**Martinez-Ortiz — Predicting**

<7 figures and equations>

**Predicting the International Appeal of Novels**

Martinez-Ortiz, C., Buschenhenke, F., Van Dalen-Oskam, K. and Koolen, M.

Globalization is a current theme in literary studies. Are authors writing increasingly for a global audience, with novels that appeal to readers in many countries, through shared ‘global’ cultural knowledge—e.g., references to concepts with which people around the globe are familiar? The Beyond the Book project aims to investigate whether we can measure how ‘international’ a novel is in its references to cultural knowledge. This could be useful in predicting the international appeal of novels and help publishers choose titles for translation that may appeal to readers in different markets.

Although many potential factors contribute to a publisher’s or editor’s decision to translate a novel, we wish to know if textual aspects are a factor, and if so, to what extent. As a first step, we focus on the named entities in a set of Dutch novels, including novels that have been translated from Dutch to English.

We link the named entities to English Wikipedia articles and use the Wikipedia edit history to measure the relative contribution of Dutch Wikipedians to each entity as a proxy for how specific the cultural references are to Dutch readers.

The main research questions addressed are

• How can we use named entities in novels as representations of cultural references?

• How can we measure the relation between cultural references and international appeal?

**The Translation Market**

*The World Book Market*

From UNESCO’s Index Translationum database, which charts worldwide book translations, Heilbron (2010) finds a four-level structure: around 55 to 60% of all translations are from English, around 10% from German and French, 13% for each of seven or eight other languages, and the rest are translations from ‘peripheral’ languages (including Dutch). For peripheral languages, translations from other languages are common, whereas in English-speaking countries, translations make up only a small fraction (around 3%) of the book market.

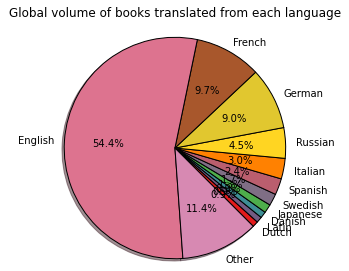


Figure 1.

*The Dutch Book Market*

Import

Across all genres, in the Netherlands only about 34% of translated books come from languages other than English (UNESCO Index, up to 2006). This percentage fluctuates over time: the report ‘Publishing Translations in Europe: Trends 1990­–2005’ by the Mercator Institute, which provides data for the Dutch market for translated fiction, states that 72.6% of all translations are from English.1

Export

The Dutch export market is very different from the import market. For spreading Dutch novels, Germany is the ‘gate market’ that opens up the possibility of translation into other peripheral markets in Europe (Heilbron, 2010). From the roughly 200 translations receiving funding from the Letterenfonds per year, 60 are adult fiction, and German takes up about 30% of those; English, only 17%.

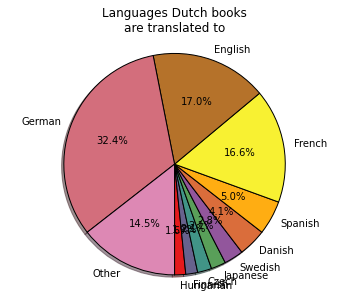


Figure 2.

*The Decision Makers*

Franssen and Kuipers (2011) asked 23 Dutch editors how they decide which foreign titles to purchase for the domestic market. When editors assessed the text, they were purely focussed on ‘universal’ manuscript quality, not international appeal. We found in our interviews with 12 Dutch decision makers that, when looking to sell international translation rights for Dutch novels, a translation into English is deemed the most desirable of all. Hence, we focus on translations to English.

**Methodology**

*Technical Pipeline*

This section describes the necessary steps to go from the full text of a novel to a meaningful measure of relative interest from a country in that novel.

Identify Words of Interest

The full text of a novel is split into arbitrarily sized units (e.g., a paragraph), which are processed using Semanticizer (Meij et al., 2012), to produce a list of Named Entities (NEs). These NEs represent the terms (*wi*) that have a high probability of being used as links to other Wikipedia pages, if the given text was an article in Wikipedia. Each of these link-terms has a probability of being a link associated with it *Plink*(*wi*); link-terms are filtered by their probability, and terms with probability below a certain threshold are discarded. The frequency of an NE in a given text *F*(*wi*) represents the relative importance of that NE.

