Royal Netherlands Academy of Arts and Sciences (KNAW)
KONINKLIJKE NEDERLANDSE AKADEMIE VAN WETENSCHAPPEN

Beyond the Book: Linking Books to Wikipedia

Martinez-Ortiz, Carlos; Koolen, Marijn; Busschenhenke, Floor; van Dalen-Oskam, K.H.

2015

document version
Publisher's PDF, also known as Version of record

document license
CC BY-NC-ND

Link to publication in KNAW Research Portal

citation for published version (APA)
Martinez-Ortiz, C., Koolen, M., Busschenhenke, F., & van Dalen-Oskam, K. H. (2015). *Beyond the Book: Linking Books to Wikipedia*. 12-21. http://conferences.computer.org/escience/2015/papers/9325a012.pdf

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the KNAW public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain.
• You may freely distribute the URL identifying the publication in the KNAW public portal.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
pure@knaw.nl

Download date: 11. Dec. 2024

Beyond the Book: Linking Books to Wikipedia

Carlos Martinez-Ortiz
Netherlands eScience Centre
Email: c.martinez@esciencecenter.nl

Marijn Koolen
Institute for Logic, Language and Computation
Universiteit van Amsterdam
Email: marijn.koolen@uva.nl

Floor Buschenhenke and Karina van Dalen-Oskam
Department of Literary Studies
Huygens ING
Email: karina.van.dalen@huygens.knaw.nl

***Abstract*—The book translation market is a topic of interest in literary studies, but the reasons why a book is selected for translation are not well understood. The *Beyond the Book* project investigates whether web resources like Wikipedia can be used to establish the level of cultural bias.**

This work describes the eScience tools used to estimate the cultural appeal of a book: semantic linking is used to identify key words in the text of the book, and afterwards the revision information from corresponding Wikipedia articles is examined to identify countries that generated a more than average amount of contributions to those articles. Comparison between the number of contributions from two countries on the same set of articles may show with which knowledge the contributors are familiar. We assume a lack of contributions from a country may indicate a gap in the knowledge of readers from that country. We assume that a book dealing with that concept could be more exotic and therefore more appealing for certain readers, while others are therefore less interested in the book. An indication of the 'level of exoticness' thus could help a reader/publisher to decide to read/translate the book or not.

Experimental results are presented for four selected books from a set of 564 books written in Dutch or translated into Dutch, assessing their potential appeal for a Canadian audience. A qualitative assessment of quantitative results provides insight into named entities that may indicate a high/low cultural bias towards a book.

I. Introduction

In literary studies, the impact of a global society is a topic of current interest. Are books (fiction and non-fiction) from different cultural backgrounds more and more alike? Or do we find the opposite: an increase in books that highlight cultural differences? Or both? We would like to investigate this by taking into account trends in the selection of books for translation. In recent decades there has been an increase in the number of books (fiction and non-fiction) which get translated; many of which are available in digital form (see Figure 1). Wikipedias in many languages have appeared. Our assumption is that we can explore globalization trends by using eScience tools to link books to Wikipedia and to analyse the results. In the *Beyond the Book* project we investigate the ways in which the cultural appeal of a book may be measured. Apart from answering the research questions above, this could also be useful for publishers who are confronted with too many books to choose from for translation into different markets. One of the challenges of our approach is that we want to compare data written in different languages. Our ultimate aim would be to find a way to estimate the different prominence of, for example, bicycles and cycling in different languages and countries as represented in Wikipedia entries in different languages and use the result in a tool that yields helpful visualizations for scholars as well as decision makers.

Warnske-Wang et al. [1] describe a way to compare different language Wikipedias based on their size, amount of shared entries, and amount of unique entries per language; they also pay attention to the flow of translations between the Wikipedias. This is a very useful approach to get an overview, but ours is focusing on a more detailed way to look at shared and unique knowledge. For this paper we have focused on proper names (named entities). Names of persons, places or organizations in one country are not always known in other countries. We assume a lack of contributions from a country to a Wikipedia entry of a certain name may indicate a gap in the knowledge of readers from that country. We assume that a book referring to that name could be more exotic and therefore more appealing for certain readers, while others therefore are less interested in the book. An indication of the 'level of exoticness' thus could help a reader/publisher to decide to read/translate the book or not.

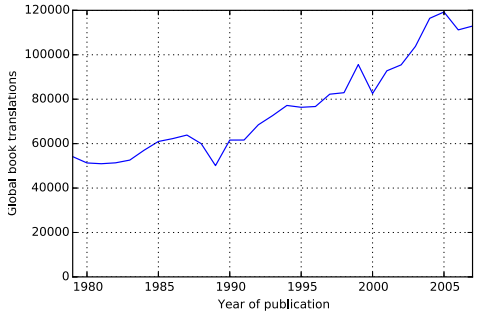


Fig. 1. Global number of books translated reported on Unesco's Index Translationum.

The project uses diverse technological means to assess the cultural content of a particular book such as natural language processing tools and linking to web resources like Wikipedia.

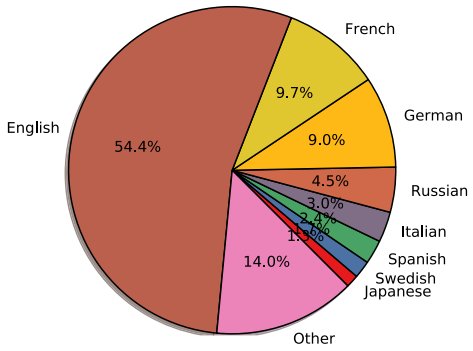
A. Book translations

Unesco's Index Translationum [2] provides an online resource for book translations worldwide. This rich source of

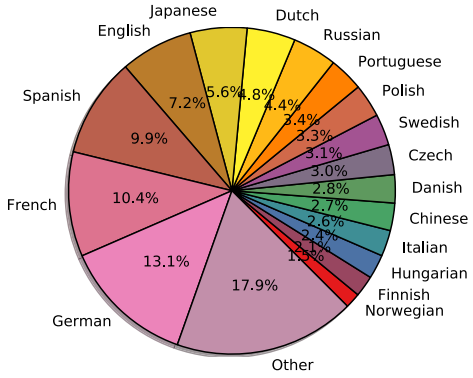
978-1-4673-9325-6/15 $31.00 © 2015 IEEE
DOI 10.1109/eScience.2015.12

IEEE computer society

information highlights the importance of understanding the reasons why a book gets translated from one language into another. The index also provides an insight into the trends of translations between languages: for example, it is possible to see that English is the largest exporting language, with 54.4% of translations having English as their source language; on the other hand, German is the largest importing language, with 13.1% of translations having German as their target language. Figure 2 illustrates the percentage of the global import and export markets which different languages represent.



