Reporting Preliminary Automatic Comparable Corpora Compilation Results

Ekaterina Stambolieva
University of Wolverhampton
West Midlands WV1 1LY
United Kingdom
ekaterina.stambolieva@euroscript.lu

Abstract

Translation and translation studies rely heavily on distinctive text resources, such as comparable corpora. Comparable corpora gather greater diversity of language-dependent phrases in comparison to multilingual electronic dictionaries or parallel corpora; and present a robust language resource. Therefore, we see comparable corpora compilation as impending in this technological era and suggest an automatic approach to their gathering. The originality of the research lies within the newly-proposed methodology that is guiding the compilation process. We aim to contribute to translation and translation studies professionals' work by suggesting an approach to obtaining comparable corpora without intermediate human evaluation. This contribution reduces time and presents such professionals with non-static text resources. In our experiment we compare the automatic compilation results to the labels, which two human evaluators have given to the relevant documents.

1 Introduction

In translation and translation studies large collections of texts serve as invaluable resources, which help translators interpret better and faster previously unseen text, extract terms, and look for context-dependent translation equivalents. These big sets of texts are referred to as corpora within professional literature. According to Aarts (1991) "a corpus is understood to be a collection of samples of running texts. The texts may be spoken, written or intermediate forms, and the samples may be of any length". Furthermore, the corpus content is collected by following unambiguous linguistic criterion (EAGLES, 1996). A widely used translator's working tool are electronic multilingual dictionaries, which store words and their equivalents in different languages. Nevertheless, the electronic dictionaries lack some translation equivalents and as they are static, this gap is not filled in.

The constant enrichment of the languages themselves results in the birth of new words, terms and translation equivalents on a regular basis. The static electronic dictionaries are difficult to update frequently, hence they are not described as a highly-robust resource for mining translation alternatives. A valuable alternative source of textual materials that aids translators is parallel corpus. The parallel corpus is compiled of snippets of text that are aligned on sentence level and are exact translations of each other in one or more languages. This kind of corpora is a perfect language resource for translators. When in doubt, the translators can explore the available parallel corpora, either with the use of specialised software or not, to analyse language structures, unknown phrases, register, and so on. Talvensaari et al. (2007) state the translation process with the use of comparable corpora as a similarity thesaurus improves the quality of the translations. However, collections of compiled parallel texts are scarce and their domain coverage is poor. Some topic specific parallel corpora exist, such as the EuroParl set (Koehn, 2005), grouping legislative documents written in one of the twenty-three official European Languages. Here comes the advantages of using comparable corpora over parallel ones or dictionaries - the comparable corpora are more robust than electronic dictionaries, and are more available than parallel corpora.

A good stimulus motivating the current research is that comparable corpora preserve the all language structures and ways of expressions, thereupon keeping all cultural aspects of the language. This is also suggested by Bekavac et al. (2004). They emphasise on the importance of

Proceedings of the Student Research Workshop associated with RANLP 2013, pages 149–156,
Hissar, Bulgaria, 9-11 September 2013.

comparable corpora with respect to the fact such collections preserve the cultural variations of the languages involved. Contrary to direct translation snippets, comparable texts can convey the most important information to the readers, following each specific language construction and structure. In this like of thought the parallel text corpora can suffer from lack of language-specific cultural marks because of the fact they require exact translations rather than preserving the language variety richness. Another idea that inspires research in the compilation of comparable corpora problem is that similar texts may be easier to find. Therefore, when a good methodology to the gathering of such collections is presented, the accessible similar texts can be collected to help researchers and professionals in translation. Likewise, Skadiņa et al. (2010 b) and Skadiņa et al. (2010a) argue that the advantages of comparable corpora in Machine Translation are considerable and more beneficial than those of parallel corpora. The researchers state that comparable corpora are a good substitute for parallel ones and they can compensate for the parallel corpora's lack.

2 Corpus. Parallel and Comparable.

Definition and explanation of the most important terms to be known is provided: a corpus and the different types of corpora that can be collected. In the work of Bowker and Pearson (2002) a detailed explanation on the importance of corpora is given. Depending on the purpose of the corpus, several different types of corpora can be categorised. Bowker and Pearson (2002) argue distinct corpora exist. Relying on the purpose they have been constructed for, the corpora can be general reference ones or specific purpose ones. Written and spoken corpora are classified depending on the electronic format data they consist of: either text or speech files accordingly. The variety of languages to be identified in the corpora group them into monolingual and multilingual. "A monolingual corpus is one that contains texts in a single language, while multilingual corpora contain texts in two or more languages." (Bowker and Pearson 2002) The corpora build from collections of documents in two languages are called bilingual, and in the cases with more than two present languages, the corpora are referred as multilingual.

