OPTIMISING KNOWLEDGE AND PRACTICE FOR OPERATIONAL
LANGUAGE RESOURCES

An LT_Observatory Workshop
co-organised with ELRA/ELDA

Date: 22 September 2016
Venue:
ELRA/ELDA
9, Rue des Cordelières
75013 Paris

BRIEFING

The following notes are intended to inform and facilitate the discussions with experts:

The LT Observatory project (LTO) and ELRA (the European Language Resource Association) are holding a workshop (“charrette”) as part of ongoing effort to consolidate and support a community platform for the production, circulation & management of parallel language resources (LRs) for commercial machine translation (MT) and other applications.

This charrette will form a critical step in our endeavour. We would like to look forward to the kind of technologies, environment and use cases that are likely to emerge in the LR-MT space in the next 2 to 5 years. We wish to sustain the assets of the project (the LR Catalogue, http://www.lt-innovate.org/Lt-observe/language-resources-observatory) and envision a specific pathway forward, in collaboration with our workshop co-organiser ELRA/ELDA. In parallel, the work being carried out by the CEF.AT programme (https://ec.europa.eu/digital-single-market/en/automated-translation) as part of the MT agenda for building a resources infrastructure for public service requirements in Europe (the so-called European Language Resources Coordination (www.lr-coordination.eu in which ELRA is involved) will be observed and collaboration opportunities identified. We shall also invite a speaker to address the related but distinct issue of circulating and enabling speech/spoken LRs, for speech to speech translation purposes in the first instance, but also as part of the same consolidated effort to promote best practices in handling LR assets in general.

A word on best practices. We interpret this well-known concept as a method for simplifying, automating and therefore lowering the cost of discovering, repurposing, repackaging, providing, and (re)using LRs. Our focus must be on the practical matters that can optimise the use (and creation) of LRs, and then retro-engineering these conditions back into existing LRs. LRs – data – are ultimately a component of a software process. We therefore need to be clear about where MT technology is going, and encourage the adaptation/tailoring of LRs to ensure their fit for purpose in an environment of fast-evolving MT and NLP technology.

We must as a result synthetise, hierarchize and enrich our understanding of the different issues/players involved in this space, which we interpret as an innovation opportunity. We are, therefore, particularly concerned to find, share and encourage new solutions to problems in this
domain that can move rapidly from research to the market and benefit from existing innovation drivers.

The focus of our workshop will therefore be on the need for:

- **Standards of best practice in LR management and associated Data Management Plans** (specify, find, create/repackage, share, evaluate, use, improve LRs, sustain)
- **Tools to enable such practices, and also to rapidly expand the range of available LRs across more (EU) language pairs but also across domains**

Below is a list of the issues we wish to address. We would ask you as an expert to reflect on one or more of these issues and be prepared to make a brief presentation – literally 5-7 minutes of bullet points - of your understanding/criticism/proposals/solutions/reactions to these questions. Then we will share, brainstorm and try to reach a constructive consensus on our findings with a path ahead.

**Collecting LRs**

**Questions**
How will new MT models impact the need for data (if at all)?

Will parallel data collection be more powerful as a driver than creating parallel LRs? In other words, what LRs do you suspect exist in the real world that have not yet been discovered – and why not?

Are there any (hidden) mechanisms that could be activated to accelerate better knowledge of what LRs exist across the EU?

How can User Generated Data be collected and used (sparsity and IPR issues)

**Evaluating LRs**

**Questions**
What is the best way to characterise domains in MT practice – semantic domain and text type?

How much does translation purpose impact data choice?

How much does the use case of gisting or search/reading for information in another language influence the kind of LR that is used in an MT context? (i.e. gisting v. human-quality translation)

As semantic and maybe domain-relevant criteria of LRs may become more important when selecting LRs, is there a compelling need for new metadata standards for identifying LRs that go beyond the current Dublin Core set, the meta-share/ELRA descriptors, etc.?