(a) Translations from source language



(b) Translations into target language

Fig. 2. Percentage of translations by source and target language. Languages which represent less than 1% have been grouped as 'Other'.

Other useful resources which provide information on book translations are OCLC WorldCat,[1] LibraryThing,[2] and GoodReads.[3] These sources provide information on the various editions of a book, such as language and date of publication, but also provide more subjective data such as reader reviews and popularity scores. This type of information provides an additional source of information which could enable researchers to learn more about the book and its acceptance in various reader circles.

The rest of this paper is structured as follows: Section II details the approaches taken to assess international appeal of a book. Section III presents experimental results and discusses their interpretation. Final remarks and future work are presented in Section IV.

[1]URL: https://www.worldcat.org/
[2]URL: https://www.librarything.com/
[3]URL: http://www.goodreads.com/

II. Methodology

This section describes the methodology used to assess the cultural content of a particular book. The starting point of this work is the assumption that Named Entities are relevant to understanding the cultural context of a book [3]: the names of individuals (e.g. historical figures or local cultural icons) may be meaningful in one culture but may bear little relevance in another. This concept is further extended to incorporate other names such as locations, organisations and events. These words comprise a set of key words which will be used as a basis to analyse the cultural setting of a book.

In this work, we want to look at cultural proximity on a country-level, as individual countries are considered as a proxy for a particular culture. Although this assumption is not strictly speaking accurate, it provides a good enough generalization to allow to make statements of the type: "Dutch people are interested in bicycles". Additionally this assumption facilitates using existing information in sources like Wikipedia which already provides country information for contribution sources (see section II-D for further details). Several examples in section III use the Netherlands and Canada as specific countries for comparison, but the method introduced here can be applied to any country.

A. Overview

With the aim of analysing the cultural bias or proximity of a book in a particular country, various tools are used to construct a technical pipeline. The pipeline takes the full text of the book as its input and will produce as its output the required information to analyse the cultural proximity.

The first stage of this pipeline is the identification of key textual elements (key words or in this case named entities) in the text (see section II-B). Cultural proximity of each key element is extracted from Wikipedia (see sections II-D and II-E). Finally the cultural proximity of all key elements is aggregated in order to produce an overall assessment for the whole book (see section II-F).

B. Key word identification

As mentioned before, the first step for analysing a book requires the identification of key words in the text. Because we want to compare content from books with entries in Wikipedia, we opted for using Semanticizer[4] [4] which is a tool for semantic linking to relevant concepts in Wikipedia. Semanticizer takes blocks of text as its input and produces a list of possible Wikipedia concepts which are considered as relevant to the given text. Each concept is assigned a probability P_x of the concept x being relevant to the text. Depending on the application, a threshold of probabilities should be assigned to accept or discard concepts; higher thresholds cause less concepts to be accepted and conversely lower thresholds

[4]https://github.com/semanticize/semanticizest

cause more concepts to be accepted. Keywords identified via Semanticizer are not assigned to one of the four types of names mentioned above, although, as mentioned before, this is not an issue since all groups are treated equally.

The next step in the process of assessing cultural content of a book is linking the identified key words to a particular culture – or in our case to a particular country. In order to do this, we assume that each identified key word corresponds to a particular Wikipedia article. This is a reasonable assumption as many concepts which may be considered *important* will have a Wikipedia page written about them. Furthermore, key words identified through Semanticizer almost certainly have a Wikipedia article associated to them, as Semanticizer uses Wikipedia to identify concepts.

C. Available information in Wikipedia

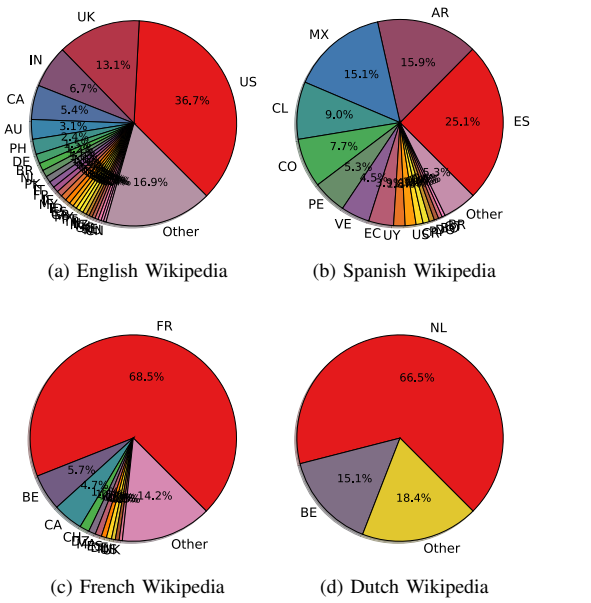
Before going into detail of how Wikipedia information was used to assess cultural proximity of a country to a particular Wikipedia article, it is worth reviewing what information is available on each article. This section explains what information is available; sections II-D and II-E explain how this information is used.

Wikipedia is a collaborative encyclopaedia, where each article is written collectively by contributors who have knowledge of, and are interested in the topic of the article. Articles are continuously edited and improved by contributors. Wikipedia is available in different languages, although there are no restrictions on origin of collaborations: users from any country can contribute to Wikipedia in any of the languages in which it is available.

Because English is such an influential language worldwide it is not surprising that the English version of Wikipedia is the largest one, representing a 48.6% share of global total volume of all combined Wikipedias, and with contributors from all over the world [5].

Although the English Wikipedia receives contributions from non-English speaking countries, it is not surprising that English speaking countries represent a larger proportion of contributions. Also, it could be expected that countries with larger populations will make a larger number of contributions on average. This bias in the distribution of contributions must be taken into account when any analysis is done on this data. Figure 3 shows the distribution of overall contributions made to various language Wikipedias by different countries [5]. As it can be seen, the major contributing country to the English Wikipedia is the US with 36.7% of the total contributions, followed by the UK with 13.1% of contributions. This implies that, for any randomly selected article, we could expect 36.7% of the contributions to come from users in the US; a higher/lower percentage of contributions would represent a deviation from the average – this fact will be further discussed in Section II-F.

