A Collection of Comparable Corpora for Under-resourced Languages

Inguna SKADIŅA [a], Ahmet AKER [b], Voula GIOULI [c], Dan TUFIS [d], Robert GAIZAUSKAS [b], Madara MIERIŅA [a], Nikos MASTROPAVLOS [c]
[a] *Tilde, Latvia*
[b] *University of Sheffield, UK*
[c] *Athena Research and Innovation Center in Information Communication and Knowledge Technologies, Greece*
[d] *Research Institute for Artificial Intelligence, Romanian Academy Bucharest, Romania*

Abstract. This paper presents work on collecting comparable corpora for 9 language pairs: Estonian-English, Latvian-English, Lithuanian-English, Greek-English, Greek-Romanian, Croatian-English, Romanian-English, Romanian-German and Slovenian-English. The objective of this work was to gather texts from the same domains and genres and with a similar level of comparability in order to use them as a starting point in defining criteria and metrics of comparability. These criteria and metrics will be applied to comparable texts to determine their suitability for use in Statistical Machine Translation, particularly in the case where translation is performed from or into under-resourced languages for which substantial parallel corpora are unavailable. The size of collected corpora is about 1million words for each under-resourced language.

Keywords. comparable corpora, under-resourced languages, comparability, metadata, crawling, statistical machine translation

Introduction

In recent decades data-driven approaches have significantly advanced the development of machine translation (MT). However, the applicability of current data-driven methods directly depends on the availability of very large quantities of parallel corpus data. For this reason the translation quality of current data-driven MT systems varies from being quite good for language pairs/domains for which large parallel corpora are available to being barely usable for languages with fewer resources or in narrow domains.

The problem of availability of linguistic resources is especially relevant for under-resourced languages, including languages of the three Baltic countries – Estonian, Latvian and Lithuanian. One potential solution to the bottleneck of insufficient parallel corpora is to exploit comparable corpora to provide more data for MT systems.

The concept of a comparable corpus is a relatively recent one in MT and NLP in general. It can be defined as 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 [1] in more than one language or variety of languages [2] that contain overlapping information [3][4]. Comparable corpora have several obvious advantages over parallel corpora – they are available on the Web in large quantities for many languages and domains and many texts with similar content are produced every day (e.g. multilingual news feeds).

Recent experiments have demonstrated that a comparable corpus can compensate for the shortage of parallel corpora. Hewavitharana and Vogel [4] have shown that adding extracted aligned parallel lexical data from comparable corpora to the training data of an Statistical Machine Translation (SMT) system improves the system's

performance with respect to un-translated word coverage. It has been also demonstrated that language pairs with little parallel data are likely to benefit the most from the exploitation of comparable corpora. Munteanu and Marcu [3] achieved performance improvements of more than 50% using comparable corpora of BBC news feeds for English, Arabic and Chinese over a baseline MT system trained only on existing available parallel data.

The FP7 project Accurat [5] [6] aims to find, analyze and evaluate methods that exploit comparable corpora in order to compensate for the shortage of linguistic resources, and to significantly improve MT quality for under-resourced languages and narrow domains. This paper presents work on the creation in Accurat of bilingual comparable corpora for 9 language pairs: Estonian-English, Latvian-English, Lithuanian-English, Greek-English, Greek-Romanian, Croatian-English, Romanian-English, Romanian-German and Slovenian-English, where in each corpus at least one languages is under-resourced. The objective was to gather texts which can be used as starting point to define criteria and metrics of comparability, i.e., determine what degree of comparability is 'preferred', 'suitable', or 'minimally acceptable' for texts used for MT. The metrics will be used to define certain routes for exploiting comparable corpora in MT.

We present an initial definition of comparability, principles used for collecting textual data, a proposal for metadata encoding and tools used for collection, and also describe useful sources and problems we experienced.

