Directions for Exploiting Asymmetries in Multilingual Wikipedia

Elena Filatova
Computer and Information
Sciences Department
Fordham University
Bronx, NY 10458, USA
filatova@cis.fordham.edu

Abstract

Multilingual Wikipedia has been used extensively for a variety Natural Language Processing (NLP) tasks. Many Wikipedia entries (people, locations, events, etc.) have descriptions in several languages. These descriptions, however, are not identical. On the contrary, descriptions in different languages created for the same Wikipedia entry can vary greatly in terms of description length and information choice. Keeping these peculiarities in mind is necessary while using multilingual Wikipedia as a corpus for training and testing NLP applications. In this paper we present preliminary results on quantifying Wikipedia multilinguality. Our results support the observation about the substantial variation in descriptions of Wikipedia entries created in different languages. However, we believe that asymmetries in multilingual Wikipedia do not make Wikipedia an undesirable corpus for NLP applications training. On the contrary, we outline research directions that can utilize multilingual Wikipedia asymmetries to bridge the communication gaps in multilingual societies.

1 Introduction

Multilingual parallel corpora such as translations of fiction, European parliament proceedings, Canadian parliament proceedings, the Dutch parallel corpus are being used for training machine translation and paraphrase extraction systems. All of these corpora are parallel corpora.

Parallel corpora contain the same information translated from one language (the source language of the text) into a set of pre-specified languages with the goal of preserving the information covered in the source language document. Translators working with fiction also carefully preserve the stylistic details of the original text.

Parallel corpora are a valuable resource for training NLP tools. However, they exist only for a small number of language pairs and usually in a specific context (e.g., legal documents, parliamentary notes). Recently NLP community expressed a lot of interest in studying other types of multilingual corpora.

The largest multilingual corpus known at the moment is World Wide Web (WWW). One part of particular interest is the on-line encyclopedia-style site, Wikipedia.[1] Most Wikipedia entries (people, locations, events, etc.) have descriptions in different languages. However, Wikipedia is not a parallel corpus as these descriptions are not translations of a Wikipedia article from one language into another. Rather, Wikipedia articles in different languages are independently created by different users.

Wikipedia does not have any filtering on who can write and edit Wikipedia articles. In contrast to professional encyclopedias (like *Encyclopedia Britannica*), Wikipedia authors and editors are not necessarily experts in the field for which they create and edit Wikipedia articles. The trustworthiness of Wikipedia is questioned by many people (Keen, 2007).

The multilinguality of Wikipedia makes this situation even more convoluted as the sets of Wikipedia contributors for different languages are not the same.

[1] http://www.wikipedia.org/

Moreover, these sets might not even intersect. It is unclear how similar or different descriptions of a particular Wikipedia entry in different languages are. Knowing that there are differences in descriptions for the same entry and the ability to identify these differences is essential for successful communication in multilingual societies.

In this paper we present a preliminary study of the asymmetries in a subset of multilingual Wikipedia. We analyze the number of languages in which the Wikipedia entry descriptions are created; and the length variation for the same entry descriptions created in different languages. We believe that this information can be helpful for understanding asymmetries in multilingual Wikipedia. These asymmetries, in turn, can be used by NLP researchers for training summarization systems, and contradiction detection systems.

The rest of the paper is structured as follows. In Section 2 we describe related work, including the work on utilizing parallel corpora. In Section 3 we provide examples of our analysis for several Wikipedia entries. In Section 4 we describe our corpus, and the systematic analysis performed on this corpus. In Section 5 we draw conclusions based on the collected statistics and outline avenues for our future research.

2 Related Work

There exist several types of multilingual corpora (e.g., parallel, comparable) that are used in the NLP community. These corpora vary in their nature according to the tasks for which these corpora were created.

Corpora developed for multilingual and cross-lingual question-answering (QA), information retrieval (IR), and information extraction (IE) tasks are typically compilations of documents on related subjects written in different languages. Documents in such corpora rarely have counterparts in all the languages presented in the corpus (CLEF, 2000; Magnini et al., 2003).

Parallel multilingual corpora such as Canadian parliament proceedings (Germann, 2001), European parliament proceedings (Koehn, 2005), the Dutch parallel corpus (Macken et al., 2007), JRC-ACQUIS Multilingual Parallel Corpus (Steinberger et al., 2006), and so on contain documents that are exact translations of the source documents.