Figure 3 also shows the percentage of contributions to the Spanish, French and Dutch language Wikipedias. These languages represent 9.8% (Spanish, the second largest Wikipedia), 3.9% (French) and 1.9% (Dutch) of the global



(a) English Wikipedia (b) Spanish Wikipedia

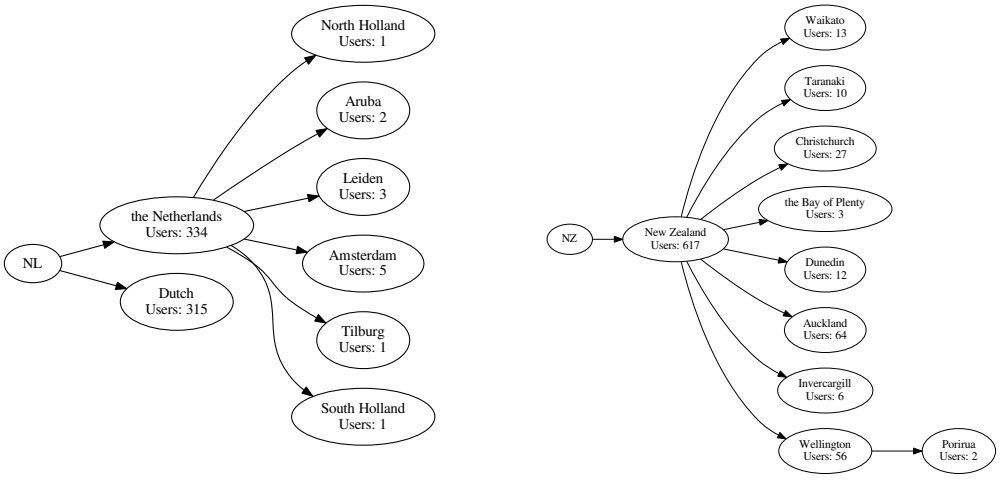
(c) French Wikipedia (d) Dutch Wikipedia

Fig. 3. Percentage of country contributions for various language Wikipedias.

volume of edits to Wikipedia. It can be observed that, although these Wikipedias still represent a significant amount of edits, contributions come mainly from countries where the Wikipedia's language is a native language, unlike the English Wikipedia which also has many contributions from countries where English is not the main spoken language. Perhaps this fact should not come as a surprise, since English can be considered as one of the most important languages for international communications [6]. Multiple language Wikipedias have previously been used to investigate cultural differences [7], although the comparison took place between Wikipedias in different languages rather than within a Wikipedia in one language.

Each page is composed of a series of revisions issued by individual contributors. A contributor can be either an anonymous contributor or a registered user; the majority of edits come from registered users [8]. For anonymous contributors, the only available information is the IP address of the contributor which can easily be resolved to a rough geographical area with sufficient precision to associate a contribution to a country. IP location is not unequivocally accurate: additional noise is introduced by users connecting via proxies, spoofing their address, etc; however, editing Wikipedia is not a task where such a practice is necessary (except perhaps on extremely controversial articles) and given the technical difficulty of identifying such users, we assume that they represent a percentage small enough not to represent an issue.

Contributions from registered users do not provide IP address information, which makes it difficult to associate contributions to a particular country. However it is possible to link a contributor to a country by examining the user's profile data. Wikipedia registered users can include category tags in



(a) Subcategories for The Netherlands

(b) Subcategories for New Zealand

Fig. 4. Category hierarchy provides geographical information about user: users which belong in one of these categories are associated with its location and by inference, within its *root* category.

their profile. Some of these categories provide specific geographical information: e.g. Category *Wikipedians in London* indicates the user lives in London, England. Categories can have subcategories (*Wikipedians in London* is a subcategory of *Wikipedians in England*, which is a subcategory of *Wikipedians in the United Kingdom*). The hierarchical organization of categories produces a category tree, where the root of the tree indicates a common geographical location applicable for the whole tree. The Wikipedia category structure may contain cycles, such that one category can be both an ancestor and descendant of another category. We avoid cycles by building a tree through subcategories and keeping a list of visited categories. Subcategories that have been visited earlier in the hierarchy are ignored.

The *Wikipedians by location* and *Wikipedians by ethnicity and nationality* categories provide a listing of all categories which can supply location information which can be used for our purpose. By associating specific categories to a country (e.g. *Wikipedians in the United Kingdom* to the UK), it is possible to associate all users in that category (or any of its subcategories) to the given country. Figure 4 shows two sample geographical category trees. These trees associate a particular country to their root category (The Netherlands and New Zealand); by extension, their subcategories also belong to these countries.

Notice that Figure 4a shows the root categories for the Netherlands from both types of categories: *Wikipedians from the Netherlands* (from *Wikipedians by location*) and *Dutch Wikipedians* (from *Wikipedians by ethnicity and nationality*).

The length of a Wikipedia page, the number of revisions in the page, and the length of each revision can also be used as indicators of the relative importance of the page: longer pages, with a larger number of revisions, may be indicative of a popular or important topic, while a smaller number of revisions may signal a less attractive topic.

Wikipedia topics which are available in multiple language Wikipedias often have links between these language versions. For example, the English Wikipedia page for *Paris*, has a link to the Dutch Wikipedia page for *Parijs*. Language links can be used to make links for the same concept between different language Wikipedias. These language links also indicate the inter-cultural relevance of a topic: topics which are more internationally appealing will be more likely to have been translated. Although it may not be immediately clear in which language the original page was written, revision timestamps and user comments on each revision may provide information regarding whether one page is a translation from the other.

Wikimedia Foundation provides an API for accessing Wikipedia data and meta-data [9]. This API allows users to query different types of information about specific pages such as a page's revision history. Several client implementations which make use of this API are available in various programming languages such as is the case of *mwclient*: a Python client for the MediaWiki API[5].

D. Cultural relevance per revision

Due to the collaborative nature of Wikipedia, it is reasonable to assume that the number of contributions to an article provides a measure of how popular/important a topic is considered to be by the community of contributors. Given the global base of contributors to the English Wikipedia, it is a useful source of information for our project. Thus, we used the English Wikipedia as a starting point for assessing cultural relevance.

[5]https://github.com/mwclient/mwclient

The Wikipedia page corresponding to each identified key word in a book is analysed to determine the interest of each country in this page. This is done through the information provided by the page's revision history.

