Locating Foci of Translation on Wikipedia: Some Methodological Proposals[1]

Mark Shuttleworth

This is the 'author accepted manuscript' – i.e. it reflects changes made in the peer review process. The following is the citation of the finalised article:

Shuttleworth, Mark (2017) 'Locating Foci of Translation on Wikipedia: Some Methodological Proposals', in *Translation Spaces* 6:2, 310-32, doi: ts.6.2.07shu.

The article is under copyright and the publisher should be contacted for permission to re-use or reprint the material in any form.

Abstract

In spite of its highly multilingual nature, it is generally accepted that most Wikipedia content is the product of original writing rather than being translated from another language version of the encyclopaedia. Wikipedia represents what is almost a complete, self-contained but vast research ecosystem; however, an unusual initial challenge for the researcher is to identify the primary data for a specific research project. The main aim of this paper is to make a number of proposals towards a possible methodology for discovering where the main foci of this new type of collaborative translation are located. Significant methods for this include the use of the encyclopaedia's list-based structure and of different features of page anatomy. The article also aims to outline specific topics for research and, during the discussion, offers some initial findings of its own, using Russian and Chinese to English translation as its main sources of examples.

Keywords: Wikipedia translation, collaborative translation, interwiki links, Revision History, Talk Pages, Chinese, Russian, Ukrainian

[1] This article has grown out of a presentation given as part of the ARTIS-sponsored "Multidimensional Methodologies: Collaboration and Networking in Translation Research" conference (http://www.ucl.ac.uk/translation-studies/artis-conference) that took place at University College London on 15-16 June 2015.

1. Introduction

If you happened to visit the British Library during the summer of 2015 you may well have seen an exact embroidered reproduction of Wikipedia's article on Magna Carta on display. Over twelve metres in length, it depicted this single article in great detail: text, formatting, images and references. Further features, on the other hand, were not so easily reproducible: there were no threads linking to embroideries of other articles, and one could say with reasonable certainty that no restitching took place every time someone performed a minor edit on the 'real' article. In spite of these inevitable limitations, this huge and imposing exhibit still managed to convey something of the sheer size and complexity of the encyclopaedia, by its simple physicality and also on a more metaphorical level, portraying this encyclopaedia entry as one large entity composed of an intricate fabric of thoughts and ideas, coming from many disparate authors and sources but together making up a single, coherent whole.

Wikipedia translation is generally perceived as a type of collaborative translation. This perception is, however, only partially correct: while there is a good deal of guidance available in terms of how and even what to translate (see for example "Wikipedia:Translation" 2016), as well as unlimited scope for informal collaboration and even one or two closely co-ordinated projects, it is also true that many editors who translate the encyclopaedia's content are likely to be acting largely under their own steam and in isolation. Nevertheless, translation as it can be observed in the encyclopaedia is without a doubt an under-researched though potentially significant type of translation. That said, there appears to be a relatively widespread (but mistaken) impression that this area is too complex (would 'haphazard' be a better word?) to research, so it is hoped that this article will demonstrate that a systematic approach to data collection is indeed possible, and that the topic should rightly be of interest to researchers from many different branches of the discipline.

As of November 2016, the "Benjamins Translation Studies Bibliography" (2017) lists only two articles with the word 'Wikipedia' in the title: Alonso (2015) and McDonough Dolmaya (2015). Of these, the first concerns itself with professional translators' perceptions of Wikipedia as a source of information and so does not have a direct bearing on the topic of this article; the second, on the other hand – an examination of Wikipedia translation quality – is certainly of relevance as it introduces an interesting approach to collecting translated material within the encyclopaedia. As research data for an investigation of revision practices this scholar sources articles translated from French and Spanish into English with the help of the Wikipedia article "Wikipedia:Pages needing translation into English" (2016); this will be outlined in Section 3.2.3 as an important variant of the second approach to identifying translated material.

To my knowledge on the other hand, none of the other main approaches outlined below have yet been documented in the translation studies literature. More generally, though, since an extravagant amount of information about the encyclopaedia is available within its pages, this article follows the principle of using Wikipedia to study Wikipedia. What this means in practice will be made clear as the article proceeds.

The multilingual Wikipedia is sometimes described as having come into being as a result of crowdsourced translation; however, it is possible to overstate the role that translation has played in shaping the content of this multilingual encyclopaedia, and indeed the level of organisation that such a statement implies, although such questions are still under investigation. If early indications turn out to be accurate, human translation is only one of a number of mechanisms facilitating the cross-language expansion of Wikipedia (others including paraphrase, non-native writing and the use of un-post-edited machine translation: see Shuttleworth 2015). With the rapid increase in the use of machine translation and also with the growing importance of crowdsourcing and other types of collaborative translation we are currently witnessing a massive shift in working practices within the translation industry, with the eventual (or even interim) destination as yet unclear. While these two approaches have much in common, one of the neatest ways of distinguishing between them is in my view that proposed by Fernández Costales, who argues that the difference is one of hierarchy, the former involving direct networking between equals and the latter presupposing some form of supervision, management or support by an organisation (2013, 96) – a distinction that is also adhered to by Jiménez-Crespo (2017, 19) – although both of course rely on the use of volunteers. A third concept, known as 'translation the wiki way', concerns the identification of efficient procedures for creating and managing multilingual wiki content (see Désilets et al. 2006). One way or another, Wikipedia is characterised by a "massive" collaborativity as pointed out by the author of "How will Massive Online Collaboration Impact the World of Translation" (Désilets 2007, n.d.).

With this distinction in mind, while some translation activity on Wikipedia (such as the medical "Translation Task Force": "Wikipedia WikiProject Medicine_Translation task force" 2016) can be thought of as crowdsourced in the standard sense, because of its lack of formal organisation and its more ad hoc (i.e. self-motivated) nature most such activity probably does not merit that denomination. This article focuses mainly on the second of these broad categories as locating material translated within one or other organised project is not usually problematic.

Finally, the ways in which factors other than translation contribute to the creation and spread of knowledge across Wikipedia are being investigated by scholars in other disciplines. Some interesting work on the interlingual pathways of influence within the encyclopaedia has been carried out by geolinguists Liao and Petzold (2010), who study interwiki links to build up a picture of the encyclopaedia's overall interlingual interconnectedness. Yasseri et al. (2014) present a comparative discussion of the distribution of controversial content across ten different language versions of the encyclopaedia.

2. Why study Wikipedia translation – and how?

This article proposes possible methods for locating and identifying material in one language version of Wikipedia that appears to have been translated from material contained in another. While it will be drawing examples mainly from Russian to English and Chinese to English translation, the approaches that it will be proposing are applicable across a wide range of language pairs and will, in brief, consist of using the features, structure, lists and methods of categorisation found within Wikipedia itself in order either to compile a body of translated material or more generally to build up a picture of the encyclopaedia's translation landscape.

There are perhaps two main uses that such material can be put to within the context of translation research. Firstly, it can simply be treated as an end in itself with the general aim of ascertaining where translated material typically tends to be located and what the main factors are which determine the likelihood of translation being used. This I consider to be of great importance as it could potentially contribute to a clear understanding of the processes of knowledge transfer that exist between different language versions of this highly multilingual on-line encyclopaedia – even if the eventual conclusion were to be that the various language Wikipedias are best described as random patchworks of material of varying provenance. The second potential application for the material that is uncovered is as raw data for researching specific aspects of this complex new type of translation. The ensuing research can in fact focus on many different topics within each of the three broad areas of product-, process- and function-oriented translation research (Holmes 2004, 184-5).