Understanding the corpus nature allows systems to utilize different aspects of multilingual corpora. For example, Barzilay *et al.* (2001) use several translations of the French text of *Gustave Flaubert*'s novel *Madame Bovary* into English to mine a corpus of English paraphrases. Thus, they utilize the creativity and language expertise of professional translators who used different wordings to convey not only the meaning but also the stylistic peculiarities of *Flaubert*'s French text into English.

Parallel corpora are a valuable resource for training NLP tools. However, they exist only for a small number of language pairs and usually in a specific context (e.g., legal documents, parliamentary notes). Recently NLP community expressed a lot of interest in studying comparable corpora. Workshops on building and using comparable corpora have become a part of NLP conferences (LREC, 2008; ACL, 2009). A comparable corpus is defined as a set of documents in one to many languages, that are comparable in content and form in various degrees and dimensions.

Wikipedia entries can have descriptions in several languages independently created for each language. Thus, Wikipedia can be considered a comparable corpus.

Wikipedia is used in QA for answer extraction and verification (Ahn et al., 2005; Buscaldi and Rosso, 2006; Ko et al., 2007). In summarization, Wikipedia articles structure is used to learn the features for summary generation (Baidsy et al., 2008).

Several NLP systems utilize the Wikipedia multilinguality property. Adafre *et al.* (2006) analyze the possibility of constructing an English-Dutch parallel corpus by suggesting two ways of looking for similar sentences in Wikipedia pages (using matching translations and hyperlinks). Richman *et al.* (2008) utilize multilingual characteristics of Wikipedia to annotate a large corpus of text with Named Entity tags. Multilingual Wikipedia has been used to facilitate cross-language IR (Schönhofen et al., 2007) and to perform cross-lingual QA (Ferrández et al., 2007).

One of the first attempts to analyze similarities and differences in multilingual Wikipedia is described in Adar *et al.* (2009) where the main goal

is to use self-supervised learning to align or/and create new Wikipedia infoboxes across four languages (English, Spanish, French, German). Wikipedia infoboxes contain a small number of facts about Wikipedia entries in a semi-structured format.

3 Analysis of Multilingual Wikipedia Entry Examples

Wikipedia is a resource generated by collaborative effort of those who are willing to contribute their expertise and ideas about a wide variety of subjects. Wikipedia entries can have descriptions in one or several languages. Currently, Wikipedia has articles in more than 200 languages. Table 1 presents information about the languages that have the most articles in Wikipedia: the number of languages, the language name, and the Internet Engineering Task Force (IETF) standard language tag.[2]

English is the language having the most number of Wikipedia descriptions, however, this does not mean that all the Wikipedia entries have descriptions in English. For example, entries about people, locations, events, etc. famous or/and important only within a community speaking in a particular language are not likely to have articles in many languages. Below, we list a few examples that illustrate this point. Of course, more work is required to quantify the frequency of such entries.

- the Wikipedia entry about Mexican singer and actress *Rocío Banquells* has only one description: in Spanish;
- the Wikipedia entry about a mountain ski resort *Falakro* in northern Greece has descriptions in four languages: Bulgarian, English, Greek, Nynorsk (one of the two official Norwegian standard languages);
- the Wikipedia entry about *Prioksko-Terrasny Nature Biosphere Reserve*, a Russia's smallest nature reserve, has descriptions in two languages: Russian and English;

[2] `http://en.wikipedia.org/wiki/List_of_Wikipedias`
Wikipedia is changing constantly. All the quotes and examples from Wikipedia presented and analyzed in this paper were collected on February 10, 2009, between 14:00 and 21:00 PST.

Number or Articles	Language	IETF Tag
2,750,000+	English	en
750,000+	German	de
	French	fr
500,000+	Japanese	jp
	Polish	pl
	Italian	it
	Dutch	nl

Table 1: Language editions of Wikipedia by number of articles.

- the Wikipedia entry about a Kazakhstani figure skater *Denis Ten* who is of partial Korean descent has descriptions in four languages: English, Japanese, Korean, and Russian.

At the same time, Wikipedia entries that are important or interesting for people from many communities speaking different languages have articles in a variety of languages. For example, *Newton's law of universal gravitation* is a fundamental nature law and has descriptions in 30 languages. Interestingly, the Wikipedia entry about *Isaac Newton* who first formulated the law of universal gravitation and who is know all over the world has descriptions in 111 different languages.

However, even if a Wikipedia entry has articles in many languages, the information covered by these articles can differ substantially. The two main sources of differences are:

- the amount of the information covered by the Wikipedia articles (the length of the Wikipedia articles);
- the choice of the information covered by the Wikipedia articles.

