ACCURAT

Analysis and Evaluation of Comparable Corpora
for Under Resourced Areas of Machine Translation

www.accurat-project.eu

Project no. 248347

Deliverable D3.6
Comparable corpora for under-resourced languages

Version No. 1.0
31/10/2011

Document Information

Deliverable number:	D3.6
Deliverable title:	Comparable corpora for under-resourced languages
Due date of deliverable:	31/10/2011
Actual submission date of deliverable:	31/10/2011
Main Author(s):	USFD
Participants:	USFD, Tilde, CTS, ILSP, FFZG, DFKI, RACAI, ZEMANTA
Internal reviewer:	FFZG
Workpackage:	WP3
Workpackage title:	Methods and techniques for building a comparable corpus from the Web
Workpackage leader:	USFD
Dissemination Level:	**PU**: public
Version:	V1.0
Keywords:	Comparable corpora, under-resourced languages

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
V0.1	16/09/2011	Draft	USFD	First internal draft	Internal
V0.2	14/10/2011	Draft	USFD	Draft	Contributions
V0.3	28/10/2011	Draft	USFD	Submitted version	Contributions
V1.0	31/10/2011	Final	TILDE	Final modifications	Submitted to PO

EXECUTIVE SUMMARY

The document provides a list of the comparable corpora collected by the ACCURAT project partners. The method for their collection is described in D3.4, and the tools employed for their gathering are summarized in D3.5. The collected corpora are stored at the ACCURAT repository and are freely available after contacting the ACCURAT consortium project@tilde.lv.

Table of Contents

Introduction

“Comparable texts are typically considered to be documents that have been produced independently in different languages, but which have the same communicative function as the source text.” (Baer and Koby., 2003)

In translation, comparable corpora are often used alongside bilingual dictionaries, however, manual creation of comparable corpora is very expensive. Much time has been devoted to aligning parallel corpora, e.g., Gale and Church (1993), however, the size of automatically created comparable corpora has been limited. For example, Talvensaari et al. (2006), using an automatic translation system combined with a nearest neighbour method to identify comparable corpora between Finnish and English, found the number of pairs to be very small. Munteanu and Marcu (2005) retrieved a large corpora in Chinese, Arabic and English, however, their maximum entropy classifier required training on a substantial parallel corpus.

1. Summary of the corpora of comparable texts

The following corpora were collected across the ACCURAT designated under resourced languages: Croatian, Estonian, German, Greek, Latvian, Lithuanian, Romanian, Slovenian and English. The size and proportion of the corpora are summarized in Table 1. Each column summarizes the amount of text collected for a particular language pair (in running words for the first language of the pair in the column heading).

The structure of the data described in this document, released as a tgz file (available to download from the ACCURAT repository upon request), is as follows:

- **method** (for example, *wikipedia-anchors*)
    - **source** (for example, *wikipedia*)
        - **language pair** (for example, *de-en)*
            - file alignment (text file, optional)
            - mapping file – filename to source (text file, optional)
            - **language** (for example, *de* or *en*)

This release is marked as Version 1.0, more data will continue to be gathered and will be added to the public release as new versions of the corpus are built.

2. Details of the Collected Corpora

In Table 1 we present the details of the corpora collected using three different techniques:

- **crawl**: Title centered alignment technique applied to crawled news text. Technique described in detail in D3.4 Section 2.1.
- **wiki-anchor**: Translation based (anchor and Google translate) alignment technique applied to Wikipedia. Technique described in detail in D3.4 Section 2.2.1.
- **wiki-topic**: Relative frequency vector based alignment technique applied to Wikipedia. Technique described in detail in D3.4 Section 2.2.2.

Table 1 Details of the corpora collected using three different techniques

Method	Source	Language pair	Document pairs	Docs in 1st lang	Docs in 2nd lang	Words in 1st lang	Words in 2nd lang
Crawl	internet news sites	en-sl	3,642	2,237	1,225	1,043,117	299,700
		en-ro	11,285	5,516	3,363	2,559,497	1,206,191
		en-lv	2,438	1,621	770	839,807	203,173
		en-lt	1,735	1,225	568	579,199	166,856
		en-hr	3,371	2,511	1,142	1,168,540	259,835
		en-et	720	661	254	292,130	37,274
		en-el	6,396	3,786	1,962	1,747,631	456,148
		en-de	29,341	12,719	8,086	5,998,058	3,181,049
wiki anchor	wikipedia	de-en	149,891	149,891	149,891	52,906,987	66,737,429
		el-en	3,668	3,668	3,668	1,094,932	3,989,099
		el-ro	841	841	841	334,304	154,234
		et-en	14,112	14,112	14,112	1,768,028	16,807,432
		hr-en	14,147	14,147	14,147	3,396,259	13,728,551
		lt-en	10,308	10,308	10,308	1,470,226	10,569,091

