TaaS

Terminology as a Service

Project no. 296312

Deliverable D3.6
Initial terminological data for Shared Term Repository

Version No. 1.0

30/11/2013

Document Information

Deliverable number:	D3.6
Deliverable title:	Initial terminological data for Shared Term Repository
Due date of deliverable:	30/11/2013
Actual submission date of deliverable:	30/11/2013
Main Author(s):	Andis Lagzdiņš, Mārcis Pinnis, Artūrs Vasiļevskis (Tilde)
Participants:	Tilde
Internal reviewer:	USFD
Work package:	WP3
Work package title:	TaaS platform and terminology services
Work package leader:	Tilde
Dissemination Level:	PU
Version:	1.0
Keywords:	Terminology collections, terminology service, parallel corpus, comparable corpus

History of Versions

Version	Date	Status	Author (Partner)	Contributions	Description/ Approval Level
0.1	12/11/ 2013	Draft	Tilde	Document fishbone and input to sections	Drafted by Tilde
1.0	30/11/ 2013	Final version	Tilde	Updated document	Revised by Tilde

EXECUTIVE SUMMARY
This is an accompanying document for the deliverable D3.6 Initial terminological data for Shared Term Repository, an integral part of the TaaS system.

Table of Contents

Abbreviations

Table 1 Abbreviations

Abbreviation	Term/definition
API	Application programming interface
CAT	Computer-aided translation
DB	Database
DC	Data category
ETB	EuroTermBank
GUI	Graphical user interface
HTML	Hypertext Markup Language
HTTP	HyperText Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
IATE	InterActive Terminology for Europe
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
MT	Machine Translation
REST	Representational State Transfer
SDB	Statistical DataBase
SMT	Statistical Machine Translation
STR	Shared Term Repository
TaaS	Terminology as a Service
TBX	TermBase eXchange
TDB	Terminology DataBase
TSV	Tab-Separated Values
BiTES	Bilingual Term Extraction System

1. Introduction

This report describes the terminological resources integrated in the TaaS Shared Term Repository (STR), which serve as initial data for the TaaS terminology services showcasing purposes. The following section summarizes the task activities related to this deliverable. The sections thereafter include a brief description of the integration procedure of the initial resources in the TaaS STR followed by a detailed list of resources. Finally, this report also includes a road map for future extensions of the repository and its resource collection.

2. Task description

The initial TaaS platform has to be able to offer a set of pre-loaded terminological resources in the form of bilingual terminology in order to be attractive to new users and to allow showcasing the capabilities of the TaaS terminology services. These initial resources allow human users to test and evaluate the system's functionality and allow users also running terminology extraction projects in a number of domains and language pairs.

3. Resource Repository

The Resource Repository of the TaaS STR consists of three main components:

- STR database – database for the user's gained terminology collections, his/her profile, group and role in it.
- File store – temporal store for term extraction process and for the files uploaded by the user.
- Statistical DataBase (SDB) – data storage and access components for raw aligned terms.

The SDB stores terms, languages, term canonical forms, morpho-syntactic information, information about sources and term contexts. The terms in SDB are grouped by canonical forms in order to provide a list of unique translation candidates for users' extraction projects.

The SDB is a constantly growing resource as the Bilingual Term Extraction System (BiTES) workflows are iteratively providing new bilingual term collections in the form of TSV (tab-separated-values) documents. The SDB is then iteratively updated with the new term collections.

Furthermore, each resource can be described by arbitrary metadata stored in the repository database. More details about the implementation and functionality of the Term Repository can be found in deliverable D3.2 Shared Term Repository. More detailed information about file processing implementation and functionality can be found in the deliverables D3.3 Facility for term candidate acquisition and term glossary creation and D3.1 Technical requirement analysis and specification.

4. Overview of Initial Resources Available in the TaaS platform

Table 2 lists all bilingual terminological data collections integrated in the TaaS SDB. The table shows the unique number of translation equivalents of term candidate surface forms extracted

from comparable resources found in the Web as well as publicly available parallel corpora. Each term collection within the SDB is identified with a unique collection URI.