For example, Wikipedia entry about *Isadora Duncan* has descriptions in 44 languages. The length of the descriptions about *Isadora Duncan* is different for every language: 127 sentences for the article in English; 77 - for French; 37 - for Russian, 1 - for Greek, etc. The question arises: whether a shorter article can be considered a summary of a longer article, or whether a shorter article might contain information that is either not covered in a longer article or contradicts the information in the longer article.

Isadora Duncan was a American-born dancer who was very popular in Europe and was married to a Russian poet, *Sergey Esenin*. Certain amount of information facts (i.e., major biography dates) about *Isadora Duncan* are repeated in the articles in every language. However, shorter articles are not necessarily summaries of longer articles. For example, the article in Russian that is almost four time shorter than the articles in English, contains information that is not covered in the articles written in English. The same can be noted about articles in French and Spanish.

In this paper, we analyze the distribution of languages used in Wikipedia for the list of 48 people in the DUC 2004 biography generation task. We analyze, the number of languages that contain articles for each of the 48 DUC 2004 people. We also analyze the distribution of the lengths for the descriptions in different languages. We believe that this statistics is important for the understanding of the Wikipedia multilinguality nature and can be used by many NLP applications. Several NLP applications that can leverage this information are listed in Section 5.

4 Analysis of Wikipedia Multilinguality

In this paper, we propose a framework to quantify the multilinguality aspect of Wikipedia. In the current work we use a small portion of Wikipedia. Analyzing only a portion of Wikipedia allows us to compare in detail the multilinguality aspect for all the Wikipedia entries in our data set.

4.1 Data Set

For our analysis, we used the list of people created for the Task 5 of DUC 2004: biography generation task (48 people).[3]

First, we downloaded from Wikipedia all the articles in all the languages corresponding to each person from the DUC 2004 evaluation set. For our analysis we used Wikitext, the text that is used by Wikipedia authors and editors. Wikitext complies with the wiki markup language and can be processed by the Wikimedia content manager system into HTML which can then be viewed in a browser. This is the text that can be obtained through the Wikipedia dumps.[4] For our analysis we removed from the wikitext all the markup tags and tabular information (e.g., infoboxes and tables) and kept only plain text. There is no commonly accepted standard wikitext language, thus our final text had a certain amount of noise which, however, does not affect the conclusions drawn from our analysis.

For this work, for each Wikipedia entry (i.e., DUC 2004 person) we downloaded the corresponding descriptions in all the languages, including simple English, Esperanto, Latin, etc. To facilitate the comparison of descriptions written in different languages we used the Google machine translation system[5] to translate the downloaded descriptions into English. The number of languages currently covered by the Google translation system (41 language) is smaller than the number of languages in which there exist Wikipedia articles (265 languages). However, we believe that using for cross-lingual analysis descriptions only in those languages that can be handled by the Google translation system does not affect the generality of our conclusions.

4.2 Data Processing Tools

After the Wikipedia descriptions for each person from the DUC 2004 set were collected and translated, we divided the description texts into sentences using the LingPipe sentence chunker (Alias-i, 2009). We apply sentence splitter only to the English language documents: either originally created in English or translated into English by the Google translation system.

4.3 Data Analysis

As mentioned in Section 1, the goal of the analysis described in this paper is to quantify the language diversity in Wikipedia entry descriptions.

We chose English as our reference and, for each DUC 2004 person, compared a description of this person in English against the descriptions of this person in other languages.

Language count: In Figure 1, we present information about descriptions in how many languages are created in Wikipedia for each person from the DUC 2004 set. All the people from the DUC 2004

