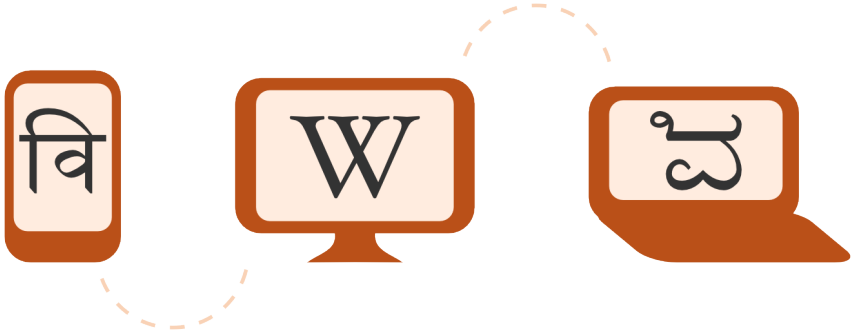


Content Translation

Computer-assisted translation tool for Wikipedia articles

Niklas Laxström[1 2], Pau Giner[1], Santhosh Thottingal[1]
[1]Wikimedia Foundation
[2]University of Helsinki

Agenda

1. Introduction
2. What we did
3. Challenges and learnings
4. Results
5. Next steps

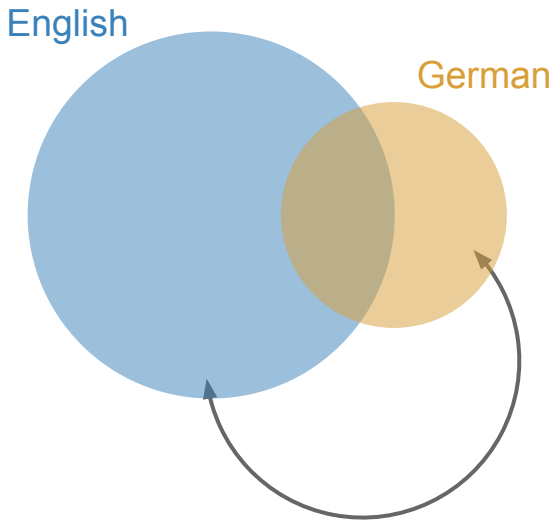

EN
4.4M


ES
1M


CA
4K

“Surprisingly small amount of content overlap between languages of Wikipedia”

The English Wikipedia contains only 51% of the articles in the second-largest edition, German.



Big opportunities for translation

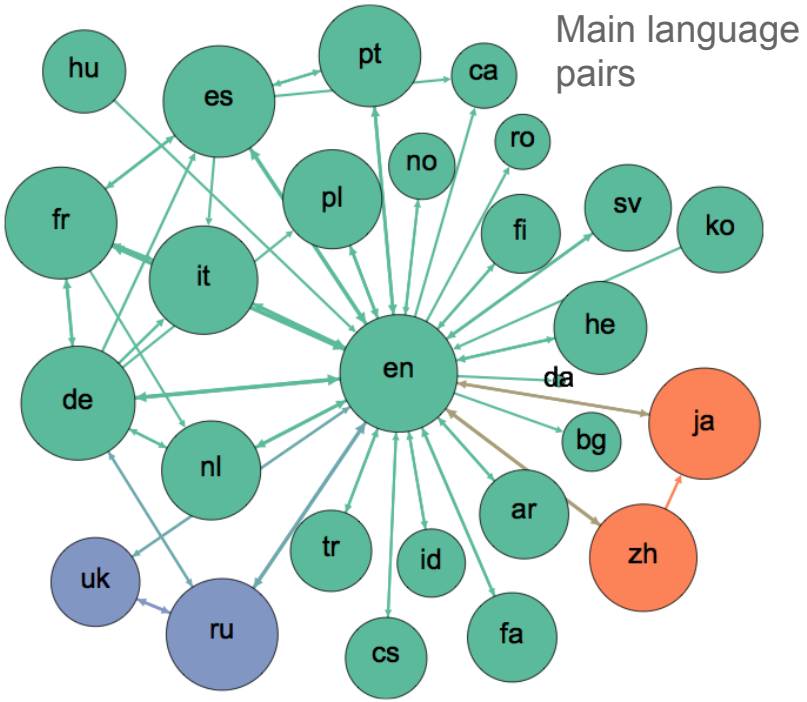
Based on the analysis of one month of all edits to the top 46 language editions of Wikipedia by Scott A. Hale, University of Oxford. More info at http://arxiv.org/abs/1312.0976