The *Wikipedians by location* and *Wikipedians by ethnicity and nationality* can be used to aggregate users by country. Note that the country where Wikipedians are located is not necessarily the same as their country of origin. If their profile contains information about both countries, their edits count in the contributions of both countries. This reflects the ties a person can have to multiple countries or cultures. A *seed* category must be defined for each country of interest, and its subcategories can be automatically fetched using the Media Wiki API.

Unfortunately, not all registered users provide location information in their profiles. For these users it is not possible to determine their location and thus link them to a particular country. We take this into account through a confidence measure, which is the fraction of edits that can be associated with a country. This indicates the confidence that the distribution of edits over countries reflects the real distribution.

This means that there will always be a fraction of contributions on any given page which cannot be associated to a country. This should be kept in mind in any conclusions drawn from this data: a lack of data completeness implies a reduction in the confidence that can be placed in produced results; e.g. if only 70% of the contributions can be associated with a country, a 0.70 confidence should be associated with the results. We label this confidence factor as k.

Another consideration that must be taken into account is the existence of *Bots*: automated software components which periodically make edits to Wikipedia pages such as spelling corrections. Contributions from bots can represent a significant number of edits to a page, but since these edits are done by non-human contributors, they can be ignored for the purpose of this study without any impact on the confidence in produced results.

One final consideration which must be addressed is the inherent bias in Wikipedia edits: there is a clear gender bias on contributions to Wikipedia as a whole, as well as a technical barrier for new contributors to join [10], [11]; previous studies suggest that different language Wikipedias have different styles of contributing to the encyclopaedia [12] – it is reasonable to deduce that this also holds true for contributions from different countries to the English Wikipedia.

E. Cultural relevance per page

IP address and user profile information can be used to identify the country of origin of each contribution as described above. After identifying the country of origin of each contribution to a page, a list of countries contributing to that page can be easily produced. Such a list thus includes a percentage of *observed* contributions per country.

Figure 3 features a distribution of country contributions to the whole of Wikipedia which can be understood as an *expected* distribution of contributions. By comparing a page's *observed* distribution and the global *expected* distribution, it is possible to determine whether a specific country has made a larger (or smaller) percentage of contributions to the page than it does on average.

More formally, the *relative interest* of a country c in a particular topic is defined as the relation between the percentage a_c of *observed* (actual) contributions made by the country to that topic, and percentage e_c of *expected* contributions made by the country:

$$I(e_c, a_c) = \begin{cases} \dfrac{a_c}{e_c - 1}, & \text{if } a_c < e_c. \\ \\ \dfrac{1 - e_c}{a_c}, & \text{otherwise.} \end{cases} \quad (1)$$

The values for e_c are constants and are illustrated in Figure 3. The value a_c has its domain in the interval $[0, 1]$, where $a_c = 0$ indicates no contributions were made by c to the page and $a_c = 1$ indicate that all contributions in the page were made by c. The values of $I(e_c, a_c)$ range in the interval $[-1, 1]$, where negative numbers represent *lower than usual* and positive numbers represent *higher than usual* interest. $I(e_c, a_c) = 0$, if the contributions made by a country to the current page match the expected contributions from that country. The measure grows linearly between $[-1, 0)$ and asymptotic between $(0, 1]$.

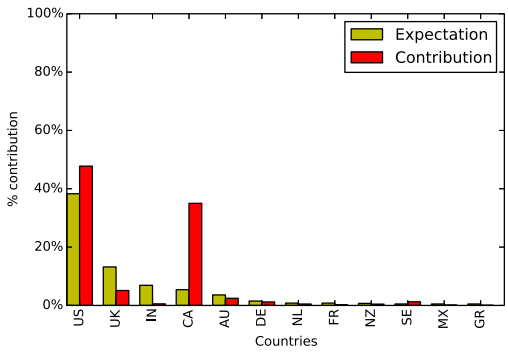
The relative interest is a consistent ratio between the number of contributions made by c and the expected number of contributions, regardless of the value of e_c. This mean that when the number of contributions made by c are $1/2$ the number of contributions c normally makes, then $I(e_c, a_c) = -0.50$. Likewise when the number of contributions made by c are twice the number of contributions c normally makes, then $I(e_c, a_c) = 0.50$.

For example, given two countries (e.g. the United States and France), with expected contributions $e_{US} = 0.40$[6] and $e_{FR} = 0.01$. Assume for a hypothetical page that these countries contribute $a_{US} = 0.60$ and $a_{FR} = 0.015$ (i.e, both countries contribution is 1.5 times their usual volume of contributions), then the relative interest for both countries would be the same: $I(e_{US}, a_{US}) = I(e_{FR}, a_{FR}) = 0.33$.

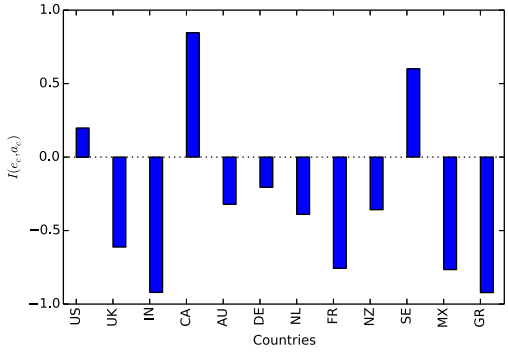
Although $I(e_c, a_c)$ is "fair" in giving the same score to both countries on the example above, it has an intrinsic limiting factor: countries which have a high contribution average (such as the US in the example above) are limited in the amount of relative interest they can express. This is due to the fact that logically no country can account for more than 100% of the contributions to a given page. Thus, it is impossible for the country with $e_c = 0.50$ to contribute more than twice its usual number of contributions. In other words, $I(e_c, a_c)$ has an upper bound $\sup I(e_c, a_c) = 1 - e_c$.

For example, given two countries (e.g. Canada and the Netherlands) with expected contributions $e_{CA} = 0.054$ and $e_{NL} = 0.008$. On the Wikipedia page for the topic *Ice*

[6]The value of e_{US} has been rounded here for the sake of the example.