As regards the first of these three, a set of data relating to Wikipedia translation would cast light on the following kinds of area:

- what gets translated;

- the characteristics of translated text in Wikipedia;

- the range – possibly a wide one – of transfer relationships – some familiar, others very new and unusual – between articles in different languages; and

- the nature of Wikipedia articles as a hybrid type of writing (i.e. in this context part translation, part original).

Data could also be used to study the following kinds of process-related phenomenon:

- translating, editing and revising practices, for example as they vary from one language Wikipedia to another;

- the evolution of Wikipedia translations (and, more generally, texts) through successive drafts;

- the different translation-related roles and the kinds of interaction and collaboration that take place;

- more generally, the respects in which Wikipedia translation is a kind of collaborative translation, leading to a broadening of our understanding of this phenomenon;

- the range of translation strategies that exist in the encyclopaedia, and also that of interlingual knowledge transfer processes that exist alongside translation; and

- the interlingual pathways typically taken in the process of knowledge transfer.

Finally, those interested in function-oriented research will find data that is useful for investigating matters such as the following:

- the role played by cultural adaptation, ideology and, possibly, political censorship in the encyclopaedia;

- the use of Wikipedia in translation pedagogy;
- the process of selecting material for translation;
- the identity, profile and individual and collective activity of translators (or, more properly, ‘editor-translators’) involved in the co-creation of articles, as well as their interests, priorities and concerns;
- the ethical considerations that appear to hold sway; and
- the formation of a concept that may be referred to as ‘translation and the web’, and how Wikipedia translation contributes to our understanding of it.

The above lists of Wikipedia-related topics for translation research are non-exhaustive, and can no doubt be quickly extended by scholars with a particular interest in one or other area. Within the context of the two uses presented above, the immediate aim of the article is to make some initial proposals towards a methodology for identifying translated material (and gathering other data relevant to translation) and for formulating generalisations as to where the translation activity is located within the on-line encyclopaedia, in terms of language pair, subject matter and other possible factors.

2.1. Wikipedia and Wikipedia translation

As should by now be abundantly clear, one of the peculiarities of Wikipedia translation is the sheer lack of clarity as to the amount, nature and location of translated material that it contains: consequently, before you can study it you need to find it. Over the fifteen years of its existence Wikipedia has evolved into a highly complex structure. Firstly, every article shares a number of common features within its ‘anatomy’, including a) the “Revision History” where every single edit is recorded in chronological order and each successive version of the evolving article is archived in an easily retrievable form; b) the “Talk Page”, which permits editors to discuss topics of relevance to the development of a particular article, including those pertaining to translation; and c) the list of “interwiki links”, which direct readers to parallel articles on the same topic in different language Wikipedias. The first two of these are located in tabs at the top of the article while the third is listed on the left-hand side of the screen below a number of other clickable links. Secondly, the encyclopaedia has evolved into a complex category structure (Suchecki et al. 2012) that permits each article to be tagged as belonging to a wide range of different subject areas, and also a grading scheme whose classification ranges from “featured article” (i.e. professional standard) to “stub” (i.e. incipient). These are both features that can be used to our advantage as will be seen below.

Thirdly, articles are marked up, often with considerable amounts of metadata (in the form of “templates”, for example) and the information that they contain can be supplemented and enriched through interaction with other Wikimedia projects (such as Wikidata and Wikimedia Commons). Fourthly, Wikipedia’s richly collaborative editorial structure (as depicted, for example, by Brandes et al. 2009) makes for an environment that permits individual editors (who may also be translators) to interact within fluid, ad hoc editing networks in order to add to and modify content in an incremental manner, to produce wish-lists of actions (including translations) that are needing to be carried out, and to patrol and monitor groups of articles in order to ensure that unhelpful edits are quickly removed.

The fifth type of complexity that we see is that Wikipedia as a whole is itself an “evolving continuity” (“Meta:Translate Extension” 2016). Not only that, but each article – or, in the present context, each source and target text – can be thought of as a ‘moving object’. Content does not stay still over time, which means that a pair of articles that display a certain type or degree of translation equivalence at a particular point in time may not continue to do so to the same degree as time passes, as the equivalence will tend to ‘decay’ over time as each text evolves largely independently from the other. Untangling the joint evolution of a pair (or group) of articles can be extremely complex and yet is a prerequisite for studying some aspects of Wikipedia translation.

2.2. The Wikipedia research ecosystem

What I refer to with this term is perhaps one of the main characteristics that sets research into Wikipedia translation apart from just about any other type of translation research: what is available within the Wikipedia sites, without the need for supplementation with external data, in effect approaches a complete, self-contained research environment for the investigator who is interested in exploring all kinds of product, process or function-related aspects of this area. Among other elements, this ecosystem includes the following:

- research material in many or all of the 295 languages in which Wikipedia versions currently exist[2];

- sets of translation guidelines, along with statistics and analysis;

- significant amounts of metadata, for example in the form of categorisation mark-up and templates; this is very important as it permits the automatic generation of vast numbers of lists, including lists of pages translated from particular languages, and also those that need to be translated (in someone's view) either from or into specific languages (see below);

- information pages for editors, including details of the etiquette, ethos and editorial structure of each language version of the encyclopaedia;

- details of and access to the translators themselves as well as full records of their activity (e.g. via individual editors' user pages);

- particulars of organised translation projects such as the "Medicine Translation task force"; and

- importantly, the enduring availability of all intermediate versions of virtually every article, along with all metadata, discussions and – sometimes – arguments and edit wars that have been generated; this in effect amounts to a vast virtual archive of every edit (including those involving translation) that allows the researcher to track with great precision the 'evolving continuity' of a particular bilingual pair of articles (for example).

As stated by Yasseri et al., the totality of this archived information provides a unique opportunity to study "the laws of peer production, the process of self-organization of hierarchical structures […] and the occurring regional and cultural differences" (2014, 25). Some but not all of the above bullet points will be elaborated on below.

The encyclopaedia makes frequent use of information boxes known as templates, which are generated by the insertion of a short text string into a page's code. According to Ayers et al., these are used "as navigational and formatting aids and to add recurring or boilerplate messages to pages in a consistent way" (2008, 270). Some of the best known of these indicate that an article is a stub, requires cleanup or has other issues. In addition, there are a number of

[2] All figures and lists taken from Wikipedia articles are accurate as of 4 November 2016.

translation-specific templates, some of which are listed at “Category:Wikipedia translation templates” (2016), while some of the approaches to data gathering discussed below are dependent on the presence of a particular template.

Combined with that, Wikipedia is a collection of lists *par excellence*, lists being automatically generated on the basis of the categories in which articles are included by editors, once again by means of the insertion of a short text string into the code. Thus for example, the article on Albert Einstein belongs to a total of 76 categories, ranging from the more obvious “Nobel laureates in Physics” and “Relativity theorists” to the less predictable “American agnostics” and “Subjects of iconic photographs”; clicking on any of these categories at the end of the article takes one to a full listing of members.

