LT_OBSERVATORY – OBSERVATORY FOR LR AND MT IN EUROPE

Acronym: LT_OBSERVATORY

COORDINATION AND SUPPORT ACTION

INFORMATION AND COMMUNICATION TECHNOLOGIES

D1.3 Best Practice Guide on LRs for Automated MT

GRANT AGREEMENT	644583
DELIVERABLE NUMBER	D1.3
DELIVERABLE TITLE	Best Practice Guide on LRs for Automated MT
DUE DATE OF DELIVERABLE	31/12/2016
ACTUAL SUBMISSION DATE	23/12/2016
START DATE OF THE PROJECT	01/01/2015
DURATION	24 months
ORGANIZATION NAME RESPONSIBLE FOR THIS DELIVERABLE	CLARIN ERIC

DISSEMINATION LEVEL		
PU	Public	☒
PP	Restricted to other programme participants (including the Commission Services)	□
RE	Restricted to a group specified by the consortium (including the Commission Services)	□
CO	Confidential, only for members of the consortium (including the Commission Services)	□

TYPE		
R	Document, report	☒
DEM	Demonstrator, pilot, prototype	□
DEC	websites, patent fillings, prototype	□
OTHER		□

LT_OBSERVATORY – OBSERVATORY FOR LR AND MT IN EUROPE

LT_OBSERVATORY

TABLE OF CONTENTS

This project is co-funded by the European Union

DOCUMENT INFO

AUTHORS

Name	Company	E-mail
Claus Povlsen	UCPH	cpovlsen@hum.ku.dk
Hanne Fersøe	UCPH	hannef@hum.ku.dk
Lina Henriksen	UCPH	linah@hum.ku.dk
Vesna Lušicky	UNIVIE	vesna.lusicky@univie.ac.at
Bente Maegaard	UCPH	bmaegaard@hum.ku.dk
Sussi Olsen	UCPH	saolsen@hum.ku.dk

DOCUMENT CONTROL

Document version	Date	Change
D1.3.1	27/04/2016	First internal draft
D1.3.2	02/05/2016	Changes by partners
D1.3 3	04/05/2016	Final version by the consortium to be submitted to the EC.
D1.3.4	20/09/2016	Changes by partners
D1.3 v1	31/12/2016	Final version submitted

This project is co-funded by the European Union

1. SUMMARY

Background information: The European Digital Single Market, one of the main goals of Europe 2020, is still fragmented due to language barriers. European society is multilingual; the diversity of its cultural heritage is an opportunity, but hampers transborder eCommerce, social communication and exchange of (cultural) content. Languages without sufficient technological support will become marginalised. These barriers must be overcome by language technology (LT) like Machine Translation (MT) solutions, a need recognized by the future Connecting Europe Facility (CEF).

To support these endeavours to reach an online EU internal market free of language barriers, it is necessary to join, benchmark the quality and facilitate the access to language resources.

Executive Summary:

These guidelines describe how to improve existing LRs by adding missing information in particular metadata that support their accessibility and usability, as well as on creating new LRs, be it bilingual corpora or terminology.

The previous reports on Language Resources (LRs) have focussed on existing LRs and catalogues, as well as on how to establish a one-stop shop, the LTO catalogue, where users can find information about relevant LRs in one single place. However, users do not always find exactly those LRs that they need. As a consequence of this observation, descriptions on best practice are given on open source tools that enable users to automatically acquire in-domain parallel data by finding domain relevant websites containing bilingual documents, evaluating the documents found, removing duplicates and excluding boilerplate elements, and how the parallel domain-specific data are sentence aligned. Furthermore, guidelines are given on how to extract domain specific terminology automatically from the internet. Finally the report has a chapter on IPR issues.

The report is based on existing knowledge and ongoing work in the field. For details see the references. We want to acknowledge in particular the PANACEA project. Additionally, we want to acknowledge direct feedback from language professional that use LRs in their daily work.

Comment:

The present version of the report is the final version of D1.3

2. VALUE-ADDING

Many LRs in existing repositories present some challenges in an operational context; even if these LRs are theoretically valuable they are also practically more or less unusable. The reason is that resources often lack basic information about e.g. copyright issues, domain, owner or resource format making it impossible, or at best difficult, to identify the resources in the first place. If they are indeed identified then to select the best suited for a particular purpose is equally difficult as a resources preview is hardly ever foreseen. Therefore valorisation of LRs through optimization of existing metadata, and sometimes addition of new metadata, will be an invaluable aid to MT developers - as this means that the developers will actually be able to identify and select the resources they need. It should be noted in this context that even if valorisation of existing resources is indeed a possibility and is the topic of this section, it is preferable to adhere to best practice suggestions already in connection with creation of language resources and the matching metadata.

In this section we will present “best practice” in relation to optimization of existing and addition of new essential metadata for LRs to be used for MT purposes. We will also present some guidelines for collection of information about best practice in relation to resource value-adding – should anyone wish to collect this type of information in some special MT context not covered by this section.

2.1 ADDITION OF METADATA

The list of metadata recommended for LRs (from D1.2) to be used for MT purposes is as follows:

Title	Name of the language resource, e.g. *Estonian-Latvian parallel corpus of building product texts*
Resource type	Some examples could be: corpus, classifications scheme, lexical resource, terminology resource, tool
Creator	Creator of the language resource**.** The creator can be an institution as well as a group or a person, e.g. *European Commission* or *Luisa Coheur*
Language(s)	The languages of the resource. The languages must be stated as ISO codes.
Availability	This data category should give basic information about the price of the resource.
Modality	Examples: text, speech, video.
URL	The URL of the language resource.
Domain	Open class containing both text type e.g. *press release, annual report* and domain e.g. *automotive, pharmaceuticals*.

This project is co-funded by the European Union

Format	Format information, e.g. *plain text, RTF*.
Size	The size of the language resource and the size unit, e.g. *300,000 words* or *10,000 sentences*.
Production date	The date when the resource was created. Production date can be stated as date or year. Some resources do not have a production data but perhaps only the date of inclusion in the catalogue. A comment about this can be made in the comment field.
Comment	Free text field where the user can provide any information on the resource that is considered of relevance that does not fit into any of the other metadata fields, e.g. alignment methods or annotations.
Description	This field adds a lot to the user friendliness of a resource. Many resources have good detailed descriptions that give the user a clear understanding of the resource. Unfortunately some resources have very little description apart from language pair and resource type.
Tags	This field is used for keyword-like values, topics that concern the resource in question. Sometimes the field can be used for domain like information that does not qualify for inclusion in the domain field.
Contact person	Can be an email address of the organisation that developed the resource, mail address of the distributor of the resource.
Format description	The specification of the format of a resource is often very poor. It should however be stressed that users find this information essential.