Table 2 Initial terminological resource statistics

Collection URI	Term pair count
http://taas.eurotermbank.com/LetsMT-Moses/DGT-TM-phrase-table	524,340
http://taas.eurotermbank.com/LetsMT-Moses/Tilde-IT-phrase-table	424,224
http://taas.eurotermbank.com/MP_Aligner/Tilde-FMC-ECONOMY-Corpus	8,012
http://taas.eurotermbank.com/MP_Aligner/Tilde-FMC-ENERGY-Corpus	21,145
http://taas.eurotermbank.com/MP_Aligner/Tilde-FMC-IT_NEWS-Corpus	582
http://taas.eurotermbank.com/MP_Aligner/Tilde-FMC-IT-Corpus	22,125
http://taas.eurotermbank.com/MP_Aligner/Tilde-FMC-LAW-Corpus	17,682
http://taas.eurotermbank.com/MP_Aligner/Tilde-FMC-MECH_ENG-Corpus	5,785
http://taas.eurotermbank.com/MP_Aligner/Tilde-FMC-MEDICINE-Corpus	22,128
http://taas.eurotermbank.com/TILDE-MPAligner/ACCURAT-Disaster-Corpus	83,216
http://taas.eurotermbank.com/TILDE-MPAligner/ACCURAT-ITLocalisation-Corpus	14,722
http://taas.eurotermbank.com/TILDE-MPAligner/ACCURAT-Political-Corpus	76,493
http://taas.eurotermbank.com/TILDE-MPAligner/ACCURAT-RenewableEnergy-Corpus	68,575
http://taas.eurotermbank.com/TILDE-MPAligner/ACCURAT-Sport-Corpus	33,212
http://taas.eurotermbank.com/TILDE-MPAligner/ACCURAT-Technology-Corpus	57,313
http://taas.eurotermbank.com/TILDE-MPAligner/MP-Medicine-Corpus	32,993
http://taas.eurotermbank.com/TILDE-MPAligner/TILDE-Automotive-Corpus	25,785
http://taas.eurotermbank.com/TILDE-MPAligner/TILDE-IT-Corpus	9,493
http://taas.eurotermbank.com/TILDE-MPAligner/USFD-News-Week12-Automotive-Corpus	31
http://taas.eurotermbank.com/TILDE-MPAligner/USFD-News-Week12-IT-Corpus	137
http://taas.eurotermbank.com/TILDE-MPAligner/USFD-News-Week3-Automotive-Corpus	63
http://taas.eurotermbank.com/TILDE-MPAligner/USFD-News-Week3-IT-Corpus	299
http://taas.eurotermbank.com/TILDE-MPAligner/USFD-News-Week4-Automotive-Corpus	38
http://taas.eurotermbank.com/TILDE-MPAligner/USFD-News-Week4-IT-Corpus	206
http://taas.eurotermbank.com/TILDE-MPAligner/USFD-TAAS-Wiki-Corpus	68,402
Total	1517001

The collection URIs' allow us to uniquely identify the corpora, from which the bilingual terminology has been extracted, what bilingual terminology alignment tools have been used and provide us with a version number of the particular bilingual term collection. The corpora from which the bilingual terminology has been extracted are summarised in Table 3.

Table 3 Parallel and comparable corpora description

Corpora	Additional information
DGT-TM-phrase-table	The parallel corpora DGT-Translation Memory[1] are publicly available parallel resources developed by the European Commission's Directorate-General for Translation. Within the TaaS project we use the parallel resources in order to extract terminological phrases and their translation equivalents from the phrase tables produced by the Moses statistical machine translation platform.
Tilde-IT-phrase-table	The Tilde IT parallel corpus is a proprietary parallel corpus owned by one of the TaaS consortium members – Tilde. The corpus has been used similarly to the DGT-Translation Memory corpora in order to extract bilingual terminology from the parallel corpus.
ACCURAT-Disaster-Corpus ACCURAT-ITLocalisation-Corpus ACCURAT-Political-Corpus ACCURAT-RenewableEnergy-Corpus ACCURAT-Sport-Corpus ACCURAT-Technology-Corpus	The comparable corpora have been produced in the ACCURAT project (www.accurat-project.eu). More details on the corpora statistics and collection procedures can be found in the Deliverable D2.4 "Aligned comparable corpora" of the ACCURAT project (http://www.accurat-project.eu/uploads/Deliverables/D2.4_Aligned comparable corpora_v1.0_final.pdf).
MP-Medicine-Corpus	The comparable corpora in the Medicine and Pharmacy domains (TaaS-2200) have been created by Pinnis (2013)[2] in order to evaluate the bilingual term alignment tool MPAligner. More details on the corpora can be found in the scientific paper by Pinnis (2013) published in the RANLP 2013 conference proceedings.
TILDE-Automotive-Corpus	The comparable corpus in the Mechanical Engineering domain (TaaS-1504) was collected for purposes of evaluation of terminology integration scenarios in statistical machine translation. Further details on the corpus can be found in the paper by Pinnis and Skadiņš (2012)[3].