2.1 A Parallel Corpus

The multilingual corpora are divided into subcategories that are parallel and comparable corpora. Bowker and Pearson (2002) restrict the monolingual corpora in the sense they do not dissemble them into parallel and comparable. In translation and translation studies a monolingual corpus can be built to be comparable but not parallel. The definition of parallel corpora according to Bowker and Pearson (2002) is "parallel corpora contain texts in language A alongside their translations into language B, C, etc." Thus a corpus build from documents in the same language cannot contain more than one ways of presenting the same exact information, meaning that the only translation a snippet of text can have in the same language is the initial snippet of text itself.
In other hand, the comparable corpora consist of texts in several languages that are not exact interpretations of one another, but having the same communicative function. Some comparable corpora indicators are listed as time-frame, topic, degree of technicality, and type of text.

2.2 A Comparable Corpus

The degree of similarity between comparable corpora documents has not yet been formalised strictly and leaves space for different interpretations of similarity, thus contributing to abundant text collections of similar or semi-similar documents. The current research endeavors to assemble a collection of comparable documents that are closely related to each other and can be used by professional translators in their everyday work. The adopted definition of comparable corpora for this work is provided by McEnery (2003) - "Comparable corpora are corpora where series of monolingual corpora are collected for a range of languages, preferably using the same sampling and frame and with similar balance and representativeness, to enable the study of those languages in contrast" (McEnery 2003).
Otero and López (2010) provide a simplified description of comparable corpora than McEnery (2003). Their definition is "a comparable corpus is one which selects similar texts in more than one language or variety".

In like manner, Talvensaari et al. (2007) interpret comparable corpora. In their views, "comparable corpora consist of document pairs that are not translations of each other but share similar topics." According to Tao and Zhao (2005) "Comparable text corpora are collections of text documents in different languages that are similar

about topics; such text corpora are often naturally available (e.g., news articles in different languages published in the same time period)". In a like manner they argue that "comparable text corpora are collections of text documents in different languages that are about the same or similar topics." Fung and Cheung (2004) define comparable corpora as being noisy-parallel: "A noisy parallel corpus, sometimes also called a 'comparable' corpus, contains non-aligned sentences that are nevertheless mostly bilingual translations of the same document. Another type of corpus is one that contains non-sentence-aligned, non-translated bilingual documents that are topic-aligned. For example, newspaper articles from two sources in different languages, within the same windows of published dates, can constitute a comparable corpus." (Fung and Cheung 2004).

Skadiņa et al. (2010a) describe comparable corpora in a slightly different manner. They are referring to a comparable corpus as a "collection of similar documents that are collected according to a set of criteria, e.g. the same proportions of texts of the same genre in the same domain from the same period ... in more than one language or variety of languages ... that contain overlapping information." The comparability features they are using hence are genre, domain, size, and time span. The level of comparability of corpora can be distinct depending on the texts in the documents. An important is that Skadiņa et al. (2010a) define different levels of comparability between documents. They distinguish three different types of comparable corpora or three separate levels of similarity. The first one is called strongly comparable corpora. The strongly comparable texts are "closely related texts reporting the same event or describing the same subject. These texts could be heavily edited translations or independently created, such as texts coming from the same source with the same editorial control, but written in different languages... or independently written texts concerning the same object, e.g. news items concerning the same specific event from different news agencies". The second level of similarity is when the documents are marked as weakly comparable. The weakly comparable documents are "texts which include texts in the same narrow subject domain and genre, but varying in subdomains and specific genres". Hence the similarity features of the documents collected in a weakly comparable corpus are genre and domain. The reason these types of documents are classified as weakly similar is that in the different genres of distinct domains the texts are not restricted to be describing the same event as if they were strongly comparable. The last type of comparable texts Skadiņa et al. (2010a) propose is a non-comparable corpus. The non-comparable texts are described as "pairs of texts drawn at random from a pair of a very large collection of texts (e.g. the Web) in the two languages".