**Creating LRs**

**Questions**
Which are the key variables that must be addressed to ensure domain relevance and MT-readiness/fit for purpose when creating new LRs? Are these standardisable?

When new LRs are created, what are the optimal minimum criteria for characterising them in terms of metadata?
Is rapid LR creation a natural next step in the MT process that will render obsolete the collecting and curating of LR?

Will web-crawling and LR creation require any safeguards that some sort of standard might ensure?

What are the compromises between size and quality for a robust MT system?

**Collaborating around LRs**

**Questions**
What is currently missing from the technology or translation playbook that needs to be initiated to overcome the LR deficit?

Can we expect tools for automatically preparing texts in 2 languages for LR MT readiness? For example, take a semantic/domain footprint of the source document and use it to find appropriate resources for the target(s)?

What resources/methods exist to ensure that all EU languages can eventually be covered by shared efforts? At which level (government, university, business associations, etc.) could we activate efforts to provide rich coverage? What are the brakes and accelerators?

How can the LR user/creator community best come together to resolve any/all of the issues mentioned above? What is missing in terms of organisation and protection? What is the big burning problem that haunts your LR reflections?

Is the MT-Data collections sharable between different MT engines/domains/players? From technical point of view as well as competitive advantages that data may bring in?

**OUTLOOK**

The workshop will feed into a Final Conference, to be organised in November or December 2016, at which the tools and findings of the LTO project will be presented to a broad cross-selection of stakeholders.
BEST PRACTICE IN LR MANAGEMENT

Introduction

This paper on best practices in accessing and using Language Data Resources for Translation Automation is intended to help anyone engaging on a machine translation (MT) project. It is inspired by work carried out under the LT-Observatory project funded by the European Commission (2015-2016). A wide range of translation practitioners, researchers and language service suppliers, who all need language resources in their work, have helped enrich it.

By language data resources (LRs), we mean the following:
- bilingual corpora (both aligned and comparable)
- monolingual corpora used to build a target language model
- terminology resources (termbases, lists, etc.)
that may be used by anyone managing a translation task. We want to simplify and streamline the work for anyone responsible for this process.

The paper covers the following four aspects of language resource management:

1. Adding value to existing translation resources
2. Finding useful new resources from the Web
3. Handling terminology resources
4. Understanding the legal and IPR issues around LR creation and usage

1. HOW TO ADD VALUE TO EXISTING RESOURCES

Many useful translation language resources (hereinafter we will use the abbreviation LRs) found in existing repositories present operational challenges. This is usually because these resources lack basic information about copyright issues, domain, owner or resource format, etc. This information is usually known as metadata – data about the data in question and is contained in a special file linked to the resource.

Poor quality metadata means that it may be hard to identify and locate LRs, and know whether they are best suited to a particular translation task.

1.1 Adding useful metadata

Below we give a maximalist list of metadata recommended by many authorities for LRs to be used for MT. It gives an idea of the complexity of the problem of identifying any LR. In most practical cases the list can be reduced to 6 or 7 pieces of information. However, it is worth knowing how the MT community has understood the nature of the maximal criteria for evaluating LRs:

<table>
<thead>
<tr>
<th>Title</th>
<th>Name of the language resource, e.g. Estonian-Latvian parallel corpus of building product texts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resource type</td>
<td>Some examples could be: corpus, classifications scheme, lexical resource, terminology resource, tool</td>
</tr>
</tbody>
</table>
In operational situations, what users usually need most is information on the translation **domain**, the **languages**, the **format**, the **cost** and the **legal context**. It is to be hoped that users will add any new practical/operational information about the resource in the comments or tags provided in the LT Observe Catalogue (www.lt-innovate.org/lt-observe/resources-list) so that the community as a whole can help build its own better-quality LR repository over time. LT Observe will be able to analyse these data and help improve the catalogue as users add comments.