This is a relatively long list in comparison with the metadata that are usually included in a resource and it will require some effort to collect the information. But we have found out that the information does often exist somewhere: Sometimes in various documents describing the resource or simply on the website hosting the resource. Still, it should be noted that it is not always possible to find the information you really want, but it may be possible to identify information that is closely related to the desired information. We therefore recommend inclusion of metadata information as e.g. *Tags, Comment* and *Description* as these fields can accommodate information that is really a substitute for the desired information. As an example can be mentioned that some resources do not include domain information and this is a serious flaw in a language resource for MT purposes. But the information can sometimes be derived in some form from resource descriptions - and even if it is not a real domain name and cannot figure in the *Domain* field, the information can be included in one of the fields: *Comment, Tags* or *Description*.

2.2 FOLLOWING A METADATA STANDARD

We recommend that the list of metadata corresponds to a standard and a good example of such a standard is the Dublin Core (DC) metadata set for resources, even if some adjustments are recommended:

- *Production date* instead of the DC category called *Date* in order to specify that only the production date of the resource is requested and not the *Change date* or some other date.

Inclusion of categories that are not part of the DC set:

- *Size* should be included though it is not part of the DC set as resource size is an important factor in MT system creation
- *Comment* should be included as additional information inconsistent with the other data categories is sometimes available and useful, e.g. in relation to the quality.
- *Modality* should be included as written as well as spoken resources are expected to be included in the LTO resource.
- *Availability* should be included as parallel resources often come at a price and sometimes even at a rather expensive price
- *Tags* should be included as it will often prove useful in to accommodate useful data and in connection with data search

DC data categories that are often not necessary to include:

- *Contributor, Coverage, Publisher, Relation, Rights, Source* as these are often dispensable in a context where an important aim is to develop a simple, concise and easy-to-use resource collection. Some of these information types are however at least partly covered in some of the other metadata.

2.3 RECOMMENDATIONS FOR VALUE-ADDING IN ANOTHER CONTEXT

If someone wishes to perform value-adding of LRs in some MT context not covered in this document, we also have some recommendations in relation to collection of best practice information. We recommend sending resources with initial valorisation to potential LR users, vendors and buyers and letting them examine and test resources in relation to their own work context and requirements. Later these people should be invited to a more formal event where they comment and give their best advice in relation to the valorisation. This procedure can be repeated – either with the same participants or with new participants. It can also be recommended to include a user-feedback system in a resource collection in order to collect information from users and thus continuously improve the resources. We also found out that a mere “rating system”, e.g. one to four stars, is not considered adequate or useful, as such rating may differ depending on purpose and use of the LRs.

3. GETTING STATISTICAL MACHINE TRANSLATION RELEVANT LANGUAGE DATA AND TOOLS ON THE WEB – BEST PRACTICE

In this context, it is assumed that the SMT software framework is the Moses open source package (Koehn et Schroeder, 2007) and that the general domain SMT system is based on parallel data from the Europarl corpus. More investigations point to the fact that adaptation of general-domain systems by inclusion of small amounts of in-domain parallel data can improve significantly the translation quality of domain-specific SMT systems (Pecina et al., 2012, Mastropavlos & Papavassiliou, 2011, and Wu & Zong, 2008). Bearing in mind the huge amounts of documents available online, it would be obvious to define and implement methods that identify and acquire domain specific bilingual corpora from the Web. Creating such corpora is relevant in particular when targeting less resource covered languages.

In broad terms, acquisition of in-domain parallel data can be divided into three phases. The first step consists of a focused search for and subsequently ranking of domain relevant websites. The links found at these websites are then regarded as candidate URL seeds with respect to identifying bilingual documents. After evaluation of the candidate documents detected, the next process consists of removal of duplicates and exclusion of boilerplate elements. Finally, the parallel domain-specific corpus is sentence aligned. The following section describes the stepwise approach to getting domain specific parallel data via web-crawling approaches.

3.1 DOMAIN RELEVANT AND MULTILINGUAL WEBSITES

Performance of SMT systems is dependent of how well the training data correlates with the documents that are translated regarding genre, style and in particularly domain-specific data. In order to meet the latter criterion, it would be obvious to collect domain-relevant training data by exploiting web-crawling approaches.

Before starting the actual web-crawling process, it is important that one spends some time on making a list of terms that, as accurately as possible, reflects and represents the domain area within which the SMT system is supposed to translate. The seed terms on this list will form the search pattern used to find relevant documents on the Web.

One possible source with respect to designing such a term list is Wikipedia. This approach is implemented as a facility in the Sketch Engine application[1]. Another approach would be to find and to extract domain relevant terms from the Eurovoc glossary.

The next task in the workflow is to prioritize between the search results, i.e. the collection of website links. Various ranking strategies have been used in connection with finding the most suitable links to follow. Such as modelling domain relevant pages by evolving neural networks or use Best-First algorithm that prioritizes between

[1] Sketch Engine is a commercial piece of software, https://www.sketchengine.co.uk/

This project is co-funded by the European Union

webpages assigned with various scores. Hybrids of these two methods can also be found. Implementation of these approaches, however are characterized by being research prototypes not available as open source.

The list of domain relevant URLs collected during the monolingual web-crawling process will not only constitute important links to acquiring monolingual data, it will also often form a starting point for identifying websites that can be regarded as candidates for bilingual documents.

Identification of domain-specific multilingual websites adds an additional task to the workflow. Finding parallel text on the Web consists of two general steps. After having located pages that may have parallel translations, the next task is to check the quality of the pages collected, i.e. finding out whether the candidate pairs are actually translations. One important element in the process of distinguishing between true and false positives is comparison of HTML structures in the parallel documents in question. The Web-Mining Architecture STRAND (Resnik & Smith, 2003) has as its main goal to identify pairs of Web pages that are mutual translations. The software package is not available and corpora acquired by STRAND on the Web are subject to copyright restrictions. Databases of URL pairs acquired by STRAND, however, can be downloaded for personal use,
cf. http://www.umiacs.umd.edu/~resnik/strand/. Another system that mines parallel documents from multilingual websites is Bitextor (Esplà-Gomes & Forcada, 2010). This application is based on a quantitative approach by looking at file size, text length, tag structures etc. Bitextor is an open source application and can be downloaded from, https://sourceforge.net/projects/bitextor/. A similar approach is the ILSP Web Crawler that extracts pairs of documents on the Web that are likely translations of each other cf. (Mastropavlos & Papavassiliou, 2011). An introduction to this research application can be found at, http://nlp.ilsp.gr/redmine/projects/ilsp-fc/wiki/Introduction

3.2 CLEANING UP AND PREPARING DOCUMENTS

Both monolingual and multilingual Web collections contain duplicates and near-duplicates that need to be filtered out. SpotSigs (Theobald et al., 2008) is an application that can detect duplicate and near-duplicate documents in large Web collections, https://sourceforge.net/projects/spotsigs/. When looking for automatic tools that can be adapted to one's specific needs, it would be advisable to look at the Nutch framework that offers a wide range of tools helpful in connection with cleaning collections of data from the web, cf. http://nutch.apache.org/.