3 Relevant Literature

Relevantly to the current research, Gatto (2010) gives a perspective on how comparable corpora are built and explored from translators in LSP translation. She emphasises on the fact the manual acquisition of comparable corpora "for a specific translation task ... is deemed too time-consuming, and the results are more often than not disappointing." Gatto (2010) explores the benefits of a semi-automatic comparable corpora compilation tool in a class-based environment for translators. As most of the work on building comparable corpora, for example as in Tao and Zhai (2005), Gatto is focused on bilingual document sets instead of exploring multilingual texts. She indicates the scarcity of the parallel and comparable corpora resources available ad hoc to translators. In her study she investigates the problem of building a similar document collection that is fast to assemble and in the same time beneficial and appropriate to the translators' needs. She seeks for a tool that can support translation trainees in their activities that is "primarily conceived of as a tool helping language professionals build the corpus they need, whenever they need, and as quickly as possible" (Gatto 2010). The tool Gatto evaluates with her students has web access and performs seed word searches online. Therefore, using the Web as a corpus (Kilgarriff 2003) and information retrieval techniques, a comparable corpus is assembled. The aspect that Gatto ephrasises on in her work is that at each step the tool waits for human verification of results. She argues the latter is an important contribution to more accurate comparable document selection for the reason dubious texts is manually checked for relevance and comparability. In Gatto's research, the retrieved web pages are based on automatic criterion and human intelligence selection. An important remark stated by Gatto (2010) is that a web crawling tool for building comparable corpora performs "better than a student can manually do, while still allowing significant interaction with the machine".

The conclusion is that such a semi-automatic system outperforms translation students' efforts to compiling a comparable corpus. This assumption gives motivation further research in the manners of developing software to collect similar documents to ease translator's work to be undertaken.

Corpora, being parallel or comparable, can be extracted from the Web. The work of many researchers, as Gatto (2010), Ion et al. (2010), Otero and López (2010), Skadiņa et al. (2010b), Talvensaari et al. (2008), shows different techniques to their automatic and semi-automatic gathering. Kilgarriff (2003) argues that the entire Web can be recognised as one corpus. Skadiņa et al. (2010b) note that the comparable documents are mined without great difficulty since they are more available than the parallel texts.

Contrary to mining corpora from the Web, many research papers are dedicated to employing multiple similarity metrics. These evaluate the degree of comparability between documents in the collections. Examples of works that aim to find comparability features and scores between documents are those of Bekavac et al. (2004), Sharoff (2010), Steinberg et al. (2006), and Talvensaari (2007).

As in the work of Talvensaari et al. (2008), the web crawling of the potential similar texts is initiated by providing a set of seed words to be queried to a web search engine. The results are retrieved and post-processed, and new keywords to serve as seed words for a consecutive search are extracted. The technique to election of keywords is a simple frequency word count. In the current research we concentrate on using whole documents as seeds to mine similarity.

A good stimulus motivating the current research is that comparable corpora preserve the all language structures and ways of expressions, thereupon keeping all cultural aspects of the language. This is also suggested by Bekavac et al. (2004). They emphasise the importance of comparable corpora with respect to the fact they preserve the cultural variations of the languages involved.

Skadiņa et al. (2010b) and Skadiņa et al. (2010a) argue that the advantages of comparable corpora in Machine Translation are considerable and more beneficial than those of parallel corpora. The researchers suggest that comparable corpora are easier and more available to collect online than parallel ones as one of obvious benefits. Also, they suggest the texts in the comparable corpora gather greater diversity of language-dependent phrases, terms, and ways of expression. An interesting observation that Skadiņa et al. (2010a) make is that comparable corpora are a good substitute for parallel ones and they can compensate for the parallel corpora's lack.

Concentrating on comparability metrics is vital for the research of automatic compilation of comparable corpora. Skadiņa et al. (2010b) focus additionally on relevance evaluation metric design. The aim of Kilgarriff (2003) includes the comparability evaluation between two collections of documents and the advantages/disadvantages of known evaluation metrics. Saralegi et al. (2008), as Tao and Zhao (2005), compare documents based on time-frame topic distributions delineated metric. Similarity metrics on word level are discussed by Deerwester et al. (1990); Dagan, Lee and Pereira (1999); and Baeza-Yates and Ribeiro-Netto (1999). Lee (1999) and Dagan et al. (1999) rely on word-co-occurrence text comparison. The current research incorporates a Latent Semantic Analysis (LSA) technique as in Radinsky et al. (2001) and in Deerwester et al. (1990).