### 1.2 Using standard metadata

We recommend that the list of **metadata** (information about the resource in question) corresponds to a standard. A useful standard is the **Dublin Core (DC)** metadata set for resources, but the following adjustments will make this standard more relevant to the LR context:

- **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 even though it is not part of the DC set. Resource size (how many words/phrases/sentences etc.) can be an important factor when building an MT engine. The industry tends to say, for example, that you need “one million words” of data to build a reasonable resource for MT.
- **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 of the resource.
- **Modality** (i.e. text or speech) should be included as spoken resources will eventually be included in the LT Observe resource.
- **Availability** should be included, as parallel resources are often fairly expensive. This whole area of the cost of LRs will be carefully monitored by LT Observe, and information will be provided in the catalogue.
- **Tags** should be included as they will often prove useful for searching for LRs in a large catalogue, and helping provide analytics on LR usage to the community.

We also recommend that relevant new information about resources is made available to potential LR users, vendors and buyers so they can examine and test resources in relation to their own work context and requirements. This is part of the community building and feedback service to which LT Observe is committed.

2. FINDING LANGUAGE DATA AND TOOLS ON THE WEB

There are vast quantities of potentially relevant documents on the web that can be adapted or used as LRs for machine translation. For our purposes, the main challenge is to identify and acquire domain-specific bilingual corpora, especially for languages with smaller speaker communities and fewer overall resources.

There are three steps to be taken for acquiring in-domain parallel data for MT usage.

1. A **focused search** for relevant websites (and then ranking them). The links found at these websites are then regarded as candidate URL seeds with respect to identifying bilingual documents.
2. The **removal of duplicates**, excluding boilerplate content.
3. **Aligning** the parallel domain-specific corpus.

2.1 Finding domain relevant and multilingual websites

1. Make a list of terms that represents the subject area (or domain) in which the MT system will translate. The “seed” terms on this list will form the search pattern used to find relevant documents on the web.
   One source for designing 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\(^2\).

---

\(^1\) Sketch Engine is a commercial piece of software, https://www.sketchengine.co.uk/

\(^2\) http://eurovoc.europa.eu/drupal/?q=download/subject_oriented&d=en
2. Prioritize the website link list, using any of the following methods.

   a) Model domain-relevant pages by developing neural networks, or
   b) Best-First algorithm that prioritizes between webpages assigned with various scores.

   Hybrids of these two methods can also be found. These are not yet available as open source solutions.

   The list of domain relevant URLs collected during the monolingual web-crawling process will constitute important links for acquiring monolingual data, and also often offer a starting point for identifying websites that can be regarded as candidates for bilingual documents.

   Identifying domain-specific multilingual Websites is a further task.

3. Find parallel texts on the web, using existing applications.

   a) After locating pages with parallel translations, it is necessary to find out whether the candidate pairs are actually translations. One technique to distinguish between true and false positives is to compare the HTML structures in the parallel documents in question. The Web-Mining Architecture STRAND can identify pairs of web pages that are mutual translations. However, the software package is not available and corpora acquired by STRAND on the Web are subject to copyright restrictions. However, databases of URL pairs acquired by STRAND, however, can be downloaded for personal use.

   b) Another system that mines parallel documents from multilingual websites is Bitextor. This application is based on a quantitative approach that looks at file size, text length, tag structures etc.

   c) A similar approach extracts pairs of documents on the web that are likely translations of each other

2.2 Cleaning and preparing documents

   There are a number of solutions for filtering monolingual and multilingual web collections that contain duplicates and near-duplicates. The files created will then be used to train an MT engine to translate.

   3. Removing boilerplate content is important for cleaning files. Web pages often contain HTML navigation links, commercials, and disclaimers that need to be eliminated for translation purposes. One automatic tool to remove boilerplate elements that may perform better than Sketch Engine is Boilerpipe, but it may need a license.

2.3 Generating sentence alignment

   The next step, after identifying your relevant domain documents and preparing them as parallel resources, involves sentence splitting and tokenization. The best reference for these processes is to download the Europarl tools for sentence splitting and tokenization.
Sentence splitting involves cutting up a long, complex sentence into its component parts so that the machine can process it as a series of “simple” sentences. Tokenisation involves identifying the individual words that comprise a sentence. These are both automatic processes for all languages.