[1] Further information on the DGT-Translation Memory can be found at: http://ipsc.jrc.ec.europa.eu/index.php?id=197.

[2] Pinnis, M. (2013). Context Independent Term Mapper for European Languages. In Proceedings of Recent Advances in Natural Language Processing (RANLP 2013) (pp. 562–570). Hissar, Bulgaria.

[3] Pinnis, M., & Skadiņš, R. (2012). MT Adaptation for Under-Resourced Domains – What Works and What Not. In A. Tavast, K. Muischnek, & M. Koit (Eds.), *Human Language Technologies – The Baltic Perspective* -

Corpora	Additional information
TILDE-IT-Corpus	The comparable corpus in the Information and Communication Technology domains (TaaS-1501) is a proprietary corpus owned by Tilde. This corpus is not widely available.
Tilde-FMC-ECONOMY-Corpus Tilde-FMC-IT-Corpus Tilde-FMC-IT_NEWS-Corpus Tilde-FMC-ECONOMY-Corpus Tilde-FMC-ENERGY-Corpus Tilde-FMC-MEDICINE-Corpus Tilde-FMC-LAW-Corpus Tilde-FMC-MECH_ENG-Corpus	The comparable corpora from six different domains (Law – TaaS 0200, Economics – TaaS-0300, Medicine and Pharmacy – TaaS-2200, Energy – TaaS-1200, Information and Communication Technology – TaaS-1501, and Mechanical Engineering – TaaS-1504) has been collected within the TaaS project with the TaaS project's BiTES, which will be described in more details in the Deliverable D2.4 "Final Bilingual Term Extraction System". The corpus has been collected using the Focussed Monolingual Crawler (FMC) from Web domains containing articles from the six TaaS domains. Further information on the comparable corpora will be described in the Deliverable D2.2 "Parallel and Comparable Data for Term Extraction".
USFD-News-Week12-Automotive-Corpus USFD-News-Week12-IT-Corpus USFD-News-Week3-Automotive-Corpus USFD-News-Week3-IT-Corpus USFD-News-Week4-Automotive-Corpus USFD-News-Week4-IT-Corpus	The comparable news corpora from RSS feeds have been collected within the TaaS project with the TaaS project's BiTES, which will be described in more details in the Deliverable D2.4 "Final Bilingual Term Extraction System". Further information on the comparable RSS news feed corpora will be described in the Deliverable D2.2 "Parallel and Comparable Data for Term Extraction".
USFD-TAAS-Wiki-Corpus	The comparable Wikipedia corpus has been collected within the TaaS project with the BiTES. Further details of the corpus will be described in the Deliverable D2.2 "Parallel and Comparable Data for Term Extraction".

The bilingual terminology extraction tools and methods will be described in more details in the Deliverable D2.4 "Final Bilingual Term Extraction System".

The bilingual terminology with respect to language pairs and TaaS domains is provided in Table 4.

Proceedings of the Fifth International Conference Baltic HLT 2012 (Vol. 247, pp. 176–184). Tartu, Estonia: IOS Press. doi:10.3233/978-1-61499-133-5-176

Table 4 Count of Term Pairs grouped by domains and language pairs

Domain	Language pair	Count of term pairs
Domain not specified	*Total*	592,742
	EN-LV	524,340
	LV-EN	68,402
TaaS-0102 Politics	*Total in domain*	76,493
	LT-EN	8,001
	LV-EN	19,252
	LV-LT	49,240
TaaS-1100 Environment	*Total in domain*	83,216
	LT-EN	7,703
	LV-EN	36,794
	LV-LT	38,719
TaaS-1200 Energy	*Total in domain*	89,720
	BG-EN	889
	CS-EN	790
	DA-EN	540
	DE-EN	490
	EL-EN	1,116
	ES-EN	476
	ET-EN	486
	FI-EN	202
	FR-EN	261
	HR-EN	7,273
	HU-EN	928
	IT-EN	281
	LT-EN	9,345
	LV-EN	32,920
	LV-LT	28,570
	NL-EN	339
	PL-EN	929
	PT-EN	161
	RO-EN	881