Potential users are active

Over **15% of users** edit multiple language editions.

These **multilingual users are more active** (2.3 times) than their monolingual counterparts on average.

Multilingual users made **30% of all edits**.

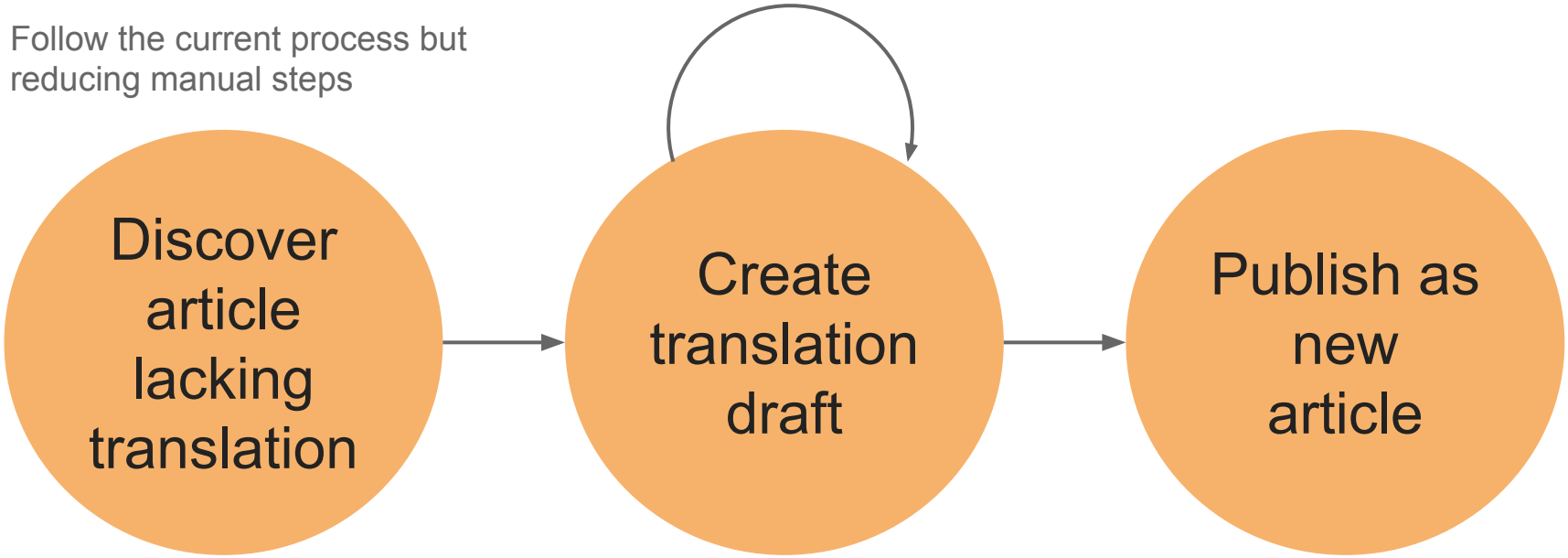


Based on the analysis of one month of all edits to the top 46 language editions of Wikipedia by Scott A. Hale, University of Oxford. More info at http://arxiv.org/abs/1312.0976

Content Translation

Workflow

Follow the current process but reducing manual steps



Entry points

Ways to make users aware of the tool

Translation view

Where translations are made

Content Translation

How to translate a Wikipedia article

Saved 2 minutes ago

Publish translation

Cupcake

español view page

2 categories

Cupcakes con glaseado de chocolate.