If you consult the “List of Lists of Lists” (2016) you will quickly gain an idea of the importance of list generation within the Wikipedia enterprise. The entries on this mega-list inevitably reflect the wide spectrum of content that has been created in the encyclopaedia, and include for example the following:

- Lists of cities by country
- Lists of solar eclipses
- Lists of Muslim scientists and scholars
- Lists of people from Quebec by region
- Lists of hospitals in Oceania
- Lists of vampires
- Lists of Star Trek planets

Interestingly, most or all of the 36 interwiki links from this page appear to lead to “lists of lists” rather than to “lists of lists of lists”.

3. Approaches to data gathering

We are now in a position to consider possible approaches to identifying pages that contain translated material and locating other translation-relevant data. These will be presented in four sections reflecting the wide range of methodologies that exist: locating relevant pages manually, via elements of page anatomy, by referring to lists of relevant pages and by consulting other resources. As discussed above, the point is not to look for random translated material for the sake of it: such a search for data will almost always be tied into a project that has specific research goals so it will be targeted as appropriate.

3.1. The manual approach

Putative translation pairs can of course be spot-checked manually, for example by following the appropriate interwiki link, although this is a painstaking, hit-and-miss approach. On the other hand, it can be used as a quick method for following up hunches or checking for the presence of translation within a specific test-case context. Additionally, the very fact that there is no guarantee of locating translated material increases the likelihood of finding other kinds of influence between articles from different language versions of the encyclopaedia; this was in fact one of the main methods that I used when preparing the list of 'theoretical transfer scenarios' that is contained in Shuttleworth (2015).

3.2. Lists of pages

True to form, Wikipedia contains a number of translation-related lists of pages, and depending on the particular aim of the research, consulting one or other of these can be a useful and relatively quick way of discovering, for example, what has been translated between two languages or what, in the view of a particular group of people (or even a single editor), is in need of translation. Once again depending on the precise nature of the research, it may need to be followed up, for example by using the third approach, in order to discover precisely what the extent of the translation is, where it is located within an article, when it took place, who performed it and so on. There are in fact at least three relevant types of list, and these will now be considered in turn.

3.2.1. "Pages translated from…". Many but by no means all versions of Wikipedia contain lists of articles translated from other language versions. In the case of the English Wikipedia, these lists are all themselves listed in "Category:Translated pages" (2016), as exemplified in Figure 1.

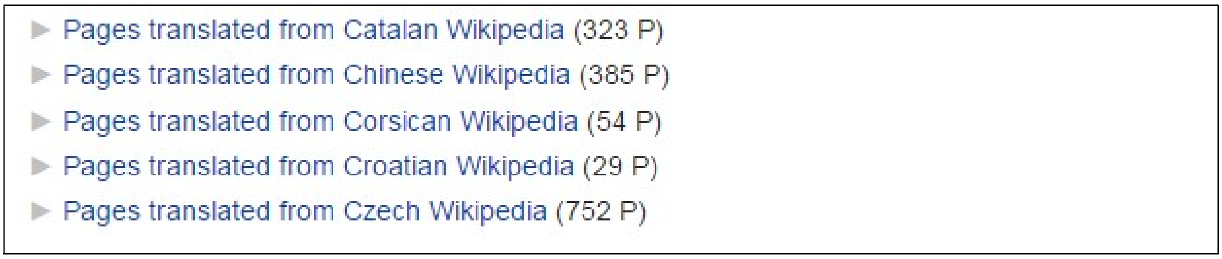

► Pages translated from Catalan Wikipedia (323 P)
► Pages translated from Chinese Wikipedia (385 P)
► Pages translated from Corsican Wikipedia (54 P)
► Pages translated from Croatian Wikipedia (29 P)
► Pages translated from Czech Wikipedia (752 P)

Figure 1: Extract from “Category:Translated pages” (2016), showing links to all the Wikipedias beginning with the letter C from which pages have been at least partially translated into English. In each case, the total number of pages is given in parentheses.

According to this page, there are in fact 68 language versions of the encyclopaedia for which pages exist listing articles that have been at least partially translated into English from the corresponding article in that language. (In practice the articles listed may only contain a small amount of translated material.) Inclusion of a particular article on a list depends on the appropriate template having been placed on the article’s Talk Page, as illustrated in Figure 2.

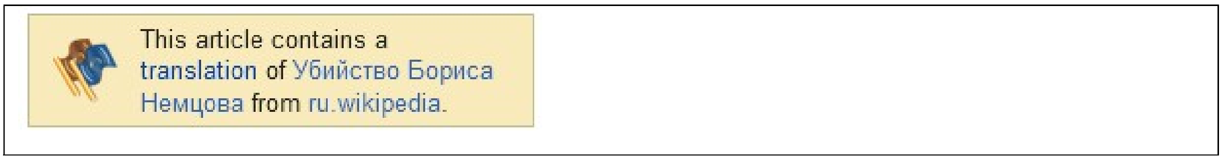
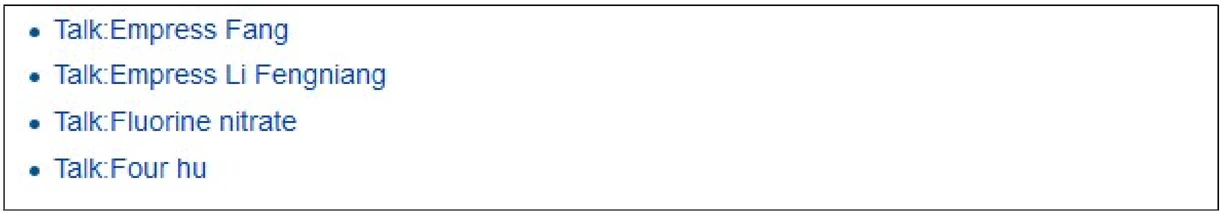

This article contains a translation of Убийство Бориса Немцова from ru.wikipedia.

Figure 2: Template added to the Talk Page of the English-language article on the “Assassination of Boris Nemtsov” (2017) to indicate that the article contains some material translated from the Russian Wikipedia article on the same topic.

The lists range in size from the single pages listed for Neapolitan, Pennsylvania German, Urdu and Võro to 4,091 articles translated from Spanish, 9,149 from French and 18,207 from German. Some are thus clearly very substantial, although because of their dependence on the presence of a particular template within an article’s Talk Page most or all of them are almost certainly incomplete, as there is nothing to force an editor to declare the use of translation in this manner. All five language versions listed in Figure 1 fall somewhere in the middle; what is immediately obvious from all these above figures, however, is the fact that there is no absolute correlation between the number of native speakers of a language and the number of pages that are translated into English from its native Wikipedia. The figures for Russian and Chinese, which are discussed later in this section, are 1532 and 385 respectively.

If you click through to a particular list you will be presented with an alphabetical listing of all the relevant pages, as exemplified in Figure 3.



- Talk:Empress Fang
- Talk:Empress Li Fengniang
- Talk:Fluorine nitrate
- Talk:Four hu

Figure 3: Extract from "Category:Pages translated from Chinese Wikipedia" (2016), consisting of links to all pages categorised under F that contain material translated from the corresponding page in the Chinese Wikipedia.

If we look in more detail at "Category:Pages translated from Chinese Wikipedia" (2016) it is interesting to observe that most if not all the pages listed relate to topics that are specific in nature (including, for example, minor historical figures, local Chinese TV stations and stub articles on obscure chemical compounds). Interestingly, few or no major topics are mentioned: to cite three more or less random examples, none of the English-language articles on any of the ten largest cities in China (as listed at "List of cities in China by population" 2016), the Four Great Classical Novels (*The Water Margin*, *The Romance of the Three Kingdoms*, *Journey to the West* and *Dream of the Red Chamber*) or the three most recent presidents of the People's Republic of China (Jiang Zemin, Hu Jintao and Xi Jinping) contain any content declared as having been translated from Chinese. One must always exercise caution before drawing conclusions based on something's absence but I believe this to be indicative of a more general trend.