[3] http://duc.nist.gov/duc2004/tasks.html/

[4] http://download.wikimedia.org/

[5] http://translate.google.com/

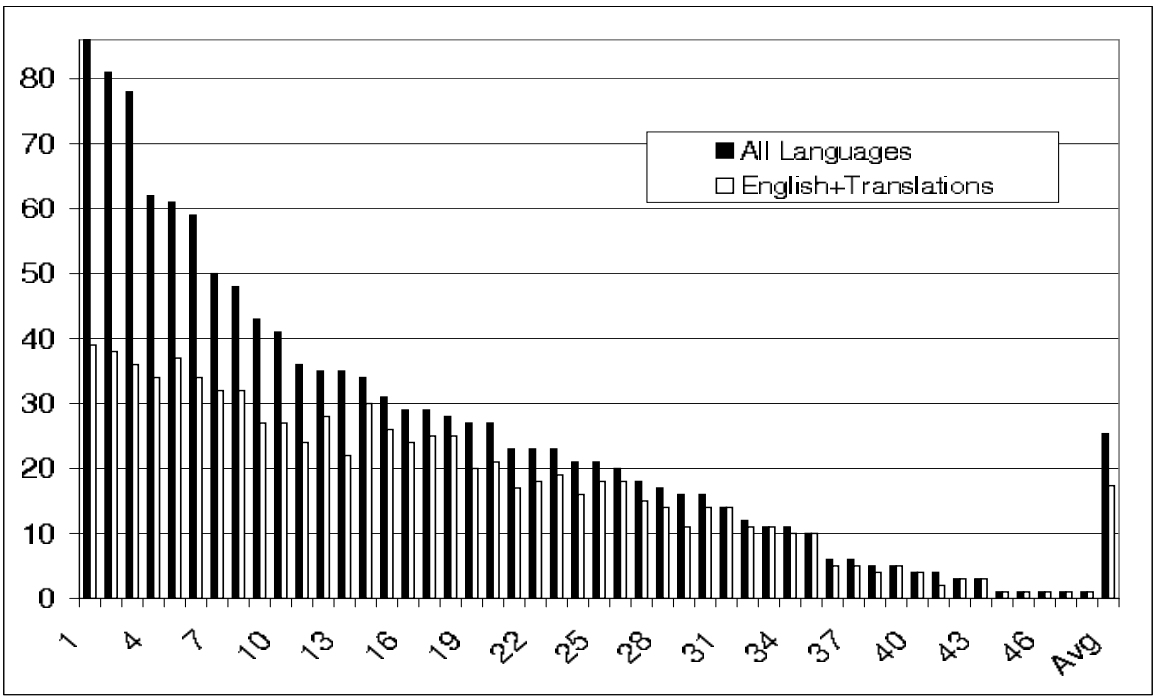

Figure 1: Number of languages for DUC 2004 people Wikipedia entries.

set have descriptions in English. The results in Figure 1 are presented in sorted order: from the Wikipedia entries with the largest number of descriptions (languages covered) to the Wikipedia entries with the smallest number of descriptions (languages covered). Five people from the DUC 2004 set have only one description (English). The person who has descriptions in the most number of languages for our data set is the former Secretary-General of the United Nations *Kofi Annan* (**86** languages). Figure 1 also has information about descriptions in how many languages were translated into English (handled by the Google translation system).

Despite the fact that English is the language having descriptions for more Wikipedia entries than any other language, it does not always provide the greatest coverage for Wikipedia entries. To show this we analyzed the length of Wikipedia entry descriptions for the people from the DUC 2004 set. For our analysis, the length of a description is equal to the number of sentences in this description. To count the number of sentences in the uniform way for as many languages as possible we used translations of Wikipedia description from languages that are currently handled by the Google translation system into English. Those five people from the DUC 2004 set that have descriptions only in English are excluded from this analysis. Thus, in the data set for the next analysis we have 43 data points.

Sentence count: For every Wikipedia entry (person from the DUC 2004 set), we count the length of the descriptions originally created in English or translated into English by the Google translation system. In Figure 2, we present information about the length of the Wikipedia entity descriptions for English and for the language other than English with the maximum description length. The results in Figure 2 are presented in sorted order: from the Wikipedia entry with the maximal longest description in the language other than English to the Wikipedia entry with the minimal longest description in the language other than English for our data set. This sorted order does not correspond to the sorted order from Figure 1. It is interesting so see that the sorted order in Figure 2 does not correlate to the length distribution of English descriptions for our data set.