Un ***cupcake*** —literalmente en español: «tarta en taza»—, es una pequeña porción de tarta para una persona. Se hornean en un molde igual que el de magdalenas y muffins. En el molde se colocan unos papeles llamados cápsulas.

Normalmente es confundido con los muffins y

Cupcake

català

No categories

Cupcakes amb setinat de xocolata.

Un ***cupcake*** —literalment en espanyol: «pastís en tassa»—, és una petita porció de pastís per a una persona. Es prepara en un motlle igual que el de magdalenes i muffins. En el motlle es col·loquen uns papers anomenats càpsules.

+ Add translation

pastís

B I

Link español

Link català

Pastís

+ Add link

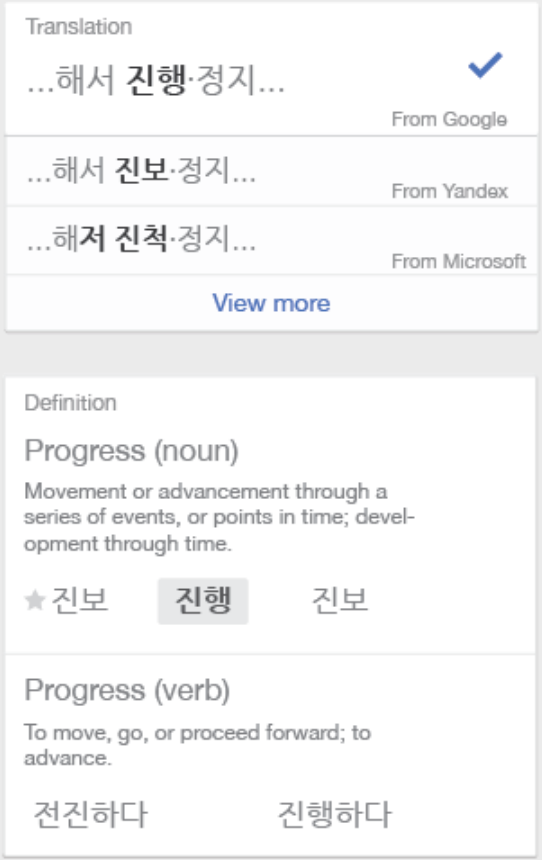
Select a paragraph and improve the initial automatic translation.
The user is not forced to translate the whole article.

Provide information at hand

Context relevant information.
Integrate information from different sources (dictionaries, glossaries, translation services).

Avoid information overload.
Make the information compact.

(Future designs)



Challenges and learnings

A simple workflow

One paragraph at a time. Provide enough freedom to rearrange sentences, but don’t force to translate the whole document.

Provide context. Visually aligning source and translations communicates what is translated and what is lacking.

Editing freedom. Don’t provide a strict workflow. Let edit the document as freely as possible.

```
Development Report |date=14 March 2013 |acc
|HDI_rank = 69th
|currency = [[Turkish lira]] [[Turkish lira
|currency_code = TRY
|time_zone = [[Eastern European Time|EET]]
|utc_offset = +2
|time_zone_DST = [[Eastern European Summer
|utc_offset_DST = +3
|date_format = dd/mm/yyyy ([[Anno Domini|AD
|drives_on = right
|calling_code = [[Telephone numbers in Turk
|cctld = [[.tr]]
|website = [www.turkiye.gov.tr]
}}

'''Turkey''' ({{IPAc-en|audio=en-us-Turkey.
'''Republic of Turkey''' ({{lang-tr|Türkiye
[[parliamentary republic]] largely located
[[Southeastern Europe]]. Turkey is bordered
(country)|Georgia]] to the northeast; [[Arm
Republic|Nakhchivan]] to the east; and [[Ir
[[Aegean Sea]] to the west; and the [[Black
```