(a) Expected and observed contributions



(b) $I(e_c, a_c)$ scores

Fig. 5. Contributions and $I(e_c, a_c)$ scores for the Wikipedia page on *Ice hockey* for selected countries.

hockey , these countries contribute $a_{CA} = 0.350$ and $a_{NL} = 0.005$. Thus they have scores of $I(e_{CA}, a_{CA}) = 0.846$ and $I(e_{NL}, a_{NL}) = -0.389$, which can be interpreted as Canada having a strong cultural connection to Ice hockey while the Netherlands has not. Figure 5 shows the comparison of expected and observed contributions and the $I(e_c, a_c)$ scores obtained by several countries for the page on *Ice hockey*. It can be observed that even despite the limitations mentioned above—e.g. the US cannot score as high in relative interest as other countries—for all countries this measure can express a countries connection to a topic: both Sweden ($e_{SE} = 0.005$) and the United States ($e_{US} = 0.383$) have a positive relative interest in Ice hockey.

F. Cultural relevance comparison

Section II-E explains how an individual word or concept can be given a cultural or country relevance score, that is, *relative interest*. It is straight forward to apply this method to a full text: key words must be extracted from the text and cultural relevance calculated for each of these words; afterwards the cultural relevance can be integrated into a single measure for the complete text.

As mentioned before, the confidence score k must be taken into account – $k(w)$ can be considered a weighting factor on the cultural relevance of word w. Additionally, one more weighting factor must be taken into account: the word frequency within the text. A word w appearing multiple times in the text may be indicative of the word being important in the text, thus the relative word frequency $f(w)$ must also be taken into account:

$$f(w) = \frac{count(w)}{\sum_{w \in T} count(w)} \tag{2}$$

Finally the cultural relevance of a given text T for a country c can be calculated as:

$$I_c(T) = \sum_{w \in T} S_c(w) \tag{3}$$

where $S_c(w)$ is the impact score of each word w:

$$S_c(w) = f(w) * k(w) * I(e_c, a_c) \tag{4}$$

Notice that, although $I_c(T)$ gives a total impact of a full text, keeping track of the impact score of each word might provide useful information regarding which words are more influential in the book's total score. For instance, two books may have a $I_c(T)$ score close to zero for different reasons: one may contain only words which are not interesting to country c, while the other may contain a mixture of many words with very high and very low $S_c(w)$ scores. When assessing the cultural content of a book for a particular country, looking at a single figure such as $I_c(T)$ on its own may be misleading; looking at the distribution of $S_c(w)$ scores might provide a better overview of the book's probable appeal to a given country. Section III provides examples which may help the reader gain a better understanding of these measures.

III. Results & Discussion

This section showcases experiments which illustrate how the proposed measure $I_c(T)$ can provide insight into the cultural relevance of a text in multiple languages. The measure is calculated for the Netherlands ($c = NL$) as the original text is in Dutch, and Canada ($c = CA$) as target country for translation into English.

In these experiments, a selected set of books (three novels and one non-fiction title) originally written in Dutch are analysed: *De aanslag* by Harry Mulisch (first published in 1982), *De bloedkoralen van de bastaard* by Rinus Ferdinandusse (first published in 1972) *Boven is het Stil* by Gerbrand Bakker (first published in 2006), and the non-fiction title *Romantische reizen: omzwervingen in de negentiende eeuw* by Peter van Zonneveld (first published in 1991). Note that at the time of writing this paper, *De aanslag* and *Boven is het Stil* had already been translated into English while *De bloedkoralen van de bastaard* and *Romantische reizen: omzwervingen in de negentiende eeuw* had not.

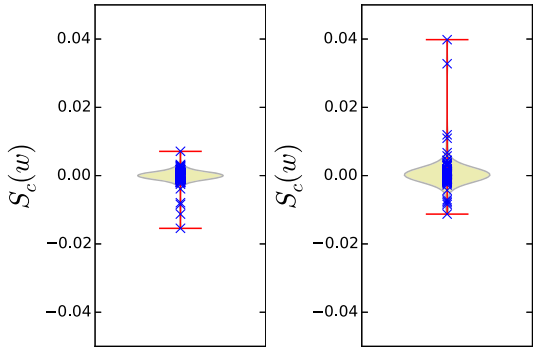
The importance of names has previously been established in a pilot project [3] that lead to the project Namescape [13]. We build on top of this work and use the Named Entities (NEs) identified in that project in our analysis. These were extracted from a collection of 564 books originally written

in Dutch between 1933 and 2008, some of which have been translated into other languages. Where possible, these NEs were linked to Wikipedia articles on those entities [13]. The NEs were classified in four groups: person, location, misc and organisation. However, in this work, entities from all four classes were treated equally. In [14] we present an analysis of the impact these classes have on cultural relevance.

To get an idea of whether concepts, like NEs, mentioned in a novel are potential cultural hurdles to understanding, e.g. that would need an explanation or extra description in a translation for a different market, the impact scores of all the identified key elements can be plotted, for instance in a distribution plot or a scatter plot. This can help translators and publishers identify how many elements are potential hurdles for each potential country, and what these elements are.

A. $S_c(w)$ *distribution plots*

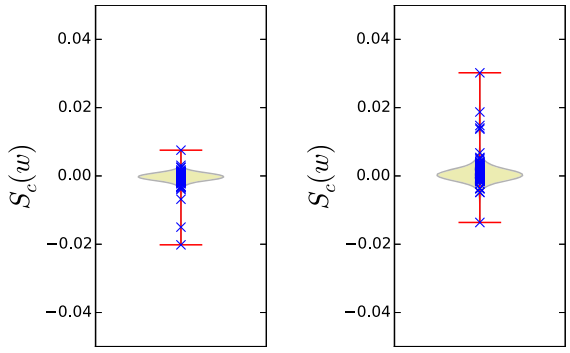
Figures 6 to 8 show the $S_c(w)$ distribution for countries $c = CA$ and $c = NL$ over a novel. Each point in this plot represents a single NE and is positioned according to its $S_c(w)$ score. The violin plot indicates the relative density of NEs along the $S_c(w)$ axis.



(a) $S_{CA}(w)$ distribution plot (b) $S_{NL}(w)$ distribution plot

Fig. 6. Comparative $S_c(w)$ distributions for *De aanslag*.