Domain	Language pair	Count of term pairs
	RU-EN	1,443
	SK-EN	614
	SL-EN	342
	SV-EN	444
TaaS-1500 Industries and technology	*Total in domain*	57,313
	LT-EN	4,203
	LV-EN	20,480
	LV-LT	32,630
TaaS-1501 Information and communication technology	*Total in domain*	471,779
	BG-EN	716
	CS-EN	1,005
	DA-EN	706
	DE-EN	2,964
	EL-EN	566
	EN-LV	424,224
	ES-EN	407
	ET-EN	455
	FI-EN	826
	FR-EN	761
	HR-EN	3,815
	HU-EN	896
	IT-EN	892
	LT-EN	754
	LV-EN	25,463
	NL-EN	1,091
	PL-EN	747
	PT-EN	798
	RO-EN	1,231
	RU-EN	1,535
	SK-EN	631
	SL-EN	557
	SV-EN	739

Domain	Language pair	Count of term pairs
TaaS-1504 Mechanical engineering	*Total in domain*	31,702
	BG-EN	110
	CS-EN	182
	DA-EN	64
	DE-EN	174
	EL-EN	115
	ES-EN	13
	ET-EN	35
	FI-EN	34
	FR-EN	27
	HR-EN	3,445
	HU-EN	28
	IT-EN	58
	LT-EN	304
	LV-EN	25,989
	NL-EN	59
	PL-EN	126
	PT-EN	43
	RO-EN	83
	RU-EN	251
	SK-EN	358
	SL-EN	112
	SV-EN	92
TaaS-0200 Law	*Total in domain*	17,682
	BG-EN	1,064
	CS-EN	1,797
	DA-EN	765
	DE-EN	459
	EL-EN	306
	ES-EN	6
	ET-EN	1,035

Domain	Language pair	Count of term pairs
	FI-EN	510
	FR-EN	303
	HR-EN	2,850
	HU-EN	574
	IT-EN	401
	LT-EN	830
	LV-EN	1,311
	NL-EN	646
	PL-EN	503
	PT-EN	527
	RO-EN	384
	RU-EN	1,686
	SK-EN	1,378
	SL-EN	93
	SV-EN	254
TaaS-0300 Economics	*Total in domain*	8,021
	BG-EN	753
	CS-EN	739
	DA-EN	357
	DE-EN	339
	EL-EN	95
	ES-EN	18
	ET-EN	269
	FI-EN	55
	FR-EN	183
	HR-EN	848
	HU-EN	138
	IT-EN	104
	LT-EN	767
	LV-EN	532
	NL-EN	129
	PL-EN	908

Domain	Language pair	Count of term pairs
	PT-EN	131
	RO-EN	354
	RU-EN	764
	SK-EN	348
	SL-EN	26
	SV-EN	164
TaaS-2007 Sports	*Total in domain*	33,212
	LT-EN	2,585
	LV-EN	10,676
	LV-LT	19,951
TaaS-2200 Medicine and pharmacy	*Total in domain*	55,121
	BG-EN	1,473
	CS-EN	719
	DA-EN	659
	DE-EN	3,861
	EL-EN	690
	ES-EN	126
	ET-EN	684
	FI-EN	390
	FR-EN	412
	HR-EN	6,204
	HU-EN	871
	IT-EN	424
	LT-EN	2,212
	LV-DE	4,061
	LV-EN	26,873
	NL-EN	321
	PL-EN	1,163
	PT-EN	654
	RO-EN	338
	RU-EN	873
	SK-EN	1,045

Domain	Language pair	Count of term pairs
	SL-EN	755
	SV-EN	313
Grand Total		1,517,001

5. Next steps

Collection of data will be continued during the lifecycle of the project. At the same time, the collected data will be uploaded to the STR and will be available for use to TaaS users.

6. Conclusions

The TaaS STR is implemented and populated with initial terminological resources. The data are available for TaaS users through the TaaS workflow of the creation of bilingual terminology collections from user-uploaded monolingual documents.

List of tables