1. Principles of Collecting Comparable Corpora

Until now there has been no agreement on the degree of similarity that documents in comparable corpora should have, or agreement about the criteria for measuring parallelism and comparability. Objective measures for detecting how similar two corpora are in terms of their lexical content have been studied only recently [7] [8]. Thus for our task we have introduced four comparability levels – parallel, strongly comparable, weakly comparable and non-comparable.

By *parallel* texts we understand true and accurate translations or approximate translations with minor language-specific variations. Typical samples of parallel texts by our definition are legal documents, software manuals, fiction translations, etc.

By *strongly comparable* texts we understand 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 (e.g., news provided by Baltic News Service in English, Latvian and Russian), or, independently written texts concerning the same subject, e.g., Wikipedia articles linked via the wiki or news items concerning the same specific event from different news agencies (e.g. the 2010 FIFA World Cup).

The third category is *weakly comparable* texts which include texts in the same narrow subject domain and genre, but describing different events, as well as texts within the same broader domain and genre, but varying in subdomains and specific genres (e.g., a user manual for LinguaType European in Latvian and a database administrator guide for MySql in English).

Finally, we can speak about *non-comparable* texts: pairs of tests drawn at random from a pair of very large collections of texts (e.g. the web) in the two languages.

Our goal was to collect 1 million running words for each language with the same distribution between domains and genres (see Table 1) and with the similar proportions between comparability levels (10% parallel texts, 40% strongly comparable texts, 50% weakly comparable texts).

Table 1. Domain and genre distribution of Accurat comparable corpora

Domain	Genre	Coverage
International news	Newswires	20%
Sports	Newswires	10%
Admin	Legal	10%
Travel	Advice	10%
Software	Wikipedia	15%
Software	User manuals	15%
Medicine	For doctors	10%
Medicine	For patients	10%

2. Metadata

For encoding both documents and the alignments between them we use *A Comparable Corpus Encoding Schema* (ACCES). It is an adaptation of the *Corpus Encoding Standard* (CES) structure and contains further metadata elements specific to the Accurat project, but potentially of use for any comparable corpus. Its structure is as shown below.

```
<cesdoc id="file123.txt" lang="el" type="text" version="1">
   ...
    <extendedsourcedesc>
      <genre>newswires</genre>
      <domain>international news</domain>
      <publicationresource>URL of text </publicationresource>
      <encoding>utf-8</encoding>
      <publicationdate>19/12/2007</publicationdate>
      <textcleaningnote>how was raw text extracted from HTML source doc </textcleaningnote>
    </extendedsourcedesc>
   …
  <htmlsource>Html source with entities encoded</htmlsource>
</cesdoc>
```

The new tags are “extendedsourcedesc” and “htmlsource”. The “extendedsourcedesc” tag encodes information about the genre, domain, source of the document, encoding of the text, date of the publication and information about any technique used to clean the original html document to obtain raw text from it.

The “htmlsource” tag includes the original html source, i.e. the entire document content found on the web. This is included because its structure may supply information that can be used in deciding whether a pair of documents are parallel, strongly or weakly comparable (of course the html source file may be useful for other purposes too). When saving the content into the XML structure, we ensure that the XML structure is still well-formed, i.e. all HTML special characters are encoded so that the HTML can be placed inside the XML without violating the structure.

It should be noted that ACCES follows the CES structure, i.e. the order of the elements occurring in the file is the same as in CES, and it also includes all mandatory elements from CES for representing a document. Thus it is possible to use a CES parser to parse the ACCES structure.