Removal of boilerplate elements is another important element in the cleaning process. webpages often contain navigation links, commercials, and disclaimers. Such HTML source elements are irrelevant in connection with establishing parallel corpora for linguistic purposes. Having made a domain specific Web collection via use of Sketch Engine, it becomes clear when you look at the somewhat poor results that boilerplate removal is both relevant and difficult to automatize. An automatic tool to remove boilerplate elements that seems to perform better than Sketch Engine is the Boilerpipe tool that can be found here, https://github.com/kohlschutter/boilerpipe.

3.3 SENTENCE ALIGNMENT

After having identified the relevant domain documents and subsequently having cleaned and prepared the parallel documents for further processing, the next step consists of sentence splitting and tokenization. At http://www.statmt.org/europarl/, Europarl tools that treat sentence splitting and tokenization can be downloaded. The final step in the pipeline of making parallel data qualified as training data for an SMT system, is to secure that the sentences extracted are aligned with the highest quality possible.

High alignment quality is crucial with respect to achieving good SMT translation results. Bearing in mind that manual checking and evaluation of automatic alignment are very time consuming, it is decisive that the alignments of parallel documents identified are of highest possible quality.

In a study (Pecina et al., 2012) three available and state-of-the art sentence aligners were evaluated, i.e. Hunalign (https://github.com/danielvarga/hunalign) (Varga et al., 2005), GMA[2] and Bilinguel Sentence Aligner, BSA (Moore, 2002). Performance of sentence alignment is usually evaluated by measuring the effect in terms of improved translation quality that is achieved when the aligned corpus is added to the training data. In (Pecina et al., 2012), the evaluation goal was to find out how well the aligners performed in an industrial scenario so it is relevant to compare the efficiency of the sentence aligners in terms of execution time and use of memory. The evaluation results revealed that the Hunalign sentence aligner regarding execution time outperformed the GMA and BSA aligners, while with respect to use of memory, the Hunalign aligner performed badly compared to the other two sentence aligners[3].

An alternative and much more resource demanding method is to evaluate the alignment results intrinsically, i.e. to measure how well the various sentence aligners can handle various bead types i.e. alignment of 1:2, 2:1 etc. In Appendix A, you will find a review of the Hunaligner and the BSAligner with a focus on how they are installed and how well they perform when confronted with parallel data consisting of various bead types.

3.4 LIST OF TOOLS

In the table below, the tools mentioned above in sections 3.1,3.2, and 3.3. are listed. It would be relevant to add that this list is not to be regarded as exhaustive.

[2] htttp://nlp.cs.nyu.edu/GMA

[3] It should be added that the Hunalign application offers preprocessing software that chunks the input data into smaller pieces, about 5.000 sentences.

Name	Functionaliy	Supported file formats	Availability	Available from:
STRAND	Identification of mutual translations on the Web. Only access to STRAND bilingual databases	The STRAND database format	GNU public license (GPL)	http://www.umiacs.umd.edu/~resnik/strand/
Bitextor	Mining of parallel documents on the Web	HTML,XHTML, XML	GNU public license (GPL)	https://sourceforge.net/projects/bitextor/
ILSP Focused Crawler	Research prototype for acquiring domain-specific monolingual and bilingual corpora	HTML, XML	GNU GPL, v. 3.0 license	http://nlp.ilsp.gr/redmine/projects/ilsp-fc/wiki/Introduction
Nutch framework	Web crawler	No information	Apache License, version 2.0	http://nutch.apache.org/
SpotSigs	Filtering near duplicates	HTML, XML	GNU public license (GPL)	https://sourceforge.net/projects/spotsigs/
Boilerpipe	Detection of Boilerplates etc	HTML	Apache License version 2.0	https://github.com/kohlschutter/boilerpipe
HunAlign	Sentence aligner	TXT	GNU LGPLv3	https://github.com/danielvarga/hunalign
Geometric Mapping and Alignment (GMA)	Sentence aligner	TXT	GNU public license (GPL)	http://nlp.cs.nyu.edu/GMA/
Bilingual Sentence Aligner (BSA)	Sentence aligner	TXT	Microsoft Research end user license agreement (MSR-EULA)	https://www.microsoft.com/en-us/download/details.aspx?id=52608

This project is co-funded by the European Union

3.5 SUMMING-UP

Seen from the LT_Observatory point of view, it would be ideal if all the software needed to generate high quality in-domain and sentence aligned corpora, was available at one single website or portal. Unfortunately, this is by far not the case. Users that want to extract and download parallel data for SMT systems have to search for and pinpoint websites having parallel documents. The most challenging task, seen from a user's point of view is, however, the lack of open source software that is required to establish the processing pipeline that takes you from the look-up via a web search engine until you possess high quality SMT training data. In other words, many useful research tools are developed often as stand-alone applications requiring that the users themselves are left to hard code the software that integrates the applications into one workflow. Having said that, many useful open source tools do exist that can help you through the pipeline steps. Examples of this are the bilingual webcrawler found at http://nlp.ilsp.gr/soaplab2-axis that finds the links within websites and the Hunalign sentence aligner http://nlp.ilsp.gr/soaplab2-axis. Again one should bear in mind that even though these tasks can be conducted automatically, human involvement is still required. For instance, generation of domain specific multilingual seed URL lists requires human effort as well as the quality evaluation of the outputs from the Hunalign sentence aligner need to be checked manually.

4. TERMINOLOGICAL RESOURCES

Terminological resources may be used in a variety of settings: customization of a machine translation system, supporting the computer-assisted translation process (multilingual terminology management), search engine optimization (SEO), etc. Lexicons and terminologies play an important role in any machine translation system, regardless of the principle on which the machine translation tool is based, resulting in hybrid MT.

As terminological resources, we usually understand structured terminological resources (termbases) and glossaries, although the term 'glossary' is often used as a generic reference to any terminological resource. A glossary is usually a mono- or bilingual collection of terms and definitions that are relevant to a particular domain. Termbases can be mono-, bi- or multilingual. They are usually concept-oriented and structured. Depending on the main purpose and application of a termbase, the entries in a termbase may include a wide range of additional information on terms, languages and concepts they designate. Terminological collections may include single word terms and multi-word terms, and their variations.

In the translation and localization industry clients require correct and accurate use of specific terminology, often the companies' in-house terminology. They may provide their own terminology collections that have to be strictly used during translation to ensure correct and consistent usage of terminology. However, in projects where the client-supplied terminology collections are not readily available, but the use of specific in-house terminology is still required, the terminology firstly needs to be extracted from documents provided by the client.