Obviously, the descriptions in English are not al-

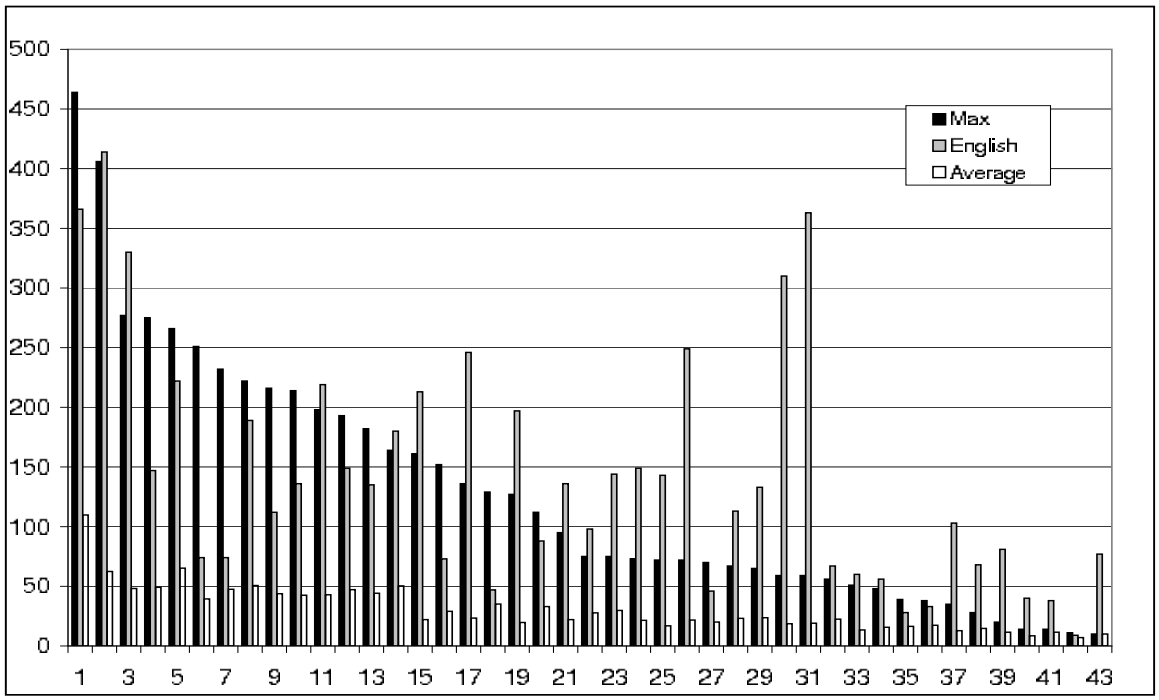

Figure 2: Number of sentences in the English description and the longest non-English description.

ways the longest ones. To be precise for 17 out of 43 people from the DUC 2004 set, the corresponding Wikipedia description in English was not the longest one. In several cases, the length of the description in English is several times shorter than the length of the longest (non-English) description. For example, the description of *Günter Grass* in German has 251 sentences while his description in English has 74 sentences.

It is safe to assume that longer descriptions have more information than shorter descriptions and 17 out of 43 English language descriptions of Wikipedia entries in our data set can be naturally extended with the information covered in the descriptions in other languages. Thus, multilingual Wikipedia gives a straight-forward way of extending Wikipedia entry descriptions.

It must be noted that the average length of Wikipedia descriptions (also presented on Figure 2) is very short. Thus, many descriptions for Wikipedia entries are quite short. The question arises how well the information covered in short descriptions corresponds to the information covered in long descriptions.

Correlation Analysis: In this paper, we present analysis for a small portion of Wikipedia. Currently, Wikipedia has more than more than $2,750,000$ articles in English alone. Thus, the question arises whether our analysis can be used without loss of generality for the complete Wikipedia (i.e., all descriptions for all Wikipedia entries).[6] To check this we analyzed the correspondence of how many Wikipedia entry descriptions are there for each language. For the Wikipedia subset corresponding to the people from the DUC 2004 set we simply counted how many Wikipedia entries have descriptions in each language. For the complete set of Wikipedia descriptions we used the Wikipedia size numbers from the *List of Wikipedias* page.[7] After getting the Wikipedia size numbers we kept the data only for those languages that are used for descriptions of Wikipedia entries corresponding to the DUC 2004 people.

To compute correlation between these two lists of numbers we ranked numbers in each of these lists. The Rank (Spearman) Correlation Coefficient for

[6] It must be noted that the notion of *complete* Wikipedia is elusive as Wikipedia is changing constantly.

[7] `http://en.wikipedia.org/wiki/List_of_Wikipedias`

the above two ranked lists is equal to 0.763 which shows a high correlation between the two ranked lists. Thus, the preliminary analysis presented in work can be a good predictor for the descriptions' length distribution across descriptions in the complete multilingual Wikipedia.

5 Conclusions and Future Work

In this papers we presented a way of quantifying multilingual aspects of Wikipedia entry descriptions. We showed that despite the fact that English has descriptions for the most number of Wikipedia entries across all languages, English descriptions can not always be considered as the most detailed descriptions. We showed that for many Wikipedia entries, descriptions in the languages other than English are much longer than the corresponding descriptions in English.