Indeed, similar tendencies can be observed in "Category:Pages translated from Russian Wikipedia" (2016). If we conduct searches that are broadly similar to those carried out above – Russia's ten largest cities ("List of cities and towns in Russia by population" 2016), four great nineteenth century novels (*War and Peace*, *Anna Karenina*, *Crime and Punishment* and *The Brothers Karamazov*) and the country's three post-1991 presidents (Boris Yeltsin, Vladimir Putin and Dmitry Medvedev) – we find that articles relating to only two of these seventeen major topics (Russia's fourth-largest city Yekaterinburg and Boris Yeltsin) are listed.

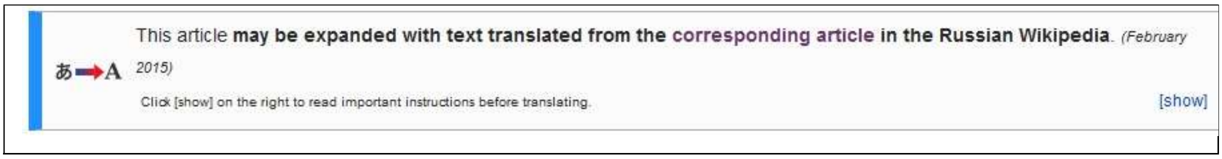
On the other hand, articles relating to specific aspects of at least some of these topics can be found on each of these two lists. For example, while Beijing itself does not feature on the Chinese list, the articles on Beijing city fortifications, the Beijing Planning Exhibition Hall and the Beijing Youth Daily are included. Along similar lines, Moscow itself is absent from the Russian list, although eight articles on topics connected to the capital, including the Moscow

City Duma elections of 2005 and 2009, the Moscow Mint and the Moscow State Art and Cultural University, can be found.

It would of course be interesting to know if this marked tendency to confine translation effort to subordinate rather than superordinate topics can be observed across all language versions of the encyclopaedia for which equivalents of this list are available – including of course the case of translation out of English or where English does not form part of the language pair. The provisional finding could simply reflect the mature state that has been reached by versions of the encyclopaedia that relate to major world languages. What would the situation have been like ten years ago, or with a version that has very few entries? Such an investigation would not be a massive undertaking, either: according to the interwiki links at "Category:Translated pages" (2016) – interestingly enough, a collection in which some of the 'larger' Wikipedia languages (such as French, German and Japanese) are not represented, while a number of 'smaller' languages (such as Min Dong Chinese) on the other hand are – a comparable list exists in a total of 25 other language versions of Wikipedia, although in the case of some of these (e.g. Portuguese, Scots, Simple English and Tagalog) the pages are not subcategorised according to source language, thus making the quick gathering of numerical information more problematic. However, pending a careful analysis of all the pages linked to by the interwiki links, and of course a more detailed investigation into the typical profile of pages translated into English, the significance of the tendency that has been provisionally identified in the previous two paragraphs appears to be that, in the case of the English Wikipedia at least, the Wikipedia of the language to which the subject matter is most closely related does not appear to be a significant source for translation (when translation is opted for at all) in the case of high-profile topics, for which alternative high quality target language information resources will presumably be readily available. On the other hand, it should of course be pointed out that other types of influence such as paraphrase cannot be ruled out. In addition, it is also possible that this finding is an artefact of topic structure – there are, after all, far fewer articles on such high-profile topics than there are on specific ones and so one would expect them to feature on such a list less prominently – although on the other hand such superordinate pages are likely to attract far more long-term interest on the part of editors. However, if it turns out that this observation can be generalised across a wide range of topics and languages it would perhaps still represent a fairly significant finding. In addition, the question of how the pages in such lists are thematically grouped has not even been touched on in this brief discussion. Quite clearly, there exists here an opportunity for further work to understand the pathways and methods of inter-Wikipedia influence and knowledge transfer; an initial step in this direction is offered in Shuttleworth (2015, forthcoming).

An additional possible research approach would be to revisit these pages at intervals to discover, for example, the rate of translation and the changing interests of the translators. If this were to be done then snapshots would need to be taken on each visit as, unlike the majority of Wikipedia pages, these lists are not continuously archived.[3]

3.2.2. "Articles needing translation from…". A way of identifying and accessing interesting data that is complementary to this is provided by "Category: Articles needing translation from foreign-language Wikipedias" (2016), which provides links to 145 language-specific pages (and also to other lists, for example of featured articles) that present a collection of English-language articles that are recommended for expansion via translation from the corresponding article in another language Wikipedia. As is the case with all the pages linked to from "Category:Translated pages" (2016), the inclusion of a particular article in one of these lists is automatically triggered by the presence of the relevant template, this time at the beginning of the article itself, as illustrated in Figure 4.



This article **may be expanded with text translated from the corresponding article in the Russian Wikipedia**. *(February 2015)*

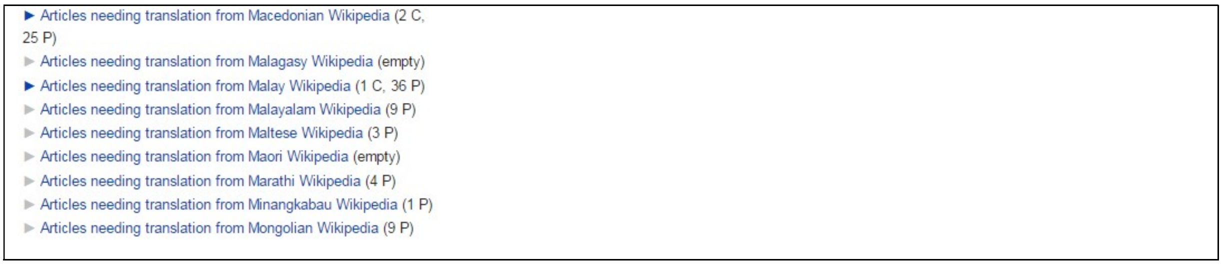
Click [show] on the right to read important instructions before translating.

[show]

Figure 4: The template placed in an article in the English-language Wikipedia to suggest that it should be expanded by means of translation.

An extract from the list is shown in Figure 5.

[3] It should be pointed out at this stage that the pages studied in this and the next section are only edited manually very rarely, while automatic updating occurs continuously. For this reason, they are being referenced by the date when they were accessed rather than by that of the last edit.



► Articles needing translation from Macedonian Wikipedia (2 C, 25 P)
► Articles needing translation from Malagasy Wikipedia (empty)
► Articles needing translation from Malay Wikipedia (1 C, 36 P)
► Articles needing translation from Malayalam Wikipedia (9 P)
► Articles needing translation from Maltese Wikipedia (3 P)
► Articles needing translation from Maori Wikipedia (empty)
► Articles needing translation from Marathi Wikipedia (4 P)
► Articles needing translation from Minangkabau Wikipedia (1 P)
► Articles needing translation from Mongolian Wikipedia (9 P)

Figure 5: Extract from "Category:Articles needing translation from foreign-language Wikipedias" (2016), consisting of links to lists of pages recommended for expansion via the translation of material from pages in versions of the encyclopaedia beginning with the letter M. Listings of Wikipedias with a number of categories as well as pages indicated can be expanded by clicking on the arrow to the left (C = categories; P = pages).