4 Approach

The proposed methodology incorporates Latent Semantic Analysis (LSA) and unsupervised machine learning (ML), the k-means algorithm, to automatically collect a comparable document corpus from a given set of texts (Stambolieva 2013). LSA is employed to identify word similarity and map this similarity to concepts. By identifying such concepts LSA reduces the space of the documents to be asserted to a two-dimensional one. In the current scenario, each concept consists of a normalized word form, a lemma, with its correspondent context-dependent part-of-speech tag. In order to for the concepts to be more context-aware, noun phrases in both languages are identified and included in the concept space with a NP part-of-speech tag.

Additionally, the ML algorithm learns from the similar concept space and predicts which documents are comparable to each other and which are not. Moreover, a possibility to identify more than one comparable corpus is presented to the learning algorithm.

To the best of our knowledge, an approach to the compilation of comparable corpora that relies on LSA with k-means has not been suggested yet. We invest into presenting a reasoned definition of the notion of comparable corpora. Accompanying to that, we perform language analy-

sis tasks such as lemmatization and noun phrase identification to investigate whether these tasks help learn comparable corpora more accurately.

5 Data

The experimental corpus is manually collected following a procedure of document collection translators follow (Zanettin 2002), when compiling their own specific purpose comparable corpora. The corpus contains documents in the narrow topic of psychometrics, in particular psychometric properties and evaluation. Noise is included in the corpora as some texts that are not on psychometrics, but still on psychology, are added. Additionally, newswire texts than have no resemblance at all with the suggested similar psychometrics documents are provided as a supplementary noise. The domain of the collected documents is psychology since psychometrics is a sub-topic of psychology. The corpus is consistent of documents written either in English or Spanish. The total number of documents, which are manually collected, is 26. We try to mimic the process translators choose related linguistics resources during translation. As time is of importance they would not invest much of it in searching for comparable documents, therefore we decided 26 is a sufficient number for the current experiment.

The distribution of topics in the psychometrics corpus is 6 psychometrics texts in Spanish, 9 psychometrics texts in English, 3 psychology but not psychometrics texts, and 8 non-psychology texts in English. Two manual evaluators label the documents in the corpus as comparable or not according to a set of evaluation guidelines. Table 1. shows the how the evaluators label the collection of Spanish and English texts.

Evaluator	Psychometrics + Psychology	Newswire
Evaluator 1	15	11
Evaluator 2	18	8

Table 1: Evaluators' manual comparability labels

6 Evaluation Metrics

The evaluation metrics used to evaluate the performance of the suggested methodology are precision, recall, purity, mutual information (MI), entropy (H) and normalized mutual information (NMI). These metrics are all explained in details by Manning et al. (2008).

7 Experiment

The aim of this experiment is to assemble a comparable corpus from different documents, in which some are found comparable and others are withdrawn from the elected comparable set due to similarity disagreement. Thus, the experimental corpus accumulates roughly two types of texts, therefore can be separated into two subsets – psychometrics (and psychology), and newswire category. Therefore, we aim at compiling a weakly-comparable bilingual corpus (Skadiņa et al. 2010a), whose domain is psychology and which contains psychology and psychometrics texts. Experiments with different k, number of resulting clusters, are performed. When k equals the number of manually evaluated number of categories, namely two, the purity of the resulting corpus is calculated. The purity score is 0.6538, which is not close to 1. Purity translates the corpus quality trade-off dependently on the number of clusters The purity result indicates that documents from both the two different labels are collected together into a comparable cluster. The precision scores of the run experiments with 2, 3, 4 and 5 clusters to be identified are shown in Table 2. The recall scores of the run experiments with 2, 3, 4 and 5 clusters to be identified are shown in Table 3. *2cl*, *3cl*, *4cl*, and *5cl* respectively show learning text comparability results when 2, 3, 4 and 5 resulting clusters are compiled.