Figure 6 shows the $S_c(w)$ distributions for *De aanslag*. This novel has $I_c(T)$ scores of $I_{CA}(T) = -0.0306$ and $I_{NL}(T) = 0.1054$, however the mean $S_c(w)$ for both countries is quite close: $\overline{S_{CA}(w)} = -0.0002$ and $\overline{S_{NL}(w)} = 0.0007$, which can be visually confirmed by the fact that both distributions are centered close to zero. The high concentration of scores close to zero indicates that most NEs mentioned in this novel get close to their expected contributions from Canada and the Netherlands. However, the distribution in (a) is relatively narrow while the distribution in (b) has a longer positive tail, reaching a maximum $S_{NL}(w) = 0.0398$ for w corresponding to the Dutch city of *Haarlem*. Interestingly, the same NE has a $S_{CA}(w) = -0.0154$, which is the minimum score in this plot. Similarly, the second and third highest $S_{NL}(w)$ scores correspond to *Amsterdam* and *Netherlands*, with scores of $S_{NL}(Amsterdam) = 0.0328$ Vs. $S_{CA}(Amsterdam) = 0.0028$ and $S_{NL}(Netherlands) = 0.0120$ Vs. $S_{CA}(Netherlands) = 0.0012$ respectively. This indicates that the novel contains more NEs for which the contributions from Dutch Wikipedians is higher than expected, e.g. their Wikipedia articles have a bias of contributions from the Netherlands.



(a) $S_{CA}(w)$ distribution plot (b) $S_{NL}(w)$ distribution plot

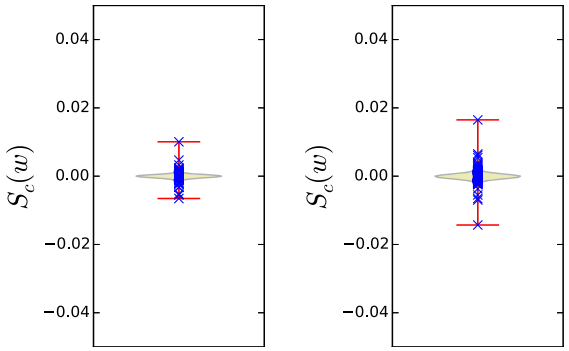
Fig. 7. Comparative $S_c(w)$ distributions for *De bloedkoralen van de bastaard*.

Figure 7 shows the $S_c(w)$ distributions for *De bloedkoralen van de bastaard*. This novel has $I_c(T)$ scores of $I_{CA}(T) = -0.0221$ and $I_{NL}(T) = 0.2254$, but again the mean $S_c(w)$ for both countries is quite close: $\overline{S_{CA}(w)} = -0.0005$ and $\overline{S_{NL}(w)} = 0.0008$. However while (a) is skewed towards negative scores, (b) is skewed towards positive scores. From the top 3 $S_{NL}(w)$ ranked NEs ($S_{NL}(Amsterdam) = 0.0302$, $S_{NL}(Middelburg) = 0.0187$[7] and $S_{NL}(FrankvanBorssele) = 0.0148$[8]), only *Amsterdam* has a relatively high $S_{CA}(Amsterdam) = 0.0026$ score, being ranked third, while the other two ($S_{CA}(Middelburg) = -0.0202$ and $S_{CA}(FrankvanBorssele) = -0.0150$) are ranked last and second to last. This seems to confirm the hypothesis that NEs which are important to one country (Dutch locations and historical figures) seem to bear less significance in other countries.

Figure 8 shows the $S_c(w)$ distributions for *Romantische reizen: omzwervingen in de negentiende eeuw*. This book has $I_c(T)$ scores of $I_{CA}(T) = -0.0087$ and $I_{NL}(T) = -0.0095$, and mean $S_c(w)$ scores of $\overline{S_{CA}(w)} = 0.0000$ and $\overline{S_{NL}(w)} = 0.0000$. In this case, both countries are equally neutral on their $I_c(T)$ score for the book. However the distribution of $S_{NL}(w)$ spans a wider range than $S_{CA}(w)$, indicating both stronger interest on some NEs while simultaneously expressing stronger disinterest on others. For example *Louis Bonaparte*[9] is the highest ranked NE for both countries, but $S_{NL}(LouisBonaparte) = 0.0165$ represents a higher interest on this NE than $S_{CA}(LouisBonaparte) = 0.0101$.

Interestingly, some NEs with high $S_{CA}(w)$ score have low $S_{NL}(w)$ and the other way around. For example,

[7]Middelburg is a municipality and city in the south-west of the Netherlands.
[8]Frank van Borssele was a Dutch nobleman of the 15th century.
[9]Louis Bonaparte was King of Holland and brother of French Emperor Napoleon I



(a) $S_{CA}(w)$ distribution plot (b) $S_{NL}(w)$ distribution plot

Fig. 8. Comparative $S_c(w)$ distributions for *Romantische reizen: omzwervingen in de negentiende eeuw*.

$S_{CA}(LordByron) = 0.0047$[10] Vs. $S_{NL}(LordByron) = -0.0143$ and $S_{CA}(Delacroix) = 0.0033$[11] Vs. $S_{NL}(Delacroix) = -0.0055$. Notice that Lord Byron and Delacroix are English and French artists and given these are the two main spoken languages in Canada, it may be understandable that these NEs are of higher importance to Canadian Wikipedians. Similarly, $S_{CA}(Leiden) = -0.0052$[12] Vs. $S_{NL}(Leiden) = 0.0061$ and $S_{CA}(PhilippFranzvonSiebold) = -0.0055$[13] Vs. $S_{NL}(PhilippFranzvonSiebold) = 0.0065$. It is reasonable that Dutch Wikipedians have more to contribute about Leiden, being a Dutch city and Philipp von Siebold, given the geographic and cultural proximity of Germany to the Netherlands.

As mentioned at the end of section II-F, this example illustrates how a single number may be insufficient to grasp the cultural proximity of a book to a country. In [14] we prepared a ground-truth data set and applied $I_{NL}(T)$ as a single number. A low correlation with translation from Dutch to English was observed, which implies one of the following causes:

- the approach of calculating a single $I_{NL}(T)$ score is not appropriate, or
- the Dutch Wikipedia contribution calculated by $I_{NL}(T)$ is a noisy signal and needs to be qualified to be used. I.e. the scatter plots are more effective for publishers than a single number, because translation is a highly complex process with decisions based on many other aspects than country-specific interest.

What the distribution plots in Figures 6 to 8 indicate is that for many of the entities mentioned in novels there is no strong country bias in the contributions on their respective Wikipedia articles, suggesting they offer no hurdle to understanding the setting of a novel in different cultures.

[10]Lord Byron was an English poet of the 19th century.
[11]Eugene Delacroix was a french romantic artist of the 19th century.
[12]Leiden is a city in the province of South Holland in the Netherlands.
[13]Philipp Franz von Siebold was a German physician of the 19th century.