As can be seen, lists are often partially classified by subject category (such as culture, geography or sports). The list includes 35 empty languages and 29 with only a single page, while twelve languages have more than 500 pages tagged as being in need of translation: Vietnamese (544), Chinese (756), Polish (674), Dutch (743), Portuguese (793), Czech (1041), Italian (1283), Japanese (1300), Russian (1403), Spanish (2893), German (4524) and French (8211). In each case the English page, which must already exist, if only in stub form, is linked to.

Such lists can potentially show us, for example, a particular source language Wikipedia community's level of activity, as well as the interests and priorities of some of its members: there is clearly no reason, for example, to expect similar selections of pages suggested for translation from language B to A as those for language A to B. What appears not to be stated, however, is the precise criteria that determine why a particular page is deemed to be "needing translation". Although the suspicion here has to be that, in many cases at least, pages are tagged for automatic addition to the appropriate list for reasons that are largely subjective and personal to an individual editor, this is quite clearly a question that deserves further investigation. What do the collective interests and priorities of each Wikipedia community appear to be? What are the implications of this? How significant a tool is translation seen to be for expanding its scope? What are the most usual source languages? Also, is there a preference for adding controversial articles, or ones with a high degree of self-focus, for example, to these lists? Indeed, I consider that a study of the precise foci within each list of desired pages would be highly revealing of the collective interests of the most active editors of a particular language version – and, consequently, of the 'image' that is projected to the world with respect to that version of the encyclopaedia.

Each language-specific page consists mainly of an alphabetical list of articles, although as mentioned this is preceded in many cases by links to pages belonging to different categories (e.g. Biography, Featured, Geography or Science). These categories contain listings of pages that are not included in the main list. As can be seen from the figures quoted above, for some languages the alphabetical listing of pages is very long. This is, for example, the case with the Chinese, Russian and Ukrainian pages, which we will now look at.

"Category:Articles needing translation from Chinese Wikipedia" (2016) contains a listing of 756 pages and includes a Featured category of 24 pages and three thematic categories (all of which are empty). None of the 'high-level' items discussed in the previous section – i.e. large cities, famous novels or recent presidents – are listed here either, possibly suggesting a similar absence of pages relating to such topics from these lists. Apart from this, few clear trends are discernible among the more than 700 pages listed, although significant numbers of geographical entities (especially within Hong Kong and Taiwan) are listed, and also of historical figures (including notably large numbers of empresses). Finally, there are a small number of pages devoted to political topics (including specific elections and anti-government protests).

"Category:Articles needing translation from Russian Wikipedia" (2016) lists a total of 1403 pages; it includes ten categories, the largest by far being Featured (55) and Geography (42). Once again, it lists no famous nineteenth-century novels or recent presidents, and only one of the top-ten cities, within the Geography category. On the other hand, the main listing includes significant numbers of military-related pages as well as articles about people, both contemporary and historical, and both Russians and representatives of the other nationalities of Russia and the former USSR (including, most notably, a relatively large number about Georgians). Like the Chinese page, it contains little material that would indicate an interest in political activism amongst the Wikipedians who have tagged articles for inclusion there.

It should be noted that an inspection of "Category:Articles needing translation from Ukrainian Wikipedia" (2016) also reveals the same tendency. This is possibly more surprising given that at present this country is to a large extent defined in the eyes of the world by political and economic turmoil and by the hybrid war of aggression that is currently being waged on its territory. No high-profile politicians or political prisoners are listed, for example, and overall there are very few pages that have any direct bearing on the conflict, a silence that is noteworthy, if only for what it tells us of the encyclopaedia's self-identified function.

In spite of these initial observations made with reference to three "Articles needing translation from…" pages within the English Wikipedia, the general conclusions seem as yet unclear, not least because all the pages looked at above relate to major languages. Once again, considerable opportunity exists for further work to untangle the interrelationship between collective identity, the use of translation and the use of Wikipedia as a possible vehicle for articles relating to a particular subject area.

What we can probably say, however, even if it does transpire that inclusion does not by and large result from a series of coherent or interlinking priorities, is that it is likely that certain clusterings of articles will be present or, at the very least, that it will be possible to make certain generalisations about the likely nature of articles included, along the lines of those that were suggested in the previous section. That said, it appears that, in many cases at least, the precise foci of the desired or requested translation effort are not always easy to identify, as the brief, very tentative initial inquiry above has testified. Revisiting the pages from time to time should also indicate what the rate of translation turnover was.

"Category:Articles needing translation from foreign-language Wikipedias" (2016) appears to be mirrored in other language versions much less readily than is the case with "Category:Translated pages" (2016), interwiki links existing only for four other versions (Italian, Japanese, Venetian and Chinese), each of which appears to have its own distinct features that would be worth investigating.

3.2.3. "Pages needing translation into English". "Wikipedia:Pages needing translation into English" (2016) has recently been discussed by McDonough Dolmaya (2015) so relatively little space is being devoted to it here. This type of page differs from the other two discussed as it exists for the purpose of problem solving: firstly, for creating English translations for pages in the English Wikipedia that are in fact either partially or wholly written in another language, and secondly, for post-editing pages that have already been translated but are in need of some clean-up. The page lists 20 of the former kind of article (added between 19 January and 4 November 2016) and 169 of the latter kind (added between 2013 and November 2016). A wide range of source languages are included: for the former kind, the first five articles listed relate to

Norwegian, Chinese, Spanish, French and Somali, while the first five of the latter type are from Catalan, Korean, Romanian and Czech. Pages can be added to the list either automatically through the use of a template or by manually including a mention on the list page itself.

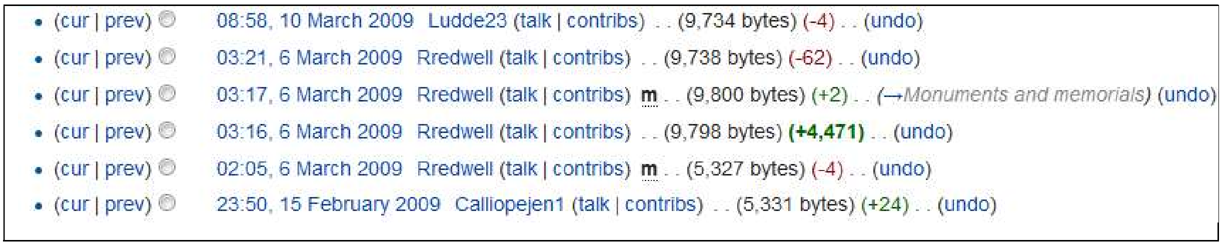
Interwiki links take the reader to 19 other language versions of this page; once again the selection is fairly random, with some minority languages represented but a number of major languages absent. Overall, these pages offer a potentially quite effective way of accessing translated material, particularly since Wikipedia's archiving system permits comparison of 'before' and 'after' versions of the same article. That said, the articles are not presented in anything like a systematic manner. McDonough Dolmaya uses the English page to study the open editing process used in the Wikipedia environment; in addition to this, it can serve as an alternative – possibly more convenient but also more limited – source of translated Wikipedia material, for whatever reason a researcher might have need of it.