Terminology collections may contain equivalents that are rated as unlikely by the SMT system models. If such an SMT system is integrated in translation service workflows, it is not possible to ensure high quality of terminology (consistency, correctness) in the SMT suggestions. Training data can contain contradicting terminology, corporate specific synonyms or vendor-biased terminology. For this reason effective adaptation of SMT systems that can profit from customized terminology collections are necessary. In terms of terminology integration, it has been shown that the introduction of a bilingual terminology in the translation model can considerably improve translation quality of an out-of-domain system (Pinnis and Skadins 2012, Arcan et al., 2014). Some anecdotal data show that in cases where one client has a significant bilingual termbase but little or no translation memory, better results can be achieved than in case of a company with large volumes of translation data (Reynolds, 2015).

For the SMT to handle the terminology correctly, it has to be able to identify terms in the translatable content. Two term identifying workflows for SMT have proven to be useful: identifying terms in SMT system training data, namely in parallel and monolingual corpora used for the creation of models, and in the translatable content prior to translation, this is done by preprocessing the text with existing terminology resources (Skadins et al., 2013).

4.1 TERM EXTRACTION

Term extraction can be defined as the operation of identifying the so-called term candidates in a given text or corpus. Term extraction generally involves four steps:

- compilation of a specialized corpus,
- extraction of term candidates,
- validation of the term candidates and
- automatic or semi-automatic creation of terminological records.[4]

The traditional way of creating terminological resources is the manual compilation by human effort. Automatic terminology extraction (ATE) is a natural language processing task that involves extracting terminology, which has been used to identify domain-relevant terms applying computational methods.

Term extraction can be monolingual or multilingual (usually bilingual). The goal of monolingual term extraction is to identify the term candidates in a monolingual specialized corpus. Human translators or terminologists can find equivalents in another language for these candidate terms. Bilingual term extraction is based on parallel (i.e. from previously translated texts) or comparable specialized corpora. Comparable corpora may prove useful for term extraction as previously translated data may be only available for some languages or completely unavailable for emerging domains (Blancafort et al., 2010).

Any ATE method has to be based on a text corpus that is *representative* of the specialized domain whose terminology is to be extracted in some ATE applications, the specialized domain is quite restricted and the relevant texts to be analyzed form a finite and well defined set (Heylen and De Hertog, 2015). When an inventory of a client’s in-house terminology is required, the text corpus corresponds to the document collection that the client provides.

The most commonly used terminology extraction methods apply the following approaches: linguistic, statistics, and hybrid. In the linguistic approach, terminology is filtered by linguistic features, by exploiting for example part-of-speech tagging, morphological analysis and shallow parsing. The linguistic approach is language-dependent, as term formation patterns differ from language to language and may therefore be unsuitable for integration into language-independent systems. On the other hand, the statistical approach is language-independent, as it is based on perusing repeated sequences of lexical items. The most common approach is the hybrid approach, combining linguistic rules and statistical filters.

[4] http://termcoord.eu/discover/free-term-extractors/ (retrieved 13.4.2015).

4.2 TERM EXTRACTION TOOLS

Term extraction tools are used to help in setting up terminology. Term extraction tools typically provide a list of potential terms, *term candidates*, from a corpus or from a text, usually to be validated by a human user. The following table contains examples of commercially and freely available term extraction tools (in alphabetical order):

Name	Languages supported	Supported file formats	Availability	Available from:
AlchemyAPI	EN, FR, DE, ES, IT, PT, RU, SV	HTML; TXT, or url	Commercial	http://www.alchemyapi.com/api/keyword-extraction
Fivefilters	Any	Plain text via web interface or url	Free	http://fivefilters.org/term-extraction/
Lexterm	Any	TXT, *.csv	Free	https://github.com/LexTerm
SDL MultiTerm Extract	Any	TXT, RTF, *.doc, *.xsl, *.ppt, HTML, TMX[5]	Commercial	http://www.sdl.com/cxc/language/terminology-management/multiterm/extract.html
TaaS	EU official languages, RU, TR	PDF, *.doc, *.xsl, *.ppt, TXT, RTF, XLIFF, HTML, XML, MIF	Basic version free; Premium version commercial	https://term.tilde.com/technology
TerMine[6]	Any	Plain text via web interface, TXT, HTML, PDF	Free	http://www.nactem.ac.uk/software/termine/#form
Terminology Extraction by Translated	EN, IT, FR	Plain text via web interface	Free	http://labs.translated.net/terminology-extraction/
SynchroTerm	All EU official languages,	*.doc, *.xsl, RTF, TXT, HTML, PDF, TMX	Commercial	http://www.terminotix.com

[5] For the full list of supported file formats, see http://downloadcenter.sdl.com/T2011/Docs/SDL_MultiTerm_2011_Extract_User_Guide.pdf (retrieved 13.4.2016).

[6] Specifically developed for the bio-medical area.

This project is co-funded by the European Union

Name	Languages supported	Supported file formats	Availability	Available from:
	except ET, GA; HR; LV; MT			

This list is by no means exhaustive.[7] Reference term lists can be used as a gold standard for the qualitative and quantitative evaluation of automatic term extraction tools (Loginova-Clouet, 2012).

4.3 VALIDATION AND CREATION OF RECORDS

In the next step, the term candidate list can be further filtered using statistical and machine learning methods, followed by validation of terminology. Validation of terminology ensures shorter translation review cycles, fewer changes, more streamlined processes, consistent and correct terminology, shorter time-to-market, and reduced costs.

Validation can be performed manually, semi-automatically or automatically. Manual validation requires a substantial human intervention. Depending on the needs and requirements of the project different types of validators may be required: terminologists, translators, domain experts etc. They may check, validate and annotate term candidates based on various criteria: terminological or non-terminological occurrence, belongingness to the domain etc. (cf. Lušicky and Wissik, 2015). Another possibility of a human intervention would be validation using crowd-sourcing (cf. Lamurias et al., 2015) and other social media components[8]. Automatic validation is mainly based on checking formal criteria.

In the last step, terminological records are compiled based on the accepted term candidates. They may be further annotated and enriched, usually by human intervention supported by terminology management tools.

4.4 SUMMING-UP

Terminological resources have been proven useful in workflows for SMT, but their creation remains labour-intensive. Terminology extraction tools can support the extraction of terminology for these workflows. Terminology extraction tools are currently mainly available either as stand-alone tools, or as an integrated part in

[7] See also http://termcoord.eu/discover/free-term-extractors/term-extraction-tools/ (retrieved 15.4.2016).
[8] See also http://termcoord.eu/people-power-crowd-powered-terminologist/ (retrieved 18.4.2016).

a computer-assisted translation tools. Ideally, the creation of tailor-made terminology resources and their integration into workflows for SMT is integrated into the toolchain, as implemented for example in Terminology as a Service (TaaS)[9] or Let's MT[10].

5. BEST PRACTICE CONCERNING THE IPR ASPECT IN CONNECTION WITH CREATION OF NEW CORPORA

5.1 OVERVIEW OF IPR ISSUES RELATED TO ACQUISITION OF TEXT CORPORA ON THE WEB

The legal framework within which agile corpus acquisition would operate is governed by two sets of legislative provisions: *copyright and database rights* (intellectual property rights, IPR), and *data protection*[11] (privacy and autonomy, i.e. confidentiality, anonymity and access arrangements). Below you will find a summary description based on (DASISH, 2013) which presents an in depth general description, and on (Panacea, 2013) which focuses on legal issues in relation to automatic retrieval from the Web.