B. $S_c(w)$ scatter plots

An alternative way of comparing the cultural relevance of a novel between two countries is via a simple scatter plot of $S_c(w)$ scores. Figures 9 to 11 show scatter plots for the same books discussed above.

One advantage of visualizing cultural relevance on a scatter plot is that this allows to identify the amount of agreement (or disagreement) in country bias between two countries: NEs with a bias towards both countries will show on the first quadrant of the plot; NEs in which neither country expresses an interest (a negative bias) will show on the third quadrant; the second and fourth quadrant will contain NEs for which Wikipedia contributions show a bias for one country and against the other. The distance from the origin provides an indication of the strength of the bias (either agreement or disagreement). NEs which are located at or close to the origin (e.g. close to both countries' expected contributions) would indicate neutrality on both parties while NEs which far away from the origin would indicate a strong agreement/disagreement.

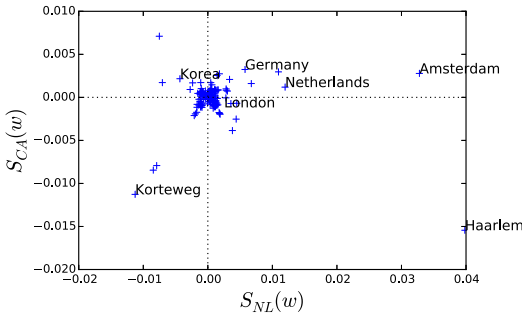


Fig. 9. Scatter plot of $S_c(w)$ for *De aanslag*.

Figure 9 shows the $S_c(w)$ scatter plot for *De aanslag*. Prominent NEs in the first quadrant include Amsterdam, Netherlands, and Germany, while NEs in the second quadrant such as Korea indicate a slight bias towards Canada; Haarlem is quite a distance away from the origin, deep into the fourth quadrant, indicating a strong bias towards Dutch contributions, therefore a high level of disagreement on this NE.

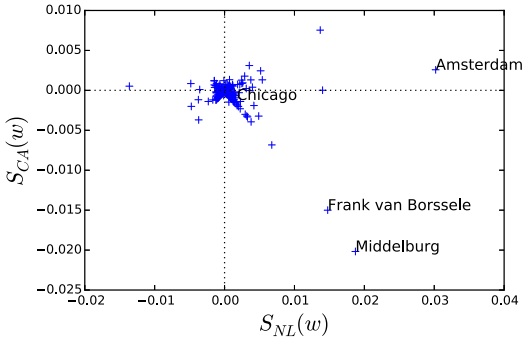


Fig. 10. Scatter plot of $S_c(w)$ for *De bloedkoralen van de bastaard*.

Figure 10 shows the $S_c(w)$ scatter plot for *De bloedkoralen van de bastaard*. Again, while both countries seem to agree in the relevance of Amsterdam, they disagree for other NEs such as Middelburg, and are equally neutral about others like Chicago.

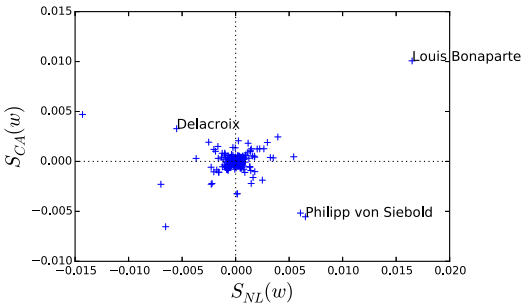


Fig. 11. Scatter plot of $S_c(w)$ for *Romantische reizen: omzwervingen in de negentiende eeuw*.

Figure 11 shows the $S_c(w)$ scatter plot for *Romantische reizen: omzwervingen in de negentiende eeuw*. Again, it is evident there is disagreement in the relevance of NEs like Lord Byron, Delacroix, and Philipp von Siebold, while there is agreement in NEs like Louis Bonaparte and Mary Shelley[14].

The scatter plots can be made between any two countries. Translators and publishers can analyse a novel considered for translation and compare the cultural relevance or bias for the country in which the novel was originally published with that of a target country.

C. Tools for book publishers

In this section we propose how the measures introduced in this paper can be of use for book publishers when deciding whether a given book is suitable for translation.

Scatter plots on Figures 9 to 11 show the $S_c(w)$ measure, which takes into account the word frequency ($f(w)$) and confidence ($k(w)$). In order to have more insight into which words have a higher cultural proximity to a country, Figure 12 shows a plot of $I(e_c, a_c)$. In this Figure as before, words fall in for quadrants: high $I(e_{NL}, a_{NL})$ and $I(e_{CA}, a_{CA})$ in the first quadrant and so on.

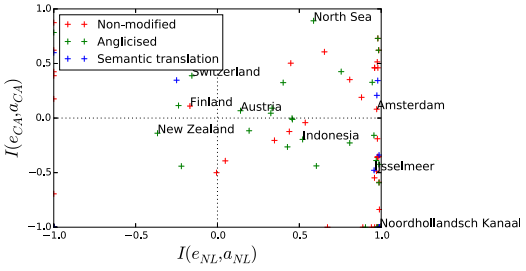


Fig. 12. Scatter plot of $I(e_c, a_c)$ for *Boven Is Het Stil*.

[14]Mary Shelly was an English writer from the beginning of the 19th century.

Figure 12 shows the distribution of $I(e_c, a_c)$ of entities in the novel *Boven is het stil*, for the NL and CA. Notice that well known entities such as *North Sea* have a high $I(e_c, a_c)$ measure for both countries, while other entities which could be understood as *"typically Dutch"*, such as *IJsselmeer* and *Noordhollandsch Kanaal*, have a high score for the NL, but a low score for CA.

Boven is het stil has been translated to English. One point to be noted about this translation is that some named entities have been altered during the translation process. Some of these entities have been replaced by the English name of places (e.g. *Nieuw-Zeeland* $\rightarrow$ *New Zealand*), while others have been semantically translated (e.g. *Noord-Hollands Kannal* $\rightarrow$ *North Holland Canal*), and yet others have been completely replaced (e.g. *Afsluitdijk* $\rightarrow$ *Lake IJssel dam*). A small number of entities have been completely removed from the translated text (*Kanis en Gunnink*, *Van Gend en Loos*), probably because the translator considered them to be very specific to Dutch culture and because they could be removed without affecting the story.

Table I shows the count of entities on each quadrant which have been translated or not. This type of analysis can provide a book translator with an indication of which named entities may require special attention during the translation process, depending on the quadrant where the entity is located.