Topic	2cl	3cl	4cl	5cl
Psychometrics + Psychology	1	1	1	1
Newswire	0.42	1	1	1

Table 2: Clustering precision

Topic	2cl	3cl	4cl	5cl
Psychometrics + Psychology	0.35	0.65	0.83	0.65
Newswire	1	0.34	0.61	0.42

Table 3: Clustering recall

The precision of most of the resulting clusters equals 1, which means the documents from the same category, psychology, are appropriately grouped together. The recall shows another fashion that is occurring in the resulting clusters. The lower the score is, the closer to 0, the larger number of documents labeled in the other category, newswire, are also grouped together with

the correctly identified psychology ones. The last observation means in the case of correctly grouped documents that are comparable to each other, texts that are not similar to them are also nominated and selected as part of the comparable corpus.

The corpus is very heterogeneous in the sense it consists of articles written in both Spanish and English in categories such as psychometric evaluation, psychometric properties, psychology, and press texts. Hence, the learning algorithm is not able to produce better results by learning from the identified concepts in the corpus. A cause for that fact, except the heterogeneity of the experimental corpus, can be the distribution of concepts over the documents. Moreover, when Spanish documents are preprocessed, a translation engine, Google Translate[1], is used as the main source of mining translation equivalents into English. Nevertheless it is constantly being enriched with new translation pairs and is a very robust source of interpretations; the translations it has provided are not to be considered perfect and can leave room for mistakes. Therefore, the translation output reflects directly on the distribution of concepts in the documents of the Spanish-English corpus.

To further explore the clustering quality of the comparable corpus selected, when two clusters are expected, *2cl*, NMI is calculated (see Table 5.). NMI requires MI, H(Ω) and H(C) calculations, which are respectfully the mutual information between the documents in the cluster, or the comparable corpus, the entropy of the documents with the same label, and the entropy between the document with the same class – comparable or non-comparable. Opposed to purity, the NMI metric is used to show the quality of clusters independently on the number of clusters. NMI is roughly 0.54, which indicates the normalized mutual information between the texts in the automatically compiled comparable collection is not high. In it, there are 11 psychology documents and 8 newswire ones, out of 18 psychology and 8 newswire texts. This results show the approach has difficulties disambiguating between the newswire and psychology texts and that text similarity is found between then when it should not. We hope further investigations will suggest improvements to the methodology in order for it to increase performance.

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

No. Clusters	MI	H(Ω)	H(C)	NMI
2	0.5025	0.8511	0.9842	0.5475

Table 5: Mutual Information, Entropy and Normalized Mutual Information over clustering results of two corpora

8 Future Work

A further improvement of the methodology is to involve human translators in judging the results of the comparable corpora compiled. The linguistic analysis tasks are prone to mistakes, which can reflect on the learning algorithm performance. Further improvement of their performance can only prove beneficial to our research. Furthermore, a new source of translation, which suggests better translation equivalents, is welcome. Recognition of diasystematic text markers, such as diachronic ones, can suggest new potential meta-information features to be considered when searching for comparability between documents.

Including all of the aforementioned, we aim at collecting a bigger initial document set on which we can evaluate our approach. Future works additionally include extending the methodology to cover other languages than English and Spanish.

9 Conclusions

This paper presents preliminary results on the automatic compilation of comparable corpora with respect to their usage in translator's work. We aim to develop a systematic methodology, which relies on LSA and a ML algorithm, to ease the comparable corpora collection by translation professional. We critically discuss our results obtained on a small experimental bilingual corpus and propose further development suggestions.

References

Jan Aarts. 1991. Intuition-based and Observation-based Grammars. In K. Aijmer and B. Altenberg, editors, *English Corpus Linguistics: Studies in Honour of Jan Svartik*, pages 44-65, Longman, London.

Ricardo Baeza-Yates and Betrhier Ribeiro-Neto. 1999. *Modern Infromation Retieval*, Addison Wesley.

Božo Bekavac, Petya Osenova, Kiril Simov and Marco Tadic. 2004. Making Monolingual Corpora Comparable: a Case Study of Bulgarian and Croa-

tian. In *Proceedings of LREC2004*, pages 1187-1190, Lisbon.

Lynne Bowker and Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London. New York: Routeledge.

Igo Dagan, Lillian Lee and Fernando Pereira. 1999. Similarity-based models of word co-occurrence probabilities. *Machine Learning*. 34(1-3):43-69.

Scott Deerwester, Susan Dumais and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the Americal Society for Information Science*. 41(6):391-407.