5.1.1 COPYRIGHT AND DATABASE RIGHTS

The basic rule in copyright law is that whether or not explicitly stated in the material, all original material published on the Web is protected by copyright[12]. The terms *Copyright* and *Intellectual Property Right* (IPR) basically mean the same. The creator of the data has the IPR on the data even if he or she transfers the ownership of the data to somebody else. The legislation, also called intellectual property law, which regulates these matters, is not identical throughout European and other countries, but the basic principle of copyright protection is implemented throughout.

The different national legislations may have implemented *copyright exceptions*. Such exceptions usually refer to rules in the copyright legislation which state that it is legal to use original material without asking permission in different cases, such as e.g. for citations, snippeting and educational activities. Neither EU as such[13] nor any European countries have copyright exceptions and limitations implemented in their legislations that would allow automatic harvesting, processing and subsequent sharing of the harvested material, neither for research purposes nor for commercial use. However, material does exists that by virtue of its nature is classified as not protected work. This type of material will be mainly works made by the public administration or legislature, both

[9] http://www.taas-project.eu/uploads/Integration%20with%20SMT%20systems.pdf (retrieved 18.4.2016)
[10] https://www.letsmt.eu (retrieved 18.4.2016)
[11] "Data Protection Directive" (95/46/EC)
[12] This is also true for material published through other media than the internet.
[13] See Article 5 of the EU Copyright Directive about exceptions and limitations.

of which are of public interest and published for the public. In some countries outside of the EU the legislation facilitates lawful text and data mining under the legal doctrine of “fair use”.

Web crawling, whether manual or automatic, for LT purposes involves the copying of content and cannot be expected to fall under copyright exceptions and limitations in most laws in the EU. Web crawling will thus require separate permission from the rights holders.

Derivative works resulting from processing of data obtained through crawling do not fall under copyright exceptions and limitations in most laws in the EU. Creation of derivative works will thus require additional separate permission from the rights holder.

Sharing of data obtained through web crawling or derivative works is restricted by copyright law, and will thus require additional separate permission from the rights holder.

5.1.2 DATA PROTECTION

These provisions concern personal data, they are very strict, and they apply to material collected from the Web where persons may be identified, including data harvested from social media such as Twitter, Facebook, etc. The data protection provisions require informed consent from the person(s) involved, before the data can be published.

In June 2015 the Council of the EU agreed on a proposal for a new EU data protection regulation. The regulation is intended “to harmonize legislation across the EU and remove unnecessary obstacles that are currently in place due to multiple legislations” cf. EU Data Protection Legislation[14]. The main elements of the agreement are a) An enhanced level of data protection, b) Increased business opportunities in the digital single market, c) More and better tools to enforce compliance with the data protection rules, and d) Guarantees regarding transfers of personal data outside the EU[15]. In conclusion, the new regulation will not make it easier to legally crawl, harvest, process, and share data for LT purposes, such as e.g. machine translation.

5.2 BEST PRACTICE TOWARDS LAWFUL DATA ACQUISITION

5.2.1 ACQUISTION OF EXISTING RESOURCES

Several European players have extensive experience in compiling catalogues and repositories containing monolingual and/or parallel corpora. Such players are for instance ELRA which provides a very comprehensive catalogue[16], and also META-SHARE[17] and OLAC[18]. CLARIN[19] offers a repository with many different materials. The catalogue of LT resources established in the LT_Observatory project represents up-to-date valuable information.

[14] www.eudataprotectionregulation.com
[15] http://www.consilium.europa.eu/en/press/press-releases/2015/06/15-jha-data-protection/
[16] http://www.elra.info/en/catalogues/catalogue-language-resources/
[17] http://www.elra.info/en/catalogues/meta-share/
[18] http://www.language-archives.org/
[19] http://www.clarin.eu/

These materials for the most part come with a price tag and/or licensing conditions governing their use. The existence of license conditions attached to the use of a specific resource means that an agreement between the IPR owner and the provider, e.g. ELRA, has been entered into about the conditions regarding distribution and use of the resource. Extensive sets of license types covering a broad range of allowed or not allowed types of usage are presented on the websites of e.g. CLARIN and META-SHARE.

5.3 SUMMING UP – HOW TO LAWFULLY ACQUIRE DATA THROUGH WEB-CRAWLING

Given that the use of Web data is restricted by the copyright legislation unless otherwise stated, and given that this type of data is extremely useful amongst others for development of statistical data driven MT systems, the question of how to go about using Web data lawfully has been investigated and discussed extensively. Unfortunately, there are no easy answers or shortcuts, but a very useful case study has been conducted in the PANACEA[20] project, and the result is described in (Arranz & Hamon, 2012) and summarized here.

The process of assessing whether to harvest Web data and obtaining permission to use these can be split up into three steps. First, you have to locate the data relevant for your MT systems in terms of number sources and especially what means and tools that are needed to handle these data (as described in section 3). The next step will be to get a clear picture of the execution costs, i.e. will it, seen from a cost-benefit point of view, be worthwhile to conduct this time-consuming procedure. The questions to be answered and issues to be examined during this analysis are,

- conditions for using the data at a Web site should be scrutinized carefully (are data use possible? Are some of the data public, i.e. freely available?)
- possibility of identifying responsible content providers that can answer questions such as, what are the terms with respect to obtaining data usage rights?

The final step will be to evaluate the pieces of information collected. Even though the immediate evaluation falls in favour of starting the negotiations about data usage rights, it is important to bear in mind that such negotiations can be cumbersome and time consuming to conduct. Having managed to identify a contact point, which in itself can be quite difficult, negotiations about usage or licensing conditions must be carried out. In that process the data owner will typically want to know what their content will be used for as they are afraid of misuse of their data, and they will often only accept a usage identical to their own intended usage. In this process it is not unusual that many data owners are not familiar with concepts such as Human Language Technologies, machine learning etc. Given the complexity of the process, such negotiations can therefore be both lengthy and complicated.

[20] http://www.panacea-lr.eu/, see Publications

This project is co-funded by the European Union

6. SUMMARY AND CONCLUSIONS OF RECOMMENDATIONS AND GUIDELINES

In this report we have provided guidelines for what we believe are the most important issues in LR collection and valorisation, namely value-adding to existing LRs (section 2), creating new LRs (section 3), creating terminology (section 4), and finally dealing with Intellectual Property Right (section 5). Each of the sections and its recommendations should be read in its own right, but for ease of reading we also provide here a short summary for each of the themes.