For expressing the alignments between documents the following structure is used. This structure is again based on CES with small extensions to meet Accurat-specific requirements.

```
<cesalign version="1">
 <linklist>
  <linkgrp targType="doc" alignmentlevel="strongly comparable" alignmentdecision="manual">
   <link xtargets="doc1.xml ; doc2.xml"></link>
  </linkgrp>
 </linklist>
</cesalign>
```

In CES each alignment is expressed in the "linkgrp" tag. It contains the alignment types (document, paragraph, sentence, etc.) and the aligned text pairs (in the "link" tag). In the example shown above we have alignment at document (doc) level where we have aligned document "doc1.xml" with "doc2.xml". We have extended the "linkgrp" tag with two further attributes which help to express the alignment level ("alignmentlevel" with possible values "parallel", "strongly comparable", "weakly comparable") and information about how the alignment level ("alignmentdecision") was determined.

3. Collection methods

3.1. Methodology adopted for collecting the ACCURAT Comparable Corpus

Methods employed by the partners for data collection heavily depend on the type of corpora, degree of comparability and, of course, availability of suitable tools. As one might expect, parallel texts were retrieved automatically by all partners from bilingual or multilingual web sources. Tools used for this purpose vary from custom made scripts that were available to partners to open-source freely available applications.

However, approaches taken for acquiring strongly and weakly comparable corpora are not uniform among partners and for all domains/genres. These corpora were to a great extent selected manually, except from for the domain/genre Software/Wikipedia which was selected automatically due to the predictable structure of Wikipedia and to the inter-linking provided among languages.

Work reported hereafter aimed at researching methods for the automatic acquisition of comparable texts from web sources. The rationale was to build on already existing open-source tools that are suitable for other types of corpora, rather than attempting to build a new harvesting application.

Depending on the approach taken, three general strategies are referred to in the literature: (a) monolingual crawling, (b) bilingual crawling, and (c) topic specific (focused) monolingual harvesting. In monolingual crawling, documents are retrieved for each language separately using either a domain specific or a general monolingual crawler, and they are classified at a later stage. In bilingual crawling, filtering techniques are employed for harvesting parallel data from bilingual/multilingual websites. Finally, topic specific (focused) monolingual crawling attempts to harvest texts belonging to pre-specified domains and narrow topics, and therefore, to directly provide corpora that are by definition at least weakly comparable. The task at hand requires the combination of the afore-mentioned techniques. Among the various candidate tools that have been considered, the following ones seemed the most promising:

- **BootCaT toolkit** [9], a well-known suite of Perl scripts for bootstrapping specialized language corpora from the web.
- **Heritrix**, an open-source, modular web crawler. Implemented in Java, it is an extremely extensible crawling tool providing many configuration settings for achieving best performance, yet it does not support focused crawling.
- **Combine** [10], an open system web crawler-indexer, implemented in Perl. It is based on a combination of a general Web crawler and an automated subject

classifier. The classification is provided by a focus filter using a topic definition in the form of a list of in-topic terms.

- **Bitextor** [11], a free/open-source application for harvesting translation memories from multilingual websites. Bitextor is based on two main assumptions: (a) parallel pages should be under the same domain, and (b) they should have similar html structure.

Within ACCURAT, the selected tools were addaptedto cater for the acquisition of corpora in two language pairs, namely Greek-English and Romanian-Greek. More particularly, parallel corpora in these language pairs were retrieved via **Bitextor**. **BootCaT** was used in order to select monolingual domain-specific corpora to initiate the acquisition of weakly comparable corpora. Seed words semi-automatically extracted from source language texts guided the acquisition of texts in the source language and in specific domains. These were consequently mapped onto their translational equivalents in the target languages in order to serve as seed terms for the selection of candidate weakly comparable texts in the target languages. **Combine** was also used to further supplement the weakly comparable part of corpora. Terms semi-automatically retrieved from the texts in the source language were also coupled with a list of seed URL lists that were manually identified as relevant to the specific domains, and Combine performed limited crawls on selected web sites (e.g. Reuters, BBC, Timesonline etc.). The highest ranking web pages were selected from the result pool and added to the weakly comparable text collection.

After several runs with the above-mentioned tools, manual validation was performed by trained annotators. Retrieved documents were grouped by topic, and annotators had to decide whether they were (a) accurately retrieved as pertaining to the specified domain/genre, and (b) correctly assigned the "weakly comparable" attribute.