Method	Source	Language pair	Document pairs	Docs in 1st lang	Docs in 2nd lang	Words in 1st lang	Words in 2nd lang
		lt-lv	1,027	1,027	1,027	264,898	166,414
		lv-en	4,273	4,273	4,273	627,481	5,925,875
		ro-de	16,246	16,246	16,246	953,539	12,608,722
		ro-en	48,880	48,880	48,880	4,827,994	27,197,314
		ro-lt	1,639	1,639	1,639	384,253	280,954
		sl-en	20,351	20,351	20,351	2,648,744	14,998,241
wiki topics	ec.europa.eu1	en-lt	137	138	138	59,706	45,554
		en-lv	137	138	138	59,706	44,879
		en-ro	137	138	138	59,706	60,414
		en-sl	137	138	138	59,706	48,616
		lt-lv	137	138	138	45,554	44,879
		lt-ro	137	138	138	45,554	60,414
		lt-sl	137	138	138	45,554	48,616
		lv-ro	137	138	138	44,879	60,414
		lv-sl	137	138	138	44,879	48,616
		ro-sl	137	138	138	60,414	48,616
	ec.europa.eu2	lt-ro	490	491	491	168,882	259,099
	euronews	en-lt	176	507	178	419,268	113,124
		en-lv	181	507	183	419,268	124,087
		en-ro	194	507	199	419,268	169,580
		en-sl	179	507	181	419,268	145,614
		lt-lv	169	178	183	113,124	124,084
		lt-ro	172	178	199	113,124	169,580
		lt-sl	165	178	181	113,124	145,614

Method	Source	Language pair	Document pairs	Docs in 1st lang	Docs in 2nd lang	Words in 1st lang	Words in 2nd lang
		lv-ro	177	183	199	124,084	169,580
		lv-sl	168	183	181	124,084	145,614
		ro-sl	177	199	181	169,580	145,614
	europarl1	en-lt	194	493	195	230,007	64,496
		en-lv	189	493	190	230,007	62,066
		en-ro	199	493	204	230,007	94,429
		en-sl	177	493	178	230,007	74,913
		lt-lv	182	195	190	64,496	62,066
		lt-ro	182	195	204	64,496	94,429
		lt-sl	170	195	178	64,496	74,913
		lv-ro	180	190	204	62,066	94,429
		lv-sl	166	190	178	62,066	74,913
		ro-sl	173	204	178	94,429	74,913
	europarl2	en-lt	174	502	175	341,902	98,364
		en-lv	132	502	133	341,902	89,293
		en-ro	207	502	217	341,902	170,547
		en-sl	169	502	170	341,902	127,346
		lt-lv	119	175	133	98,364	89,293
		lt-ro	163	175	217	98,364	170,547
		lt-sl	143	175	170	98,364	127,346
		lv-ro	130	133	217	89,293	170,547
		lv-sl	121	133	170	89,293	170,547
		ro-sl	165	217	170	170,547	170,547
	europarl3	en-lt	212	213	213	84,508	52,991
		en-lv	212	213	213	84,508	47,906

Method	Source	Language pair	Document pairs	Docs in 1st lang	Docs in 2nd lang	Words in 1st lang	Words in 2nd lang
		en-ro	212	213	213	84,508	84,513
		en-sl	212	213	213	84,508	57,094
		lt-lv	212	213	213	52,991	47,906
		lt-ro	212	213	213	52,991	84,513
		lt-sl	212	213	213	52,991	57,094
		lv-ro	212	213	213	47,906	84,513
		lv-sl	212	213	213	47,906	57,094
		ro-sl	212	213	213	84,513	57,094
	wikipedia	de-ro	3,954	3,955	3,955	14,588,797	5,110,252
		en-ro	4,228,	4,229	4,229	27,364,056	5,350,424
		lt-ro	5,154	5,155	5,155	4,078,843	6,705,527
	wikitravel	de-ro	1,360	1,361	1,361	1,640,965	401,218
		de-en	1,360	1,361	1,361	1,640,965	7,451,984
		en-ro	1,360	1,361	1,361	7,451,984	401,218

3. References

Baer, B. J. and Koby, G. S. (2003). Beyond the ivory tower: rethinking translation pedagogy. John Benjamins Publishing Company.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. Computational Linguistics, 19(1):75102.

Munteanu, D. S. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. Computational Linguistics, 31(4):477-504.

Talvensaari, T., Laurikkala, J., J□arvelin, K., and Juhola, M. (2006). A study on automatic creation of a comparable document collection in cross-language information retrieval. Journal of Documentation, 62(3):372-387.