Turkey

English view page

18 categories

Turkey (/ˈtɜːrˈkiː/; Turkish: *Türkiye* ['tyɾcijɛ]), officially the **Republic of Turkey** (Turkish: *Türkiye Cumhuriyeti* ['tyɾcijɛ d͡ʒum'huɾijeti] (listen)), is a parliamentary republic largely located in Western Asia with the portion of Eastern Thrace in Southeastern Europe. Turkey is bordered by eight countries: Bulgaria to the northwest; Greece to the west; Georgia to the northeast; Armenia, Iran and the Azerbaijani exclave of Nakhchivan to the east; and Iraq and Syria to the south. The Mediterranean Sea is to the south; the Aegean Sea to the west; and the Black Sea to the north. The Sea of Marmara, the Bosphorus and the Dardanelles (which together form the Turkish Straits) demarcate the boundary between Thrace and Anatolia; they also separate Europe and Asia.[5] Turkey's location at the crossroads of Europe and Asia makes it a country of significant geostrategic importance.[6]

Turkey has been inhabited since the paleolithic age,[7] including various ancient Anatolian civilizations, Aeolian and Ionian Greeks, Thracians, and Persians.[8][9][10] After Alexander the Great's

Apertium - open source MT

Supports only plain text translation
Annotation transfer system as pre-processing and post-processing

Apertium APY scalable web service
Self hosted high-availability web service

Good result for some language pairs
Spanish-Catalan pair was well received by translators

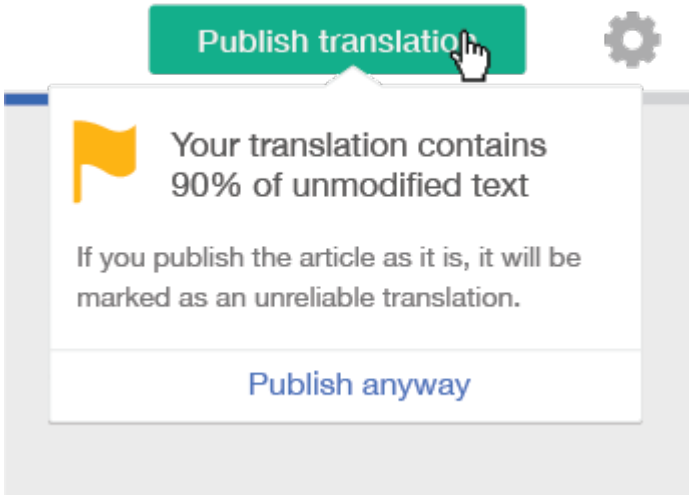
Machine translation is one of the translation tools,
often not available for a language pair.