As was expected, **BootCaT** provided satisfying results in the domains "Sports" and "Travel" with the vast majority of retrieved texts being positively validated, yet it failed in the identification of genre in documents pertaining to the domains "News", "Software" and "Medicine". **Bitextor** performance heavily depends on how well-formed a web page is (HTML structure) as well as the general structure of the web site. Testing in well structured web sites (e.g. www.setimes.com) provided quite satisfying results in terms of precision and recall, while not so well-formed web sites proved the tool's main weakness in dealing with such environments.

3.2. Visualized crawling environment

Data collection from the web is rarely a well defined job and more often than not corpus linguistics practitioners design their own scripts to provide an answer to an immediate need and as soon as the problem is solved, the scripts are forgotten. We tried to give a more principled solution to reusing the small pieces of useful software and prolonging the life-time of such scripts by the development an environment that incorporates three components: a Flow Graphical Editor which enables the user to easily create and manage workflows, a Script Editor which assists the user in defining the processing units of the workflows and a Windows Service which takes as input the chained scripts generated by the first two components and executes the entire process at a given interval. Thus the environment is not a standalone crawler but a more general program which supports high scalability and integration of modules.

The **Flow Graphical Editor** component allows the user to graphically organize the logic of the application around processing units and decision blocks. The user can alter the global application behavior by adding new blocks or modifying the way the output is being handled. The Flow Graphical Editor allows for the integration of existing modules that produce console output, but the system also supports the usage of other application types. By clicking on a block icon the user can edit its parameters (see Figure 1 and Figure 2).

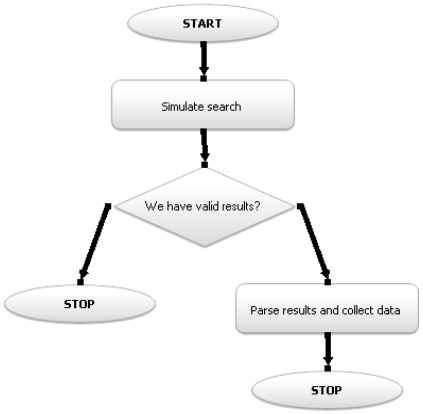

Figure 1. Simple workflow example

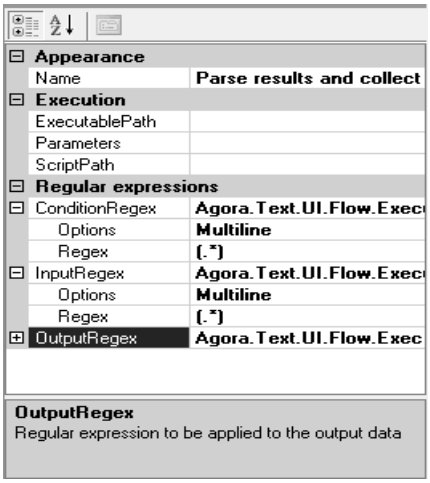

Figure 2. Block property editor

The **Script Editor** enables the creation of processing modules invoked by the active blocks. We use the tools provided by ICSharpCode (http://www.icsharpcode.net/) to enable syntax highlighting and code compilation.

The **Windows service** provides the actual functionality for the built-up processing flow. It will start at a given interval, read the flow diagram and start the execution of active blocks. The user can observe the execution progress at any time and can stop/pause/resume the process.

By means of the environment presented here we created two main processing flows, which can be further connected into a larger one. The first one was a monolingual processing chain incorporating tokenization, tagging, lemmatization and tagging. This ensemble of language tools, called TTL is written in Perl and each of its components is also a web-service [13].