3.3. Page anatomy

This third approach involves utilising particular elements of an article's structure in a systematic manner. For this we will be focusing on the Revision History and the Talk Page tabs that form part of every Wikipedia article. As a primary method of data acquisition it is not as efficient as that discussed in Section 3.2; on the other hand, it may be of more value as a kind of second stage of the process, for pinpointing the location of translation, both within the structure of an article and in time.

3.3.1. Revision history. As outlined above, by selecting the Revision History tab it is possible to access a complete listing of every edit performed on almost any Wikipedia article, from its very inception right up to the current version; in other words, with very few exceptions, not a single intermediate version of an article ever goes beyond recall.

When you select this tab you will be presented with what is likely to be a very long list of intermediate article versions, as illustrated in Figure 6 for the article on the Road of Life.

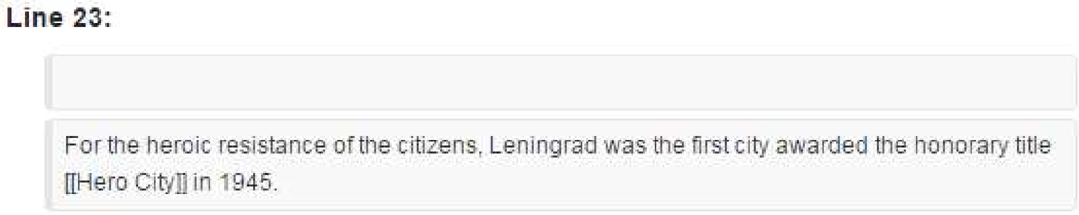


- (cur | prev) ○ 08:58, 10 March 2009 Ludde23 (talk | contribs) . . (9,734 bytes) (-4) . . (undo)
- (cur | prev) ○ 03:21, 6 March 2009 Rredwell (talk | contribs) . . (9,738 bytes) (-62) . . (undo)
- (cur | prev) ○ 03:17, 6 March 2009 Rredwell (talk | contribs) m . . (9,800 bytes) (+2) . . (→*Monuments and memorials*) (undo)
- (cur | prev) ○ 03:16, 6 March 2009 Rredwell (talk | contribs) . . (9,798 bytes) **(+4,471)** . . (undo)
- (cur | prev) ○ 02:05, 6 March 2009 Rredwell (talk | contribs) m . . (5,327 bytes) (-4) . . (undo)
- (cur | prev) ○ 23:50, 15 February 2009 Calliopejen1 (talk | contribs) . . (5,331 bytes) (+24) . . (undo)

Figure 6: Short extract from the list of intermediate versions of the article on the "Road of life" (2017), showing versions produced between 23:50 on 15 February 2009 and 08:58 on 10 March 2009.

As can be seen from the figure, among the information provided is the exact timing of the edit, the editor's username (if this is not available then the IP address of the computer that was used will be supplied instead), the resulting size of the article in bytes and, importantly, the increase or reduction in size caused by the edit (indicated in green or red respectively). If a sudden jump in size is indicated – as occurs in the above figure against the edit performed at 03:16 on 6 March 2009 – this is possible evidence that a significant act of translation has occurred even if it is not explicitly marked as such. In this particular case the edit will only be revealed to consist of an insertion of translated material when the researcher takes the trouble to investigate, although acts of translation are sometimes explicitly tagged, most usually by the presence of the word "translated".

Once a possible translation event has been identified (however this may have been achieved), two – usually adjacent – revisions can be selected and compared, as shown in Figures 7 and 8. The first of these shows an extract of the article before the edit.

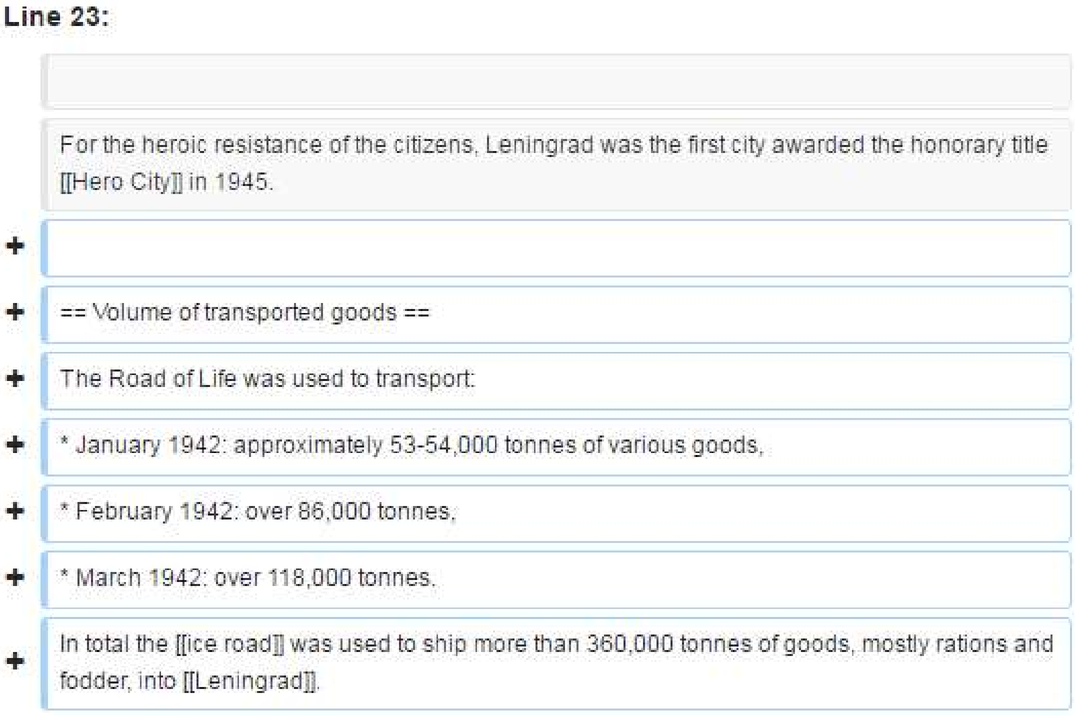


Line 23:

For the heroic resistance of the citizens, Leningrad was the first city awarded the honorary title [[Hero City]] in 1945.

Figure 7: A small extract from the article as it existed as of 02:05, 6 March 2009.

The second indicates the precise textual modifications that were implemented in the edit.


Line 23:

For the heroic resistance of the citizens, Leningrad was the first city awarded the honorary title [[Hero City]] in 1945.

+

+ == Volume of transported goods ==

+ The Road of Life was used to transport:

+ * January 1942: approximately 53-54,000 tonnes of various goods,

+ * February 1942: over 86,000 tonnes,

+ * March 1942: over 118,000 tonnes.

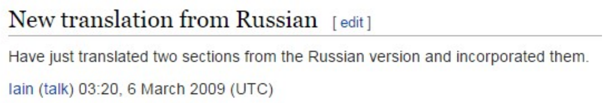
+ In total the [[ice road]] was used to ship more than 360,000 tonnes of goods, mostly rations and fodder, into [[Leningrad]].

Figure 8: The appearance as of 03:16, 6 March 2009 of a new section in the Road of Life article consisting of material translated from the contemporaneous version of the corresponding Russian article.