Quality matters

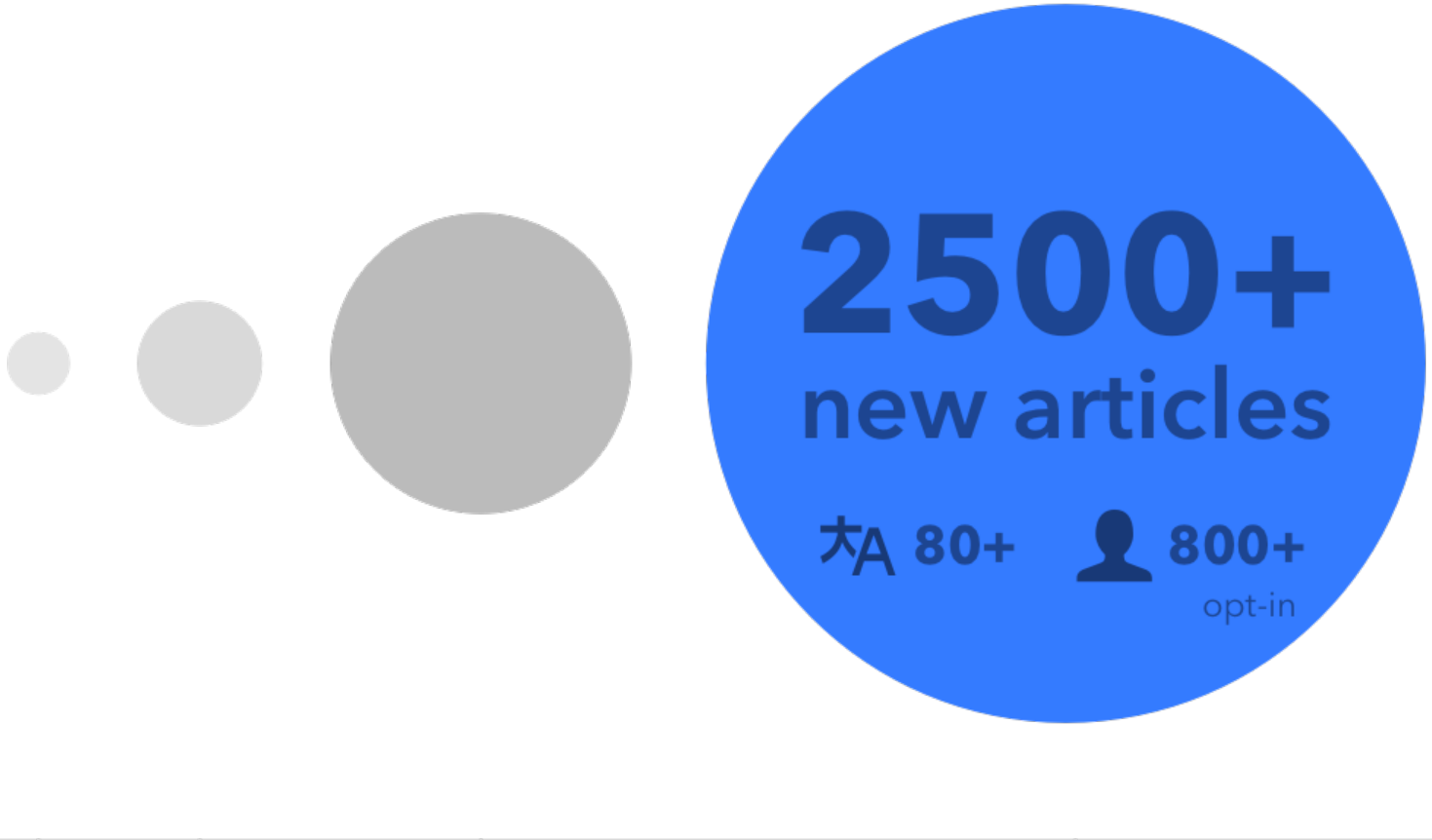
Educate. Convey that the focus is more on quality than on quantity.