IV. Conclusion

This paper has introduced an automated method for measuring the potential (absence of) cultural interest for a novel in different countries. With this measure, publishers and translators can identify key elements in a novel that may be difficult to understand in countries considered as target markets. This method relies on public domain data and open source tools and thus is available to anyone interested in translations. And although the method introduced here is aimed for translation of books, it can also be used for translation of other textual sources such as song lyrics, theatre play or movie scripts, etc.

Identifying textual features (such as we have done with named entities) to analyse a novel is a common practice in literary studies. Linking novels to Wikipedia is not a new concept, but using Wikipedia as a proxy for gauging the cultural significance of specific terms is an innovative technique which, given its limitations, provides an interesting way to analyse the cultural resonance with a novel in different regions.

Whether (cultural) familiarity and (global) appeal converge or not may vary depending on the book market, the genre (commercial vs literary) and the combination of the source and target cultures. (Dutch may be into the exoticism of Japan but not the exoticism of Albania, for example). Our tool offers a quantitative way of looking at questions on the globalisation of literature and the (shifting, decreasing?) *otherness* of literature from different cultures. We wonder if in the future of publishing, besides using the two favourite tools of the trade (gut feeling and sales figures [15]) there will also grow a kind of *data-driven market research* using tools such as the one we are developing. Publishers could also

TABLE I
ENTITY COUNT ON EACH QUADRANT ON FIGURE 12

$I(e_c, a_c)$	Non-modified	Anglisiced	Semantically translated	Left out
High NL, High CA	11	9	2	1
High NL, Low CA	20	15	5	2
Low NL, High CA	6	3	2	0
Low NL, Low CA	15	6	4	0
Total	52	33	13	3

start using these types of cultural interest measures to enrich their e-books with helpful annotations of difficult/unfamiliar concepts.

A. Future work

The next step will be a more thorough validation of $S_c(w)$ scores. For example, the *highly Dutch* cultural significance of the Dutch cites like *Haarlem*, *Middelburg* and to a lesser extent *Amsterdam* can be seen as *matching with common sense expectations*. What is needed is a more objective and quantifiable way to test the usefulness of the measure.

Other information sources (OCLC WorldCat, LibraryThing, and Goodreads) provide additional book meta data which can be incorporated in the analysis to enhance the understanding of the decision process for translating a book. For instance, these sources can be mined to produce a list of books which have been translated from one language to another and of books which have not yet been selected for translation. The $I_c(T)$ could be applied to the text of these books and the similarities between translated/non-translated books can be analysed.

Additionally, these sources can provide information regarding the typical translation paths for a book. For example, whether Dutch books typically get translated to German first and then to other languages, etc. These research questions will be the topic of future research.

REFERENCES

[1] M. Warncke-Wang, A. Uduwage, Z. Dong, and J. Riedl, "In search of the ur-wikipedia: Universality, similarity, and translation in the wikipedia inter-language link network," in *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, ser. WikiSym '12. New York, NY, USA: ACM, 2012, pp. 20:1–20:10. [Online]. Available: http://doi.acm.org/10.1145/2462932.2462959

[2] UNESCO. Index translationum - world bibliography of translation. [Online]. Available: http://www.unesco.org/xtrans/

[3] K. van Dalen-Oskam, "Names in novels: an experiment in computational stylistics," *Literary and Linguistic Computing*, 2012. [Online]. Available: http://llc.oxfordjournals.org/content/early/2012/03/09/llc.fqs007.abstract

[4] D. Odijk, E. Meij, and M. de Rijke, "Feeding the second screen: Semantic linking based on subtitles," in *Open research Areas in Information Retrieval (OAIR 2013)*, Lisbon, Portugal, 05/2013 2013.

[5] E. Zachte. (2014) Wikimedia statistics. [Online]. Available: http://stats.wikimedia.org

[6] I. N. Khokhlova, "Cross-cultural communication: Euro-english," *Life Science Journal*, vol. 11, no. 12s, 2014. [Online]. Available: http://www.lifesciencesite.com/lsj/life1112s/169_26493life1112s14_784_786.pdf

[7] Y. Eom and D. Shepelyansky, "Highlighting entanglement of cultures via ranking of multilingual wikipedia articles," *PLoS ONE*, vol. 10, 2013. [Online]. Available: http://arxiv.org/abs/1306.6259

[8] K. Panciera, A. Halfaker, and L. Terveen, "Wikipedians are born, not made: A study of power editors on wikipedia," in *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, ser. GROUP '09. New York, NY, USA: ACM, 2009, pp. 51–60. [Online]. Available: http://doi.acm.org/10.1145/1531674.1531682

[9] MediaWiki, "Mediawiki, the free wiki engine," 2014, [Online; accessed 2-February-2015]. [Online]. Available: http://www.mediawiki.org/w/index.php?title=API:Main_page

[10] J. Antin, R. Yee, C. Cheshire, and O. Nov, "Gender differences in wikipedia editing," in *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, ser. WikiSym '11. New York, NY, USA: ACM, 2011, pp. 11–14. [Online]. Available: http://doi.acm.org/10.1145/2038558.2038561

[11] P. Vora, N. Komura, and S. U. Team, "The n00b wikipedia editing experience," in *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, ser. WikiSym '10. New York, NY, USA: ACM, 2010, pp. 36:1–36:3. [Online]. Available: http://doi.acm.org/10.1145/1832772.1841393

[12] U. Pfeil, P. Zaphiris, and C. S. Ang, "Cultural differences in collaborative authoring of wikipedia," *Journal of Computer-Mediated Communication*, vol. 12, no. 1, 2006. [Online]. Available: http://jcmc.indiana.edu/vol12/issue1/pfeil.html

[13] K. van Dalen-Oskam, J. De Does, M. Marx, I. Sijaranamual, K. Depuydt, B. Verheij, and V. Geirnaert, "Named entity recognition and resolution for literary studies," *Computational Linguistics in the Netherlands Journal*, vol. 4, pp. 121–136, 12/2014 2014.

[14] C. Martinez-Ortiz, F. Buschenhenke, K. van Dalen-Oskam, and M. Koolen, "Predicting the international appeal of novels," in *abstract accepted for DH2015, Sydney Australia (forthcoming).*, 2015.

[15] T. Franssen and G. Kuipers, "Coping with uncertainty, abundance and strife: Decision-making processes of dutch acquisition editors in the global market for translations," *Poetics*, vol. 41, no. 1, pp. 48 – 74, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0304422X12000691