In this way, versions resulting from different edits can be compared to pinpoint any precise differences that may exist. Of course, given that this is a translation act, depending on the precise aims of the research it is very possible that the source text version that is exactly contemporaneous with the translation edit will also need to be consulted in a similar manner in order to identify a precise source-target pair. Alternatively, if the researcher's interest is less text-orientated, this approach can yield some tantalising insights into Wikipedia-style collaborativity such as the comment "Please keep Template:Expand Russian until the article is fully translated" recorded at 01:40 on 28 February 2015 for the 'Assassination of Boris Nemtsov' article.

Given the fact that translation equivalence can be edited away from in the course of source and target texts' independent movement through progressively evolving versions, the advantage of this method is that it permits the researcher to discover translated portions of articles as they existed at a particular point in time. On the other hand, it is to a large extent dependent on the presence of particular verbal triggers and can be somewhat hit-and-miss in nature.

3.3.2. Talk pages. Occasionally, translation activity is declared or discussed in an article's Talk Page. This is the case with the Road of Life edit just discussed, as shown in Figure 9.



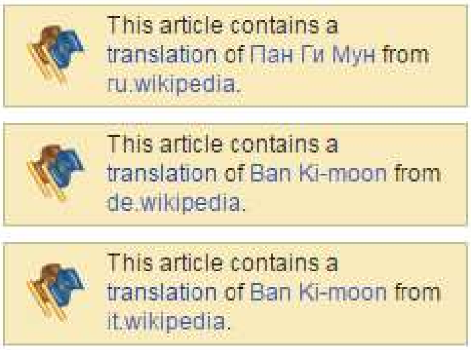
New translation from Russian [edit]

Have just translated two sections from the Russian version and incorporated them.

Iain (talk) 03:20, 6 March 2009 (UTC)

Figure 9: An act of translation is declared on the Talk Page of the Road of Life article.

Besides this brief insertion, a template is also added. Many discussions focus on very specific points and some can be quite extended. The Talk Page of the English Tianxia article, for example, discusses using translation from the Chinese version to redress the balance of the English version, and then documents the translation as it is carried out. Additionally, some articles contain fragments of text that are translated from different Wikipedia sources, as seen for example in Figure 10.



This article contains a translation of Пан Ги Мун from ru.wikipedia.

This article contains a translation of Ban Ki-moon from de.wikipedia.

This article contains a translation of Ban Ki-moon from it.wikipedia.

Figure 10: Ban Ki-moon Talk Page.

This suggests that in this respect – and, no doubt, in others too – some Wikipedia articles are formed like patchworks of textual fragments originating from a variety of sources in different languages.

Unlike Revision History, the contents of Talk Pages and their equivalents in other Wikipedias can easily be probed via a search engine, although the automatic harvesting of results (a practice known as "scraping") would be a breach of most companies' terms and conditions.

3.4. Other approaches

Besides these major resources, Wikipedia also boasts a number of other pages that offer the researcher quick access to translated material. These include the following:

- "Wikipedia WikiProject Medicine_Translation task force" (2016) gives details of the Translation Task Force, an on-going project that is currently aiming to translate a thousand key medical articles into a hundred languages.

- "Translation of the week" (2016) describes a project that has been running since 2004. The weekly translation generally consists of "a stub or the first paragraph of an important article" ("Translation of the week" 2016). According to a list of links, this project also operates in 43 other languages. The page also lists past translations (for 2016) and links to an archive of older ones (2004-2016).

- "Wikipedia:WikiProject Intertranswiki" (2016) provides information on a collaborative effort to "improve Wikipedia by importing and translating content from foreign language Wikipedias", drawing up a directory of content missing from different versions of the encyclopaedia in order to facilitate this.

- In terms of automatic identification of existing translated material, Plamadă and Volk (2013), for example, describe a method for searching for parallel text on the sentence (or segment) level. While this kind of approach has not yet reached maturity, it promises much for the future: it may well be that its eventual perfection and application will be the only way we can build up a complete picture of translation in the encyclopaedia.

- "Wikipedia:Translation" (2016) is one of the main sources for guidance on translating for Wikipedia. The page also includes details of different "userboxes" that editors can add to their User page to indicate their interest in translation. Pages parallel to "Wikipedia:Translation" exist in a total of 49 languages.

Additionally, it is possible that factors other than article topic may have a bearing on the activity of editor-translators. These might include the level of controversy, self-focus or topicality represented by articles, or the relative size of the source and target Wikipedias, for example. Luckily, features such as these can be used to identify sets of articles to be checked for the presence of translation activity. In addition, in the context of a specific research project it is envisaged that the approaches outlined above will be used together in the most expedient combination that suggests itself.

That said, one problem that besets many of the above approaches is the fact that we simply do not know how much undeclared translation exists. This material is like the ‘dark matter’ of Wikipedia translation: it is quite possible that it is present in the encyclopaedia in far greater quantities than the material that is marked as such. The only methods that stand a chance of identifying such translation are the manual approach and one or two of the above bullet-points.

4. Conclusions

The embroidered fabric of Wikipedia does indeed contain a subset of threads that are added by translators, although all too often these are hidden among surrounding text and their presence can often only be revealed through painstaking work on the part of the researcher. Besides the embroidery trope, in this article I have introduced a number of others: Wikipedia as a research ecosystem, articles as moving objects, undeclared translation as dark matter. All of these, I would argue, throw important theoretical light on Wikipedia translation. However, one further one emerges from the discussion, and this is ‘Wikipedia as a labyrinth’: there are numerous pathways for negotiating this huge encyclopaedia, only some of which will bring us to our goal, and unfortunately we will not know in advance which pathway to choose.

The approaches covered above have at least some bearing on all the items listed under the three headings of product, process and function in Section 2 above. Although there has been insufficient space to consider them all in equal depth, these approaches can be used (for example) to accomplish the following:

- identify translated material, desired translation and routes of influence between Wikipedias;
- investigate the main foci of the translation effort;
- determine the many factors determining the likelihood of translation being used in a particular context;
- ascertain the kinds of translation and translation-related phenomenon that can be found on Wikipedia;
- examine the activity of translators both individually and collectively;
- explore the light that the encyclopaedia versions cast on collaborativity in translation; and

- understand the complex joint evolution of a multilingual group of pages.

These are initial, fairly general suggestions and this list will no doubt be developed and refined as the research effort gets properly underway. The emphasis in this article has been to propose methodologies and outline research agendas. That said, one initial research finding that appears to have emerged along the way is that inter-Wikipedia translation seems to be found in specific rather than general topics. Further findings, based on translated material I have identified, are reported on by Shuttleworth (2015, forthcoming).

References

Alonso, Elisa. 2015. "Analysing the Use and Perception of Wikipedia in the Professional Context of Translation". *The Journal of Specialised Translation* 23: 89-117. Accessed 4 November 2016. www.jostrans.org/issue23/art_alonso.pdf.

"Assassination of Boris Nemtsov". 2017. *Wikipedia*. Accessed 13 August 2017. https://en.wikipedia.org/wiki/Assassination_of_Boris_Nemtsov, date updated 13 July 2017.

Ayers, Phoebe, Charles Matthews, and Ben Yates. 2008. *How Wikipedia Works and How You Can Be a Part of It*. San Francisco: No Starch Press, Inc. [Also available at http://en.wikibooks.org/wiki/How_Wikipedia_Works.]

"Benjamins Translation Studies Bibliography". 2017. Edited by Yves Gambier and Luc van Doorslaer. Accessed 13 August 2017. https://www-benjamins-com.libproxy.ucl.ac.uk/online/tsb/.