Country Interest

Next, the relative interest from a country in each link-term is determined. This is estimated on the basis of the number of edits from a given country to the Wikipedia page corresponding to the NE. The English version of Wikipedia is used for this study, since it contains edits from many countries, including the Netherlands. Each Wikipedia article is built from contributions of several users over time. The interest in a particular topic is gauged by the number of contributors from a specific country. It is possible to determine the country of origin of the contributor, either by IP address (recorded for anonymous contributors) or by analysing the geographical categories the contributor belongs to (e.g., Wikipedians in the Netherlands).

This proposed methodology has limitations. Contributions from ‘bots’ (i.e., contributions made automatically by programs) need to be ignored. Registered users can state their nationality on their user profile, but not all users provide such information. The fraction of unknown contributions is treated as an uncertainty *C*(*wi*). For instance, if 70% of the contributions to a given page are from known sources, then the interest from a given country can be assessed with a 0.7 confidence.

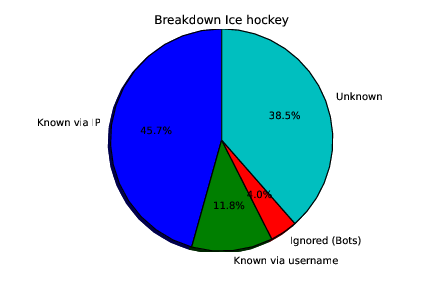


Figure 3.

Each country provides a different volume of contributions to Wikipedia. Thus, country contributions on a specific topic must be normalized by that country’s contributions to the whole English Wikipedia. The interest of a given country on a particular NE *IQ (wi)* is then the normalized percentage of contributions from that country to the Wikipedia article for the NE in question.

To illustrate these steps, the following graphs show the percentages of contributions from various countries to the English Wikipedia,2 the expected and observed contribution percentages to the English Wikipedia article on ‘ice hockey’, and the relative country contributions to the ‘ice hockey’article.

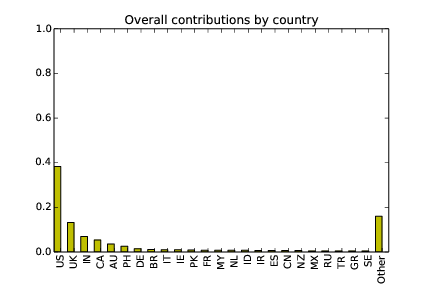


Figure 4.

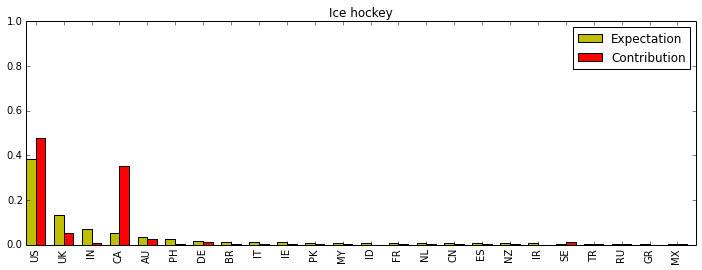


Figure 5.

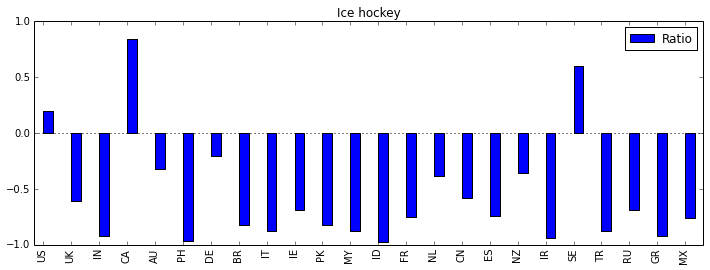
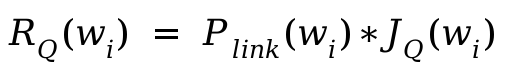
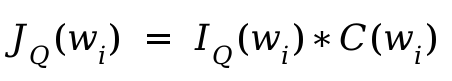


Figure 6.