The second application is a web harvester for collecting parallel and strongly comparable corpora from the seed web-pages. We applied the process of collecting strongly comparable documents from the news section of The European Parliament website in 22 languages. Within the months of May and June 2010 195 short articles were harvested, not all of them available in 22 languages. Because we wanted to preserve the structure of the multilingual strongly comparable corpus for the few articles which were not translated in some languages, empty content has been created for the missing languages in these cases.

4. Collected corpora

Using the different approaches described in Section 3, we collected comparable corpora for 9 language pairs: Estonian-English, Latvian-English, Lithuanian-English, Greek-English, Romanian-Greek, Croatian-English, Romanian-English, Romanian-German and Slovenian-English. Almost each language pair corpus consists of approximately one million words for the under-resourced part of corpus (see Table 2). The Romanian-Greek corpus is approximately 130 000 words short of the target one million words which can be explained by the difficulty of collecting appropriate comparable corpora for under-resourced language pairs.

Although all the Accurat languages are under-resourced, the collection process revealed significant differences in relation to availability of parallel and strongly comparable texts. For example, for Balkan languages news can be easily collected from

the SETimes portal, while for languages of the Baltic countries such a resource is not available. Also texts in the domain "International News" or Wikipedia, present a significant disproportion in terms of document size and content among languages.

Table 2. Collected corpora

	Parallel		**Strongly comparable**		**Weakly comparable**		**Total**
	Words	**%**	**Words**	**%**	**Words**	**%**	
ET-EN	101 884	9,48	548 764	51,06	424 022	39,46	**1 074 670**
LV-EN	122 581	11,82	389 127	37,51	525 681	50,67	**1 037 389**
LT-EN	553 747	46,17	261 841	21,83	383 819	32	**1 199 407**
EL-EN	191 843	13,33	294 554	20,47	952 534	66,2	**1 438 931**
RO-EL	282 213	32,62	267 897	30,96	315 108	36,42	**865 218**
HR-EN	418 752	39,51	100 000	9,44	541 085	51,05	**1 059 837**
RO-EN	186 682	6,94	459 458	17,07	2 045 631	76	**2 691 771**
RO-DE	117 281	8,52	449 942	32,67	809 929	58,81	**1 377 152**
SL-EN	462 514	40,17	322 243	27,98	366 759	31,85	**1 151 516**
All language pairs	**2 018 745**	**20,49**	**2 993 826**	**26,01**	**5 823 483**	**53,5**	**11 895 891**

Although the collection process was performed independently in five countries by different project partners, several common resources were identified:

- The SETimes (http://www.setimes.com/) news portal is a source of news about Southeastern Europe in ten languages: Albanian, Bosnian, Bulgarian, Croatian, English, Greek, Macedonian, Romanian, Serbian and Turkish. The portal is updated every day and is an excellent resource of parallel texts for the above mentioned languages.
- The *JRC Acquis* corpus [14] (http://wt.jrc.it/lt/Acquis/) contains selected EU legal texts in all EU official languages, except Irish. It is a widely used source of parallel texts for the legal domain. However, it lacks texts in Croatian (because Croatia is an accession country).
- EMEA corpus [15] (http://urd.let.rug.nl/tiedeman/OPUS/EMEA.php) contains European Medicines Agency documents in 22 languages. The corpus has no texts in Croatian or Slovenian.
- Wikipedia (http://www.wikipedia.org/) is a well known source of comparable texts in more than 270 languages. However, the size of Wikipedia differs from language to language. E.g., for the Accurat languages Wikipedia contains the following number of articles: Croatian – 82 952, Estonian – 76 334, Greek – 53 546, Latvian 28 483, Lithuanian – 110 799, Slovenian – 88 129, Romanian – 146 418 articles (05.07.2010). Also the level of comparability of Wikipedia articles varies a lot.
- European Commission News (http://ec.europa.eu/news) is good resource of strongly comparable texts for EU official languages, especially those which have no other parallel news texts available. The articles describe different topics of interest in the EU, e.g. business, culture, science and technology.
- The software domain is well covered by manuals of open source resources, e.g. Linux; also Web pages of software companies are a good resource for collecting comparable corpora.