Brandes, Ulrik, Patrick Kenis, Jürgen Lerner, and Denise van Raaij. 2009. "Network Analysis of Collaboration Structure in Wikipedia". Paper given at the *World Wide Web Conference 2009*, April 20-24 2009, Madrid. Accessed 4 November 2016. http://cs.smith.edu/classwiki/images/a/a9/Network_analysis_collaboration_wikipedia.pdf.

"Category:Articles Needing Translation from Chinese Wikipedia". 2016. *Wikipedia*. Accessed 4 November 2016. https://en.wikipedia.org/wiki/Category:Articles_needing_translation_from_Chinese_Wikipedia.

"Category:Articles Needing Translation from Foreign-language Wikipedias". 2016. *Wikipedia*. Accessed 4 November 2016.

https://en.wikipedia.org/wiki/Category:Articles_needing_translation_from_foreign-language_Wikipedias.

"Category:Articles Needing Translation from Russian Wikipedia". 2016. *Wikipedia*. Accessed 4 November 2016. https://en.wikipedia.org/wiki/Category:Articles_needing_translation_from_Russian_Wikipedia.

"Category:Articles Needing Translation from Ukrainian Wikipedia". 2016. *Wikipedia*. Accessed 4 November 2016. https://en.wikipedia.org/wiki/Category:Articles_needing_translation_from_Ukrainian_Wikipedia.

"Category:Pages Translated from Chinese Wikipedia". 2016. *Wikipedia*. Accessed 4 November 2016. https://en.wikipedia.org/wiki/Category:Pages_translated_from_Chinese_Wikipedia.

"Category:Pages Translated from Russian Wikipedia". 2016. *Wikipedia*. Accessed 4 November 2016. https://en.wikipedia.org/wiki/Category:Pages_translated_from_Russian_Wikipedia.

"Category:Translated Pages". 2016. *Wikipedia*. Accessed 4 November 2016. https://en.wikipedia.org/wiki/Category:Translated_pages.

"Category:Wikipedia Translation Templates". 2016. *Wikipedia*. Accessed 4 November 2016. http://en.wikipedia.org/wiki/Category:Wikipedia_translation_templates.

Désilets, Alain, Lucas Gonzalez, Sébastien Paquet and Marta Stojanovic. 2006. "Translation the Wiki Way". *Proceedings of WikiSym 2006 – The 2006 International Symposium on Wikis*, Odense, Denmark, August 21-23, 2006. Accessed 4 November 2016; log-in needed. http://dl.acm.org/citation.cfm?id=1149464.

Fernández Costales, Alberto. 2013. "Crowdsourcing and Collaborative Translation: Mass Phenomena or Silent Threat to Translation Studies?' *Herměneus. Revista de Traducción e Interpretación*, 15: 85-110. Accessed 4 November 2016. http://recyt.fecyt.es/index.php/HS/article/viewFile/30295/15892.

Holmes, James S. 2004. "The Name and Nature of Translation Studies". in *The Translation Studies Reader*, 2nd Edition, edited by Lawrence Venuti, 180-192. London: Routledge.

"How will massive online collaboration impact the world of translation". n.d. Accessed 4 November 2016. http://www.wiki-translation.com/How+will+massive+online+collaboration+impact+the+world+of+translation.

Jiménez-Crespo, Miguel A. 2017. *Crowdsourcing and Online Collaborative Translations: Expanding the Limits of Translation Studies*. Amsterdam & Philadelphia: John Benjamins.

Liao, Han-teng and Thomas Petzold. 2010. "Analysing Geo-linguistic Dynamics of the World Wide Web: The Use of Cartograms and Network Analysis to Understand Linguistic Development in Wikipedia". *Journal of cultural science*, 3 (2): 1-18. Accessed 4 November 2016. http://cultural-science.org/journal/index.php/culturalscience/article/view/44/128.

"List of Cities and Towns in Russia by Population". 2016. *Wikipedia*. Accessed 4 November 2016. https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_Russia_by_population, date updated 19 October 2016.

"List of Cities in China by Population". 2016. *Wikipedia*. Accessed 4 November 2016. https://en.wikipedia.org/wiki/List_of_cities_in_China_by_population, date updated 28 August 2016.

"List of Lists of Lists". 2016. *Wikipedia*. Accessed 4 November 2016. https://en.wikipedia.org/wiki/List_of_lists_of_lists, date updated 29 October 2016.

McDonough Dolmaya, Julie. 2015. "Revision History: Translation Trends in Wikipedia". *Translation Studies* 8(1): 16-34. Accessed 5 November 2016. DOI 10.1080/14781700.2014.943279.

"Meta:Translate Extension". 2016. *Wikimedia*. Accessed 4 November 2016. http://meta.wikimedia.org/wiki/Meta:Translate_extension, date updated 2 March 2016.

Plamadă, Magdalena and Martin Volk. 2013. "Mining for Domain-specific Parallel Text from Wikipedia". *Proceedings of the 6th Workshop on Building and Using Comparable Corpora*, Sofia, Bulgaria, 8 August 2013, 112-120. Accessed 4 November 2016. www.aclweb.org/anthology/W13-2514.

"Road of life". 2017. *Wikipedia*. Accessed 13 August 2017. https://en.wikipedia.org/wiki/Road_of_Life, date updated 11 August 2017.

Shuttleworth, Mark. 2015. "Wikipedia Translation: Collaborativity, Translation and the Web." *IATIS 5th International Conference*, Belo Horizonte, Brazil, 7-10 July 2015, slides available at https://www.academia.edu/13808267/5th_IATIS_Conference_Wikipedia_presentation. Accessed 24/4/2017.

Shuttleworth, Mark. Forthcoming. "Translation in Wikipedia Articles and its Contribution to the Spread and Creation of Knowledge". To appear in *Alif* 38.

Suchecki, Krzysztof, Alkim Almila Akdag Salah, Cheng Gao, and Andrea Scharnhorst. 2012. "Evolution of Wikipedia's Category Structure". *Advances in Complex Systems*, 15 supp01. Accessed 4 November 2016; log-in needed). DOI 10.1142/s0219525912500683.

"Translation of the week". 2016. *Wikipedia*. Accessed 4 November 2016. http://meta.wikimedia.org/wiki/Translation_of_the_week, date updated 31 October 2016.

"Wikipedia:Pages Needing Translation into English". 2016. *Wikipedia*. Accessed 4 November 2016. https://en.wikipedia.org/wiki/Wikipedia:Pages_needing_translation_into_English, date updated 4 November 2016.

"Wikipedia:Translation". 2016. *Wikipedia*. Accessed 4 November 2016. http://en.wikipedia.org/wiki/Wikipedia:Translation, date updated 27 October 2016.

"Wikipedia:WikiProject Intertranswiki". 2016. *Wikipedia*. Accessed 4 November 2016. https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Intertranswiki, date updated 2 November 2016.

"Wikipedia WikiProject Medicine_Translation task force". 2016. *Wikipedia*. Accessed 4 November 2016. http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Translation_task_force, date updated 14 October 2016.

Yasseri, Taha, Anselm Spoerri, Mark Graham, and János Kertész. 2014. "The Most Controversial Topics in Wikipedia: A Multilingual and Geographical Analysis". in *Global Wikipedia: International and Cross-cultural Issues in Online Collaboration*, edited by Pnina Fichman and Noriko Hara, 25-48. Lanham, Maryland: Rowman and Littlefield.