ACCURAT

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

www.accurat-project.eu

Project no. 248347

Deliverable D3.2

Test corpora of comparable texts to evaluate comparability metrics

Version No. 1.0
01/04/2010

Document Information

Deliverable number:	D3.2
Deliverable title:	Test corpora of comparable texts to evaluate comparability metrics
Due date of deliverable:	31/03/2010
Actual submission date of deliverable:	01/04/2010
Main Author(s):	Voula Giouli, Radu Ion, Gasper Koren, Madara Mieriņa, Serge Sharoff, Inguna Skadiņa, Marco Tadic, Gregor Thurmair
Participants:	Tilde, ILSP, FFZG, RACAI, Zemanta, LT
Internal reviewer:	USFD
Workpackage:	WP3
Workpackage title:	Methods and techniques for building a comparable corpus from the Web
Workpackage leader:	USFD
Dissemination Level:	PU
Version:	V1.1
Keywords:	Test corpora, comparability metrics

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
0.1	09/03/2010	Draft	Tilde	Draft	First internal draft
0.2	29/03/2010	Draft	All partners	Final draft	For internal review
0.3	29/03/2010	Draft	Tilde	Revised final draft	Revised at Tilde
0.4.	30/03/2010	Draft	USFD	Revised final draft	Revised at USFD
1.0	01/04/2010	V1.0	Tilde	Revised draft	Public deliverable

EXECUTIVE SUMMARY

The document provides an overview of test corpora of comparable texts collected by the Accurat project partners Tilde, ILSP, FFZG, RACAI, Linguatech and Zemanta. The corpora have been collected according to the same principles as the Initial comparable corpora, described in detail in D3.1. The size of the test corpora is about 25 000 running words for each Accurat language. The collected test corpora are stored at the Accurat repository and are freely available after contacting the Accurat consortium project@tilde.lv

Table of Contents

List of tables

1. General principles for collecting test corpora

The main goal of creating test corpora is to provide a collection of comparable texts for the evaluation of the first version of the comparability metrics. Since the initial comparable corpora, described in *D3.1 Initial comparable corpora* are the main resource to create the initial version of the comparability metrics, the test corpora have been collected using the same principles and proportions as the Initial comparable corpora (see Table 1), however texts of test corpora are distinct from initial comparable corpora.

Domain	Genre	% of corpus
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%

Table 1. Domain, genre and coverage of initial comparable corpora and test corpora

The size of the test corpora for major language pairs was set to 25 000 running words (2.5% of the initial comparable corpora). The size requirement for minor language pairs, e.g. Romanian-Greek and Romanian-German, is smaller.

For the German-English language pair narrow domain texts of Accurat specified domains, i.e. software, medicine and automotive, were collected. The narrow domain texts were collected automatically thus the size of corpora exceeds 25 000 running words.

2. Summary of the Test corpora of comparable texts

2.1. Test corpora for under-resourced languages

The test corpora were collected across the Accurat designated under resourced languages: Croatian, Estonian, German, Greek, Latvian, Lithuanian, Romanian, Slovenian and English. The Corpora were also collected for two minor language pairs: Greek-Romanian and Romanian-German. The size and proportions of the test corpora are summarized in Table 2. 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).

Domain	**Genre**	**Comparability**	**ET-EN**	**LV-EN**	**LT-EN**	**EL-EN**	**RO-EL**	**HR-EN**	**RO-EN**	**RO-DE**	**SL-EN**
International news	Newswires	Parallel	977	643	511	587	621		5000		
		strongly comparable	2 628	2215	909	2529		5071			
		weakly comparable	3 023	2568	892	3503					1251
Sports	Newswires	Parallel	318			502	494				
		strongly comparable	1 400			0			2500	10000	
		weakly comparable	1 833	2976	2532	0		3009			2200
Admin	Legal	Parallel	280	686	555	2978	2312	2223	5000		9758
		strongly comparable	1 154	1118	1 308	3332					
		weakly comparable	1 520	1303	1 527	12329					
Travel	Advice	Parallel	307	1025		0		4769			2602
		strongly comparable	1 213	986		0					

		weakly comparable	1 533	1226	2 690	0					
Software	Wikipedia	Parallel				0					
		strongly comparable	3 004	1000	619	0			2500	2500	1266
		weakly comparable				0		3523			
Software	User manuals	Parallel	698	1448		0			5000		
		strongly comparable	1 881	1900	3 869	0					3757
		weakly comparable	2 782	2216	1 503	0		4687			
Medicine	For doctors	Parallel	315	350	7 852	565					2757
		strongly comparable	1 336	1149	1 288	1102					
		weakly comparable	1 565	1298	1 545	0		5615			
Medicine	For patients	Parallel	582	563	21 973	584					
		strongly comparable	1 279	1085		1019			5000	10000	
		weakly comparable	2 061	1312		1501		2622			2867
		Total	**31689**	**27067**	**49573**	**30531**	**3427**	**31519**	**25000**	**22500**	**26458**

Table 2. Size and proportions of test corpora for ACCURAT languages

2.2. Test corpora for narrow domains

The German-English narrow domain test corpora were collected for the software, medicine and automotive domains. The narrow domain test corpora were collected automatically, the size and proportions of the corpora are shown in Table 3 below.

Domain	Genre	Parallel	Strong	Weak
Software	Manuals	532K	734K	69K
Software	Wikipedia	--	2118K	--
Medical	pharma (emea)	22800K	--	--
Automotive	Transmission	47K	659K	765K

Table 3. Size and proportions of test corpora for narrow domains

3. Conclusions

The test corpora were collected in accordance with the same principles as the initial comparable corpora (D3.1). For all the language pairs, except the minor pairs, i.e. Romanian-German and Romanian-Greek, the test corpora exceed the planned 25 000 running words. The Estonian-English test corpus has 6 000 words more than was initially specified, the Latvian-English corpus has 2 000 words more, the Lithuanian-English corpus has 14 000 words more, the Greek-English 5 000 words more, the Croatian-English corpus has 6 000 words more, and the Slovenian-English 1 000 words more.

In total the test corpora contain 247 000 running words. The corpus consists of 34% parallel texts, 33% strongly comparable texts and 33% weakly comparable texts. The domain and genre coverage are shown in the table below:

Domain	Genre	Coverage (%)
International news	Newswires	13
Sports	Newswires	11
Admin	Legal	19
Travel	Advice	7
Software	Wikipedia	6
Software	User manuals	12
Medicine	For doctors	11
Medicine	For patients	21

Table 4. Test corpora domain coverage

The collected test corpora are publicly available from the Accurat repository. Access to the repository is available following registration and notification to Tilde (project@tilde.lv), who will notify partners on access requests to their collections.