The guidelines for **adding value** to already existing LRs focus on adding missing metadata as this will greatly improve the value for users. The guidelines also describe how to search for additional information, and how to use alternative, less formal metadata like *Tags* when e.g. a value for *Domain* is missing. This section also includes recommendations for adhering to standards, in particular the Dublin Core standard for metadata.

The guidelines for finding **relevant tools and data for creating SMT relevant collections** describe which tools are relevant, and points to the fact that no single place exists where open source software for the full task of creating parallel corpora can be found. A few open source tools for this purpose are recommended. The guidelines also provide recommendations on how to locate relevant and multilingual websites, cleaning up and preparing documents and performing sentence alignment. These are crucial tasks and they cannot be performed without human intervention; the guidelines mention where the human intervention is advisable and necessary.

The guidelines for **terminology** focus on terminology extraction: the creation of new term lists, be it monolingual or bilingual. Automatic term extraction can be done with the help of tools, and the guidelines contain an alphabetical list of tools. It is recommended to always have humans validate automatically created term lists, as this significantly improves the quality of translations.

The guidelines for **IPR in relation to the creation of new corpora** recommend using data where the rights are cleared – e.g. LRs from ELRA, and other providers where license is regulated. CLARIN, META-SHARE and others also use licenses so that the user can safely use data in accordance with the license. However, when you create new corpora from the Web or from other text sources, the rights to use this data have to be negotiated and this can be very time-consuming. In some countries there is copyright exception for specific purposes, and in the US they apply the concept of "fair use". Also, this section mentions that materials exist which do not fall under the copyright protection law. This mainly concerns texts from public administration – such texts can therefore be very useful in this context.

As a summary of the guidelines for these four issues, described in section 2-5, we can say that

- there are pretty firm guidelines for value-adding, and that value-adding is mostly manual
- tools exist that can help the (semi-)automatic creation of new corpora, but it requires some technical skills and some work to use them
- tools exist that can help the (semi-)automatic creation of new term lists
- the issues of Intellectual Property Rights hamper the free use of materials from the web. Guidelines on exceptions are made.

7. REFERENCES

Arcan, Michael, Claudio Giuliano, Marco Turchi and Paul Buitelaar. 2014. Identification of Bilingual Terms from Monolingual Documents for Statistical Machine Translation. In: *Proceedings of CompuTerm 2014*. Available at http://www.aclweb.org/anthology/W14-4803.

Arranz V, Hamon O. (2012). On the Way to a Legal Sharing of Web Applications in NLP. 2012. In: *Proceedings of LREC 2012*, Istanbul, Turkey.

Blancafort, Helena, Béatrice Daille, Tatiana Gornostay and Claude Méchoulam. 2010. TTC: Terminology Extraction, Translation Tools and Comparable Corpora. In: *Proceedings of the 14th EURALEX International Congress*.

DASISH report. 2013. DASISH Deliverable D6.1, http://dasish.eu/publications/projectreports/D6.1_final.pdf.

Esplà-Gomis, M. and M. L. Forcada. 2010. Combining Content-Based and URL-Based Heuristics to Harvest Aligned Bitexts from Multilingual Sites with Bitextor. In: *The Prague Bulletin of Mathemathical Lingustics*, 93:77–86.

Heylen, Kris and Dirk de Hertog. 2015.. Automatic Term Extraction. In: *Handbook of Terminology,* Kockaert, H. J. and Steurs, F. (Eds.), John Benjamins, 203-221.

Koehn, P. and J. Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In: *Proceedings of the Second Workshop on Statistical Machine Translation*, pp 224–227, Prague, Czech Rep.

Loginova-Clouet, Elizaveta, Anita Gojun, Helena Blancafort, Marie Guegan, Tatiana Gornostay, et al. 2012. Reference Lists for the Evaluation of Term Extraction Tools.In: *Terminology and Knowledge Engineering Conference (TKE), Jun 2012*, Madrid, Spain. Available at https://hal.archives-ouvertes.fr/hal-00816566

Lamurias Andre, Vasco Pedro, Luka Clarke and Francisco M. Couto .2015.. Annotating biomedical ontology terms in electronic health records using crowd-sourcing. In: *Proceedings of the International Conference on Biomedical Ontology*. Available at http://ceur-ws.org/Vol-1515/early1.pdf

Lušicky, Vesna and Tanja Wissik. 2015. *Procedural Manual on Terminology*. Translation-oriented Terminology Work. GIZ/SEA.

This project is co-funded by the European Union

Mastropavlos N. & Papavassiliou V. 2011. Domain Adaptation of Statistical Machine Translation using Web-Crawled Resources: A case study. In: *Proceedings from the 10th International Conference of Greek Linguistics*.

Melby, Alan (2012). Terminology in the age of multilingual corpora. In: *JosTrans – Journal of Specialised Translation. 18/2012*. Available at: http://www.jostrans.org/issue18/art_melby.php

Moore, C., Robert: Fast and Accurate Sentence Alignment of Bilingual Corpora. 2002. In *Machine Translation: From Research to Real Users, the proceedings of the 5th conference of the Association for Machine Translation in the Americas.*

8. APPENDIX A

Hunalign and BSA

1. Introduction

This appendix to the report describes two sentence aligner tools, Hunalign and BSA, on the basis of a simple test material containing about 100 sentences in German and English. The description mainly concentrates on the following aspects: the installation procedure, user interface functionality, alignment algorithm principles, and alignment results for different bead types (alignment of sentences where the texts to be aligned have a different number of sentences) including an evaluation of the results.

2.Hunalign

2.1 Hunalign and LF aligner

Hunalign (Varga et al., 2005) aligns bilingual text on the sentence level. The input to the alignment tool must be tokenized and sentence segmented in both languages. The tokenization and segmentation tasks are not included in the Hunalign tool per se. These routines can be executed by some tool preferred by the user or they can be executed by LF Aligner - which is a wrapper created for Hunalign by Andras Farkas[21].

LF Aligner deals with document format conversions, tokenization, sentence segmentation and other stuff making the use of Hunalign much easier for the user. LF Aligner's format conversion scripts allow input in e.g. txt, doc, docx, rtf, html and pdf format and output can be e.g. TMX and xls. The documentation however recommends that the input should be plain text with UTF-8 encoding.

2.2 Hunalign alignment algorithm

The alignment algorithm employed by Hunalign is based on both Gale-Church sentence length (Gale and Church, 1993) and on lexical similarity. The Gale-Church part of the algorithm is based on the assumption that sentences in one language tend to be translated into sentences of a similar length in the target language. The lexical part of the Hunalign algorithm makes use of a dictionary. If an existing dictionary is not available Hunalign will create a dictionary from the data set itself using the following method: 1) create an initial alignment based on sentence length, 2) create a dictionary combining the words of this alignment and 3) re-align with the use of the new dictionary.

[21] http://sourceforge.net/projects/aligner

This project is co-funded by the European Union

2.3 Installation of Hunalign/LF Aligner and the user interface

Hunalign and LF Aligner can be used on a Linux, Mac or Windows platform. We have used a Mac platform for our test of Hunalign/LF Aligner.