Our estimation is that even though Wikipedia entry descriptions created in different languages are not identical, they are likely to contain information facts that appear in descriptions in many languages. One research direction that we are interested in pursuing is investigating whether the information repeated in multiple descriptions of a particular entry corresponds to the pyramid summarization model (Teufel and Halteren, 2004; Nenkova et al., 2007). In case of the positive answer to this question, multilingual Wikipedia can be used as a reliable corpus for learning summarization features.

Also, our preliminary analysis shows that Wikipedia entry descriptions might contain information that contradicts information presented in the entry descriptions in other languages. Even the choice of a title for a Wikipedia entry can provide interesting information. For example, the title for the Wikipedia entry about *Former Yugoslav Republic of Macedonia* in English, German, Italian, and many other languages uses the term *Republic of Macedonia* or simply *Macedonia*. However, Greece does not recognize this name, and thus, the title of the corresponding description in Greek has a complete formal name of the country: *Former Yugoslav Republic of Macedonia*.

Multilingual Wikipedia is full of information asymmetries. Studying information asymmetries in multilingual Wikipedia can boost research in new information and contradiction detection. At the same time, information symmetries in multilingual Wikipedia can be used for learning summarization features.

References

ACL. 2009. Workshop on building and using comparable corpora: from parallel to non-parallel corpora.

Sisay Fissaha Adafre and Maarten de Rijke. 2006. Finding similar sentences across multiple languages in wikipedia. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, Workshop on New Text – Wikis and blogs and other dynamic text sources*, Trento, Italy, April.

Eytan Adar, Michael Skinner, and Dan Weld. 2009. Information arbitrage in multi-lingual Wikipedia. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, Barcelona, Spain, February.

David Ahn, Valentin Jijkoun, Gilad Mishne, Karin Müller, Maarten de Rijke, and Stefan Schlobach. 2005. Using Wikipedia at the TREC QA track. In *Proceedings of the Text REtrieval Conference (TREC 2004)*.

Alias-i. 2009. Lingpipe 3.7.0. (accessed January 19, 2009). `http://alias-i.com/lingpipe`.

Fadi Baidsy, Julia Hirschberg, and Elena Filatova. 2008. An unsupervised approach to biography production using wikipedia. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-2008)*, Columbus, OH, USA, July.

Regina Barzilaya and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001)*, Toulouse, France, July.

Davide Buscaldi and Paolo Rosso. 2006. Mining knowledge from wikipedia for the question answering task. In *Proceedings of The Fifth international Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy, May.

CLEF. 2000. Cross-language evaluation forum (CLEF). `http://www.clef-campaign.org`.

Sergio Ferrández, Antonio Toral, Óscar Ferrández, Antonio Ferrández, and Rafael Munoz. 2007. Applying Wikipedia's multilingual knowledge to cross-lingual question answering. *Lecture Notes in Computer Science (LNCS): Natural Language Processing and Information Systems*, 4592:352–363.

Ulrich Germann. 2001. Aligned hansards of the 36th parliament of Canada. Website.

http://www.isi.edu/natural-language/download/hansard/.

Andrew Keen. 2007. *The Cult of the Amateur: How Today's Internet is Killing Our Culture*. Doubleday Business.

Jeongwoo Ko, Teruko Mitamura, and Eric Nyberg. 2007. Language-independent probabilistic answer ranking for multilingual question answering. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007)*, Prague, Czech Republic, June.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Machine Translation Summit (MT-2005)*, Phuket Island, Thailand, September.

LREC. 2008. Workshop on building and using comparable corpora.

Lieve Macken, Julia Trushkina, and Lidia Rura. 2007. Dutch Parallel Corpus: MT corpus and translator's aid. In *Proceedings of the Eleventh Machine Translation Summit (MT-2007)*, pages 313–320, Copenhagen, Denmark, September.

Bernardo Magnini, Simone Romagnoli, and Ro Vallin. 2003. Creating the DISEQuA corpus: A test set for multilingual question answering. In *Proceedings of the Cross-Lingual Evaluation Forum (CLEF-2003)*, Trondheim, Norway, August.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The Pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, 4(2).

Alexander Richman and Patrick Schone. 2008. Mining Wiki resources for multilingual named entity recognition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-2008)*, Columbus, OH, USA, July.

Péter Schönhofen, András Benczúr, István Bíró, and Károly Csalogány. 2007. Performing cross-language retrieval with wikipedia. In *Proceedings of the Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, September.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufis, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of The Fifth international Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy, May.

Simone Teufel and Hans Van Halteren. 2004. Evaluating information content by factoid analysis: Human annotation and stability. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL-2004)*, Barcelona, Spain, July.