EAGLES, The European Advisory Group on Language Engineering Standards. 1996. Available at <http://www.ilc.cnr.it/EAGLES/home.html>.

Pascale Fung and Percy Cheung. 2004. Mining Very Non-Parallel Corpora: Parallel Sentence and lexicon Extraction via Bootstraping and EM. In *Proceedings of EMNLP*, pages 57-63, Barcelona, Spain.

Maraistella Gatto. 2010. From language to culture and beyond: building and exploring comparable web corpora. In R. Rapp, P. Zweigenbaum, and S. Sharoff, editors, *Proceedings of the Third Workshop on Building and Using Comparable Corpora: applications of Parallel and Comparable Corpora in Natural Language Engineering and the Humanities (BUCC 2010)*, pages 72-78, Paris, France.

Radu Ion, Dan Tufiș, Tiberiu Boroș, Alexandru Ceaușu and Dan Ștefănescu. 2010. On-line Compilation of Comparable Corpora and Their Evaluation. In *Proceedingds of the 7th International Conference of Formal Approaches to South Slavic and Balkan Languages (FASSBL7)*, pages 29-34, Dubrovnic, Croatia.

Phillip Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit*, pages 79-86, Phuket, Thailand.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Lingusitics*, 6(1):97-133.

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of ACL 1999*, pages 25-32.

Christopher D. Manning, Prabhakan Raghavan and Hinrich Schűtze. 2008. *Introduction to Information Retrieval*, pages 356-358, Cambridge University Press.

Tony McEnery. 2003. Corpus Linguistics. In R. Mitkov, editor, *The Handbook of Computational Lingustics*. pages 448-464, Oxford University Press, Oxford.

Pablo Gamallo Otero and Isaac González López. 2010. Wikipedia as Multilingual Source of Comparable Corpora. In *Proceedings of the 3rd workshop on BUCC(LREC 2010)*, pages 21-25.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich and Shaul Markovitch. 2011. A word at a time: Computing Word Relatedness using Temporal Semantic Analysis. In *WWW'11*, pages 337-346.

Xabier Saragueli, Inaki San Vincente and Antton Gurrutxaga. 2002. Automatic Extraction of bilingual terms from comparable corpora in a popular science domain. In *Proceedings of the workshop on Comparable Corpora, LREC'08*.

Serge Sharoff. 2010. Analysing similarities and differences between corpora. *In Proceedings of the 7th Conference of Language Technologies (Jezikovne Tehnologije)*, pages 5-11, Ljubljiana. Slovenia.

Inguna Skadiņa, Ahmet Aker, Voula Giouli, Dan Tufiș, Robert Gaizauskas, Madara Mieriņa and Nikos Mastropavlos. 2010a. A Collection of Comparable Corpora for Under-Resourced Languages. In I. Skadiņa and D. Tufiș, editors, In *Proceedings of the 4th International Conference Baltic HLT 2010*, pages 161-168.

Inguna Skadiņa, Andrejs Vasiljeiv, Raivis Skadiņš, Robert Gaizauskas, Dan Tufiș and Tatiana Gornostay. 2010b. Analysis and Evaluation of Compoarable Corpora for Under Resourced Areas of Machine Translation. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora. Applications of Parallel and Comparable Corpora in Natural Language Engineering and the Humanities*, pages 6-14.

Ekaterina Stambolieva. 2013. Learning Comparable Corpora from Latent Semantic Analysis Simplified Document Space. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora (BUCC 2013)*, pages 129-137, Sofia, Bulgaria.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiș and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pages 2142-2147, Genoa, Italy.

Tuomas Talvensaari, Jorma Laurikkala, Kalevro Järvelin, Martti Juhola, and Heikki Keakustalo. 2007. Creating and Exploiting a Comparable Corpus in

Cross-language Information Retrieval. *ACM Translations on Information Systems*, 25(1).

Tuomas Talvensaari, Ari Pirkola, Kalevro Järvelin, Martti Juhola and Jorma Laurikkala. 2008. Focused Web Crawling in the acquisition of comparable corpora. *Information Retrieval*, 11:427-445.

Tao Tao and Cheng Xiang Zhai. 2005. Mining Comparable Bilingual Text Corpora for Cross-Language Information Integration. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 691-696.