Installation of Hunalign without the wrapper is not complicated at all, but some knowledge of the terminal window and basic UNIX commands are definitely qualifications that will facilitate the installation procedure. After installation, Hunalign is also run via the command line using a variety of arguments and options; i.e. the tool has no user interface as such. If a user is not familiar with the terminal window and the structure of UNIX commands this will require a learning period, albeit a relatively short one.

Installation of Hunalign in conjunction with LF Aligner is very easy and requires no prior knowledge of command line programming. Running the programs is also totally uncomplicated. LF Aligner contributes with a very simple command line user interface that interactively obtains the necessary information from the user. It appears from the documentation that the Windows user interface is more sophisticated than the Mac interface (which is used in our test). It is recommended to use Hunalign in conjunction with LF Aligner. The below descriptions and evaluations are based on the functionality of Hunalign including the wrapper LF aligner.

Hunalign is written in C++, GNU public license and LF Aligner is described as open source and mainly written in Perl.

2.4 Hunalign and lexicons

Alignment procedures based on sentence length is generally considered relatively fast - and lexical methods are usually slower and more accurate. The lexical methods are however dependent on a certain text quantity. The performance of lexical methods will degrade, as the texts to be aligned are shorter because there is less data to support statistical inference of a bilingual lexicon (Abdul-Rauf et al., 2012).

We have not found guidelines on how to best create a lexicon for Hunalign, and how size and quality of the lexicon relate more precisely to the quality of alignments. For our bilingual text examples used in this description we have not employed an external lexicon, but only the lexicon automatically created during the alignment process. It should be noted though that the automatically created lexicon combines primarily single words in source and target languages whereas alignments should sometimes be between phrases. Often relatively long phrases are probably the most reliable parallel sentence indicators. This means that an external dictionary containing phrases would probably yield better results than what our tests show.

This project is co-funded by the European Union

2. 5 Hunalign efficiency

This experiment includes only a very modest sized corpus far from challenging the efficiency of Hunalign. Other experiments have led to the conclusions that Hunalign shows very good performance in comparison with other alignment tools when the total number of sentences is 20k or less. With an input larger than 20k sentences, memory issues will make alignments more or less impossible (Toral et al., 2012, Sneha & Kumar, 2014). LF Aligner facilitates alignment of more than 20k sentences by splitting the corpora into smaller chunks, but the downside is that this will result in worse lexicon estimates.

3. Microsoft's Bilingual Sentence Aligner (BSA)

3.1 The BSA alignment algorithm

The BSA aligner is an implementation of the method described in Moore (2002). The algorithm described here is a three-step process. First the parallel corpus is aligned using the sentence-length-based model described in Brown et al. (1991). Then the aligned sentence pairs assigned the highest probability are used as training data for a lexical model that is a modified version of the IBM Translation model, cf. (Brown et al, 1991). Finally, the parallel data are realigned based on the acquired knowledge of both sentence length and word correspondences. No external sources such as bilingual dictionaries are involved in the training algorithm, and the BSA aligner is language independent. The main difference between BSA and Hunalign is thus that BSA employs a trained lexical model, whereas Hunalign employs a crude word-by-word dictionary and Hunalign permits the flexibility of different dictionaries for different tasks. The BSA application can be downloaded from the following address, https://www.microsoft.com/en-us/download/details.aspx?id=52608. The restrictions for use of the BSA tool are stated in the Microsoft Research End User License Agreement (MSR-EULA).

3.2 Installation of BSA and the user interface

The system requirement for the BSA application is Windows 7, 8, or 10. Since the downloaded data are compressed as a Tar Gz file, data must be unzipped using e.g. an open source decompression program such as 7zip that can be downloaded here.

According to the BSA Readme file, the application is implemented as Perl scripts. As a consequence of this software choice, the appropriate version of Perl must be downloaded, such as Strawberry Perl 5.24.01. This version of Perl can be found at http://strawberryperl.com/. The BSA Readme also informs that the two input files of parallel data must be preprocessed/segmented so that each sentence occurs a on separate line in the file. The applications used for this exercise were found amongst the CST online tools cf., https://cst.dk/tools/index.php?lang=en. It is also required that the parallel data are stored as plain text with UTF-8 encoding.

BSA is run from the DOS command-line and does not have a user interface as such. When sentences of the parallel data are segmented, the following command is used that also invokes several other Perl program files:
align-sents-all.pl <source file> <target file>[22]
The alignment result is then stored in two separate files:
<source aligned> <target aligned>

4. Test material

The parallel texts used in this evaluation context consist of approx. 100 sentences extracted from Grimm's fairytale, *Eine Gesicthe über Angst*, 1837 and a translated version in English.

The overall approach was first to establish a test suite that both the BSA and the Hunalign applications aligned correctly and then subsequently construct some test suites that included the following bead types,

- 0-1 (a test suite in which sentences from the source language were deleted)
- 1-0 (a test suite in which sentences from the target language were deleted)
- 1-2 (a test suite in which 1 sentence from the source language was split into two target sentences)
- 2-1 (a test suite in which 2 sentences from the target language were concatenated into one sentence)

Using a text from the 19th century posed some challenges for the segmentation/alignment tools. Since the tools described here are designed to process modern texts, the use of ancient orthography in terms of punctuation markers do not comply with the functionality of the tools. In addition, the German and English fairytale versions use different punctuation principles, which also results in bad alignment results.

For this test we therefore decided to change the punctuation markers to modern versions and we harmonized the punctuation principles of the two texts. The main test set thus consists of German and English text versions with exactly the same number of equivalent sentences in exactly the same order (resulting in only 1:1 alignments). In the next steps we created the other different configurations described above: 1:0 and 0:1 (when a sentence in one language does not have an equivalent sentence in the other language) and 1:2 and 2:1 (when one sentence is represented as two sentences in the other language).

[22] Another version of the code is available that makes it possible for the aligner to get access to parallel data stored elsewhere than in the working directory

4.1 Hunalign test results

Our test shows that Hunalign performs very well on material where all sentences can be aligned 1:1. Hunalign made no errors.

The test material containing 0:1 and 1:0 examples shows that Hunalign is not always successful when a particular sentence has no equivalent sentence in the other language. The below example shows a 1-0 configuration as the German sentence *Das wird wohl eine Kunst sein, von der ich auch nichts verstehe* has no equivalent English sentence. The tool has handled this inconsistency by combining the two preceding German sentences with one English sentence and by combining the “extra” German sentence with an “irrelevant” English sentence (figure 1). The result is two misaligned segments. This example of a misalignment also shows that Hunalign/LF Aligner does not somehow mark a suspected inconsistency - this makes the error virtually impossible to detect in most cases (the yellow marker is our emphasis). As the third segment shows Hunalign falls in step again immediately afterwards.