The final step in the pipeline of qualifying parallel data as training data for an SMT system is to make sure that the sentences that have been extracted from the files are aligned with the highest possible quality equivalence.

High-quality alignment – i.e. best possible alignment between Language 1 and Language 2 sentences - is crucial for achieving good results. However, the manual checking and evaluation of automatic alignment is very time-consuming.

Sentence alignment is usually evaluated by measuring the output achieved when the aligned corpus is added to the training data. An alternative and much more resource demanding method is to evaluate the alignment results intrinsically, i.e. to measure the alignment error rate.

This latter approach requires a gold standard to be established, which is very labour-intensive. Evidence from evaluation results suggests that the Hunalign sentence aligner with respect to execution time was better than the GMA and BSA aligners, while for memory issues, the Hunalign aligner performed badly compared to the other two sentence aligners.

If you want to extract and download parallel data for SMT systems you have to find websites that publish parallel documents – i.e. the same document in multiple languages. The most challenging task is the lack of open source software 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.

Many useful research tools have been developed as stand-alone applications. But users themselves have to hard code the software that integrates the applications into a single workflow.

However, there are many useful open resource tools that can help you through the pipeline steps. Examples include:

- the bilingual webcrawler that find the links within Websites
- the Hunalign sentence aligner. This still requires human effort. For example, you will have to proceed manually when generating domain specific multilingual seed URL lists; and to evaluate the quality of the outputs from the Hunalign sentence aligner, you will need to check them manually.

3. TERMINOLOGICAL RESOURCES

Terminological resources can be used in a variety of operations:

- Customising a machine translation system,
- Supporting the computer-assisted translation process (multilingual terminology management),
- Optimising search engine (SEO).

---

3 NB: the Hunalign application offers preprocessing software that chunks the input data into smaller pieces, about 5,000 sentences.
Lexicons and terminologies play an important role in any machine translation system, regardless of the principle on which the machine translation tool is based.

Terminological resources include:

- **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.
- **Extracted terms**: in the translation and localisation industry, clients usually require correct and accurate use of specific terminology, often their own in-house terminology. However, in projects where the client terminology is not readily available, although specific in-house terminology must be used, the terminology can be extracted from documents provided by the client.

Terminology collections may contain language equivalents that are rated as unlikely by the SMT system models. If this kind of SMT system is integrated into a translation service workflow, then high quality terminology (i.e. consistent and correct for the client) will not be offered by the SMT suggestions.

Training data for MT engines can also contain contradictory terminology, corporate specific synonyms or vendor-biased terminology. This is why it is important to use customized and validated terminology collections in SMT engines.

In terms of terminology integration, it has been shown that the introduction of bilingual terminology in the translation model can considerably improve translation quality of an out-of-domain system. Anecdotal sources have shown that when a client has a significant bilingual termbase but little or no translation memories, better translation results can be achieved than in the case of a company with large volumes of translation data but no structured terminology (Reynolds 2015).

For an SMT engine 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:

a) identifying terms in SMT system training data, namely in parallel and monolingual corpora used for the creation of models,

b) Identifying terms in the translatable content prior to translation, by pre-processing the text using existing terminology resources.

### 3.1 Extracting terms to feed a translation engine

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:

1. compilation of a specialized corpus,
2. extraction of term candidates,
3. validation of the term candidates and
4. Automatic or semi-automatic creation of terminological records.\(^4\)

The traditional way of creating terminological resources is to do it manually. However, **automatic terminology extraction (ATE)** is a natural language processing task that can extract terminology by using computational methods to identify domain-relevant terms.

Term extraction can be monolingual or multilingual (in fact, this usually means bilingual). The goal of monolingual term extraction is to identify the term candidates in a single-language, specialised corpus. Human translators or terminologists can find then 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.

Any ATE method has to be based on a text corpus that is **representative** of the specialized domain in question. In some ATE applications, the specialized domain is quite restricted and the relevant texts to be analyzed form a finite and well defined set