5. Conclusions and future work

We collected comparable corpora for 9 language pairs: Estonian-English, Latvian-English, Lithuanian-English, Greek-English, Romanian-Greek, Croatian-English, Romanian-English, Romanian-German and Slovenian-English. Every corpus, except

Romanian-Greek, consists of approximately one million words for each language. Taken together the collected corpora consist of 11,8 million words for Croatian, Estonian, Greek, Latvian, Lithuanian, Romanian and Slovenian.

Currently the corpora are used for two tasks. First, they are being used to develop criteria and automated metrics to determine the kind and degree of comparability of comparable corpora and parallelism of individual documents. Secondly they are serving to evaluate the applicability of existing alignment methods to comparable corpora.

The collected corpora are available for Accurat consortium currently, more texts will be collected through the project lifetime and thus publicly available comparable corpora will be released by the end of the project.

Acknowledgements

The research within the project Accurat leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement n^{o} 248347. Many thanks for collecting corpora data to colleagues in ACCURAT partner organizations: Serge Sharoff from University of Leeds (UK), Gregor Thurmair from Linguatec (Germany), Marko Tadić from University of Zagreb (Croatia) and Boštjan Špetič from Zemanta (Slovenia).

References

[1] A.M. McEnery, R.Z. Xiao, Parallel and comparable corpora: What are they up to? *Incorporating Corpora: Translation and the Linguist. Translating Europe. Multilingual Matters* (2007), Clevedon, UK.

[2] EAGLES, Preliminary recommendations on corpus typology (1996), electronic resource: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

[3] D. Munteanu, D. Marcu. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4) (2005), 477-504.

[4] S. Hewavitharana, S. Vogel, Enhancing a Statistical Machine Translation System by using an Automatically Extracted Parallel Corpus from Comparable Sources. *Proceedings of the Workshop on Comparable Corpora, LREC'08* (2008), 7-10.

[5] A. Eisele, J. Xu. Improving machine translation performance using comparable corpora. *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora. Applications of Parallel and Comparable Corpora in Natural Language Engineering and the Humanities* (2010), 35-39.

[6] I. Skadiņa, A. Vasiļjevs, R. Skadiņš, R. Gaizauskas, D. Tufis, T. Gornostay, Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation. *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora. Applications of Parallel and Comparable Corpora in Natural Language Engineering and the Humanities* (2010), 6-14.

[7] A. Kilgarriff, Comparing Corpora. *International Journal of Corpus Linguistics 6 (1)* (2001), 1-37.

[8] P. Rayson, R. Garside, Comparing corpora using frequency profiling. *Proceedings of the Comparing Corpora Workshop at ACL'00* (2000), 1-6.

[9] M. Baroni, S. Bernardini, .Bootcat: Bootstrapping corpora and terms from the web. *Proceedings of Language Resources and Evaluation Conference LREC'04* (2004).

[10] Ardo, "Combine web crawler," Software package for general and focused Web-crawling (2005), electronic resource: http://combine.it.lth.se/.

[11] M. Gomis, M. Forcada, Combining Content-Based and URL-Based Heuristics to Harvest Aligned Bitexts from Multilingual Sites with Bitextor. *PBML No 93*(2010), 77–86.

[12] J. Cho, H. Garcia-Molina, L. Page, Efficient crawling through URL ordering. *Proceedings of the seventh international conference on World Wide Web* (1998), 161–172.

[13] D. Tufiş, R. Ion, A. Ceauşu, D. Ştefănescu, RACAI's Linguistic Web Services. *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC'08)*, (2008)

[14] R. Steinberger, B.Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, D. Varga, The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation: LREC'06* (2006).

[15] J. Tiedemann, News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. *Recent Advances in Natural Language Processing* vol. V (2009), 237-248.