Integrate over Complete Novel

The relative importance *R* of a given NE *wi* for country *Q* is calculated as follows:

where *JQ*(*wi*) is the weighted interest of country *Q* on the word:

And the interest from that country for a whole novel can be calculated as the average interest of the identified NEs in the novel:

Where *N* is the total number of NEs in the novel.

**Experiments**

For the experiments, 492 Dutch (mostly literary) novels are used, all published between 1933 and 2008. Of these, 318 have been translated into other languages, but only 27 have been translated into English (based on data from WorldCat).

The following features are determined:

• The **number of translations** to English.

• The **number of entities** in four categories: person, location, organization, and miscellaneous.

• The relative **Dutch contribution** to Wikipedia articles on entities in each category.

The correlations between the number of English translations and the two features per entity category are shown in the figure below:

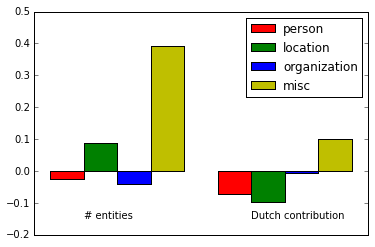


Figure 7.

There is a weak correlation (0.38) between a Dutch novel getting translated into English and its number of miscellaneous entities (events, buildings, etc.). These can be interpreted as cultural references that readers from different countries may be familiar with. The number of locations has a positive but very weak correlation with getting translated, while the number of person, location, and organisation names has a very small negative correlation.

The number of translations to English is weakly correlated with the relative contribution from Dutch Wikipedians to the entities in the miscellaneous category. These results suggest that novels that get translated from Dutch to English contain many references to Dutch culture. The locations in translated novels have a relatively low Dutch contribution, which suggests that their settings are perhaps more international. It is intriguing that the international appeal of these novels may lay just as much in their ‘Dutchness’ as in their ‘global’ references.

**Conclusions**

The aim of this paper is to investigate the role the text of a novel plays in determining whether it gets translated.

Cultural references seem to contribute to a novel’s chances of getting translated. The classic writers’ adagium, ‘Show, don’t tell’, may actually work; cultural references provide concrete, familiar details that anchor the story to the ‘real world’. Their realism could lead to a greater appeal for readers.

These are preliminary findings. The current set of books is very small, and the methods are crude. The next steps will be to refine our methods by incorporating more textual features, look in more detail at how they relate to translation decisions, and apply our methods to a larger set of novels, in different genres and languages.

**Notes**

1. There are gaps in the available data of 1993–1999, so the picture is not entirely accurate.

2. See http://stats.wikimedia.org/wikimedia/squids/SquidReportPageEditsPerLanguageBreakdown.htm.

**References**

**Franssen, T. and Kuipers, G.** (2011). Overvloed en onbehagen in de mondiale markt voor vertalingen. *Sociologie,* **7(**1) (2011): 67–93.

**Heilbron, J.** Structure and Dynamics of the World System of Translation. *UNESCO International Symposium on Translation and Cultural Mediation*, Paris, 22–23 February 2010.

**Meij, E., Weerkamp, W. and Rijke, M. de.** (2012). Adding Semantics to Microblog Posts. *WSDM 2012: Fifth ACM International Conference on Web Search and Data Mining*, Seattle, WA, 8–12 February 2012, pp, 563–72, http://ilps.science.uva.nl/biblio/adding-semantics-microblog-posts.

**Publishing Translations in Europe: Trends 1990–2005.** (2010). Making Literature Travel Series, Literature across Frontiers. Mercator Institute for Media, Languages, and Culture, Aberystwyth University, Wales.

**Van Dalen-Oskam, K. Does, J. de., Marx, M., Sijaranamual, I., Depuydt, K., Verheij, B. and Geirnaert, V.** (2014). Named Entity Recognition and Resolution for Literary Studies. *Computational Linguistics in the Netherlands Journal,* **24**: 473–80.

**Vertalingendatabase, Letterenfonds** [Dutch Institute for the Promotion of Dutch Literature Abroad]. (2014). https://letterenfonds.secure.force.com/vertalingendatabase/zoeken.