1	"Immer sagen sie, es gruselt mir, es gruselt mir! Mir gruselt's nicht.	"They are always saying 'it makes me shudder, it makes me shudder!'.
2	Das wird wohl eine Kunst sein, von der ich auch nichts verstehe."	It does not make me shudder," thought he.
3	Nun geschah es, daß der Vater einmal zu ihm sprach: "Hör, du in der Ecke dort, du wirst groß und stark, du mußt auch etwas lernen, womit du dein Brot verdienst.	Now it came to pass that his father said to him one day "Hearken to me, thou fellow in the corner there, thou art growing tall and strong, and thou too must learn something by which thou canst earn thy living.

FIGURE 1 1:0 CONFIGURATION - MISALIGNED

The test sentences with 1:2 and 2:1 configurations were all successfully aligned with Hunalign (figure 2).

1	Mir gruselt's nicht.	It does not make me shudder," thought he.
2	Das wird wohl eine Kunst sein, von der ich auch nichts verstehe", nun geschah es, daß der Vater einmal zu ihm sprach: "Hör, du in der Ecke dort, du wirst groß und stark, du mußt auch etwas lernen, womit du dein Brot verdienst.	"That, too, must be an art of which I understand nothing." Now it came to pass that his father said to him one day "Hearken to me, thou fellow in the corner there, thou art growing tall and strong, and thou too must learn something by which thou canst earn thy living.
3	Siehst du, wie dein Bruder sich Mühe gibt, aber an dir ist Hopfen und Malz verloren."	Look how thy brother works, but thou dost not even earn thy salt."

FIGURE 2: 1:2 CONFIGURATION – SUCCESSFUL ALIGNMENT

This project is co-funded by the European Union

Hunalign cannot handle alignment of sentences occurring in opposite order i.e., segments X and Y in one language corresponding to segments Y' X' in the other language.

4.2 BSA test results

BSA performs well on material where all the sentences in parallel data can be aligned 1:1. As mentioned above, the evaluation scenario that was established consists of four test suites. In relation to the 1:0 and 0:1 scenarios, the BSA aligner performed perfectly. In these situations the BSA aligner simply chose to remove this "redundant" sentence resulting in a correct alignment of the remaining sentences.

Regarding the other the two test suites, the BSA aligner performed in approximately the same way. In the case of either concatenated source sentences (1-2), or target sentences (2-1), the BSA aligner removes the sentences in focus on both parts of the parallel data resulting in a correct alignment of the rest of the sentences.
Similar to Hunalign, BSA cannot handle alignment of sentences occurring in opposite order i.e., segments X and Y in one language corresponding to segments Y' X' in the other language.

5. Summing-up

User-friendliness is an important criterion when reviewing and assessing software applications. Both Hunalign and BSA score relatively low in this respect as both aligners make use of a command-line user interface (even if the LF Aligner provides Hunalign with a somewhat better interface, especially in the Windows version). If these applications had addressed the commercial market, the aligners would undoubtedly have been embedded in a graphical interface.

With respect to aligning 1:1 parallel sentences, the test shows that both aligners perform well. Regarding the other bead types included in the test (0:1, 1:0, 1:2, 2:1), the BSA aligner performs well by excluding the sentences that were redundant and by keeping the remaining and correctly aligned sentences in the result files. The test results indicate that Hunalign performs well when confronted with the bead types 2:1, 1:2, while the bead types 1:0 and 0:1 are treated less successfully resulting in some misaligned sentences.

As mentioned above in section 3.5 the lack of software, that integrates different existing research tools into new software packages, has the effect that users must create their own workflows combining the tools into meaningful pipelines. Seen from this perspective, Hunalign (in conjunction with LF aligner) outperforms the BSA application. Not only does LF aligner offer format conversion, but it does also, in contrast to BSA, take care of the necessary preprocessing steps before aligning sentences, i.e. tokenization and sentence segmentation.

6. References

Abdul-Rauf S., M. Fishel, P. Lambert, S. Noubours, R. Senrich. 2012. Extrinsic Evaluation of Sentence Alignment Systems. In: *Proceedings of LREC workshop*.

Brown, P.F., Lai, J.C., Mercer, R.L. Aligning Sentences in Parallel Corpora. 1991. In: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California (1991) 169–176.

Moore, C., Robert: Fast and Accurate Sentence Alignment of Bilingual Corpora. 2002. In *Machine Translation: From Research to Real Users, the proceedings of the 5th conference of the Association for Machine Translation in the Americas.*

Sheha D., G. B. Kumar. 2014. Comparison of Various Bilingual Sentence Alignment Methods for Parallel Corpora Development. In: *International Journal of Engineering Research & Technology* (IJERT) ISSN: 2278-0181 Vol. 3 issue 4.

Toral A., M. Poch, P. Pecina, G. Thurmair. 2012. Efficiency-based evaluation of aligner for industrial applications. In: *Proceedings of the 16th EAMT Conference*.

Varga, D., L. Nemeth, P. Halacsy, A. Kornai, V. Tron, and V. Nagy. 2005. Parallel corpora for medium density languages. In: *Recent advances in natural Language Processing* (RANLP 2005).

William A. Gale and Kenneth W, Church. 1993. A program for aligning sentences in bilingual corpora. In: *Computational Linguistics*.

Panacea report (2013), PANACEA Project, Deliverable D2.4 Annex 1, January. Not publicly available. The report can be obtained by requesting it from the project.

Pecina P., Toral T., Papavassiliou V., Prokopidis P., & Genabith van J. (2012). Domain Adaptation of Statistical Machine Translation using Web-Crawled Resources: A case study. *Proceedings from the 16th EAMT Conference.*

Pinnis, Marcis and Raivis Skadins (2012). *MT Adaptation for Under-Resourced Domains – What Works and What Not. Human Language Technologies – The Baltic Perspective*, A. Tavast et al. (Eds.). Available at http://ebooks.iospress.nl/publication/7500

Resnik, P. & N. A. Smith. 2003. The Web as a parallel corpus. In: *Computational Linguistics*, 29:349–380.

Reynold, Peter. 2015. Machine translation, translation memory and terminology management. In: *Handbook of Terminology,* Kockaert, H. J. and Steurs, F. (Eds.), John Benjamins, 276-287.

Skadins, Raivis, Marcis Pinnis, Tatiana Gornostay and Anrejs Vasiljevs .2013. Application of Online Terminology Services in Statistical Machine Translation.In: *Proceedings of the XIV. Machine Translation Summit*, 281-286.

Theobald, M., Siddharth, J., Paepcke, A. 2008. SpotSigs: Robust and Efficient Near Duplicate Detection in Large Web Collections. In: *31st annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR 2008).

Varga D., Németh L., Halácsy P., Kornai A., Trón V., & Nagy V. 2005. Parallel corpora for medium density languages. In: *Proceedings of the RANLP*.

Wu, H., & Zong C. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In: *Proceedings of the 22nd International Conference on Computational Lingustics – Volume 1*.

This project is co-funded by the European Union