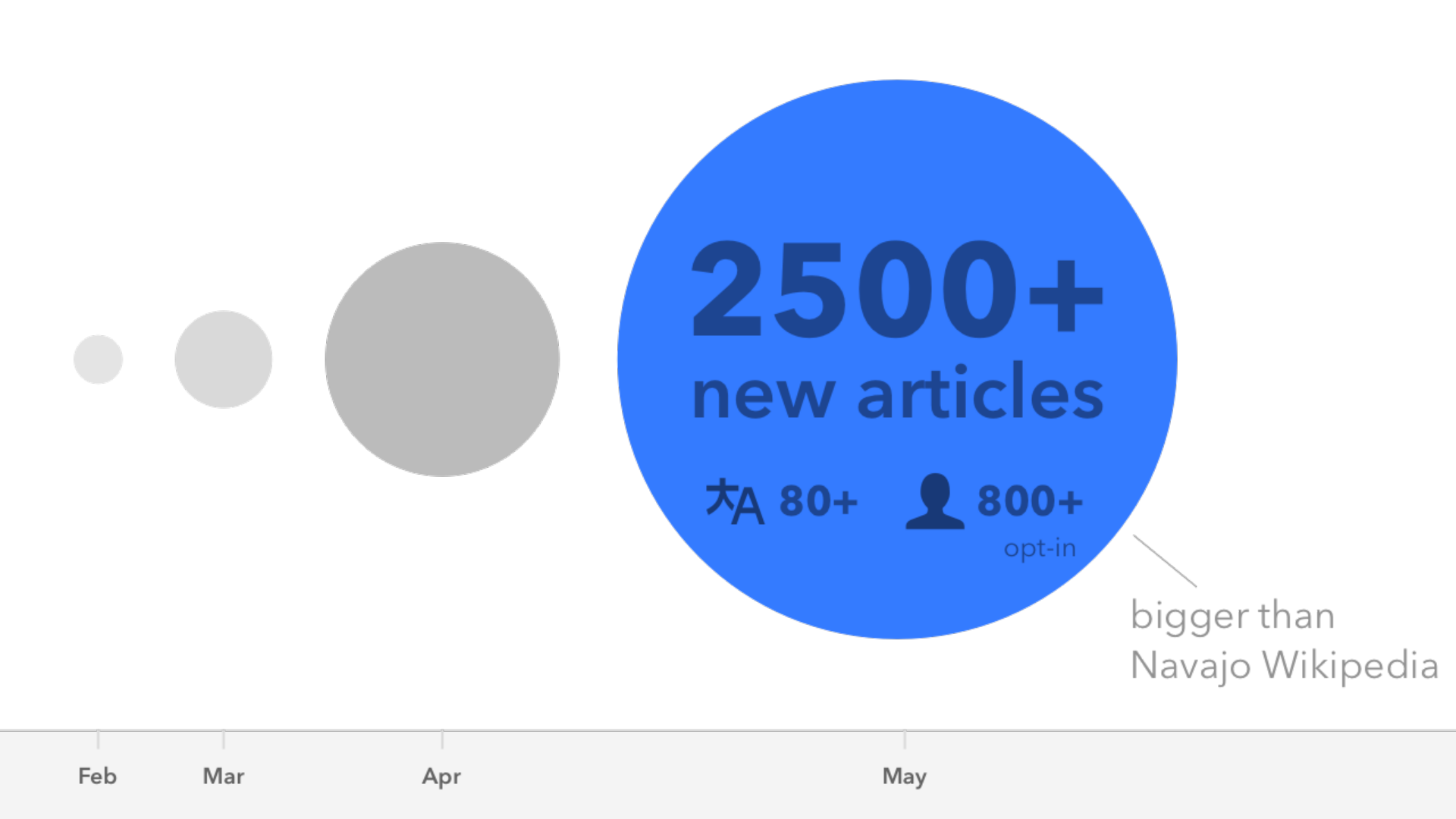
Warn. Detect potential patterns that lead to low-quality (unmodified automatic translation or pasted text).

Inform the community. Allow other users to easily find potential problematic content.



Results


2500+
new articles
80+
800+
opt-in
Feb
Mar
Apr
May


2500+
new articles
80+
800+
opt-in
bigger than
Navajo Wikipedia
Feb
Mar
Apr
May

Next steps

Parallel corpora

Freely available
CC BY-SA

Comparable corpora or translation corpora?
From section level alignment up to sentence level alignment

For improving and developing MT engines

Better understanding

Why translation happens more in some languages and less in others?

What is the effect of MT availability?

MT does not guarantee large amount of translations.

When and how to encourage people to translate?

Target multilingual people with interesting content.

Content Translation is reusable and extensible.

Not limited to just Wikipedia

Code available under GPL-2.0+

Extensible with more translation tools

Contact us!

More details

http://mediawiki.org/wiki/CX

Machine translation

Availability

Not all languages have MT support.

Quality

Quality of MT varies a lot across languages.

Open source and closed source

Apertium is the prominent open source MT system.

Machine translation is one of the translation tools,
often not available for a language pair.

2500+ new articles
by **800+** opt-in translators
in **3** months
across **50+** languages

Together about the size of Old English Wikipedia & bigger than 80 other wikipedias.