014|year=1956|publisher=Blacksleet River|isbn=978-1-4499-7722-1|pages=124-125}}</ref>

```

Figure 10: Edit textbox for Wikipedia articles. Note the complex markup.

1. We detect links in the original page (e.g., *dogs*).
2. We retrieve the linked page
3. We check the language links on that page to see if an article in the target language exists (e.g., *Hunde*).
  - if such a link exists, we automatically insert it into the resulting translation — the user only sees `tt {{link}}`.
  - if such link does not exist, we remove it to avoid dead links.

## 5 Usage

The integration of the Global Voices platform was finished first, and has been promoted to the Global Voices translator community, especially the translators working on translation into German, who requested additional language pairs (Portuguese-German and Spanish-German).

The total number of sentence translation requests are summarized in Table reftab:usage. Usage is currently still relatively light, but given that the support will remain available after the project is expected to increase

Language Pair	en-da	en-de	en-el	en-es	en-fr	en-pt	es-de	es-en	fr-en	pt-de
Translations	43	287	150	86	22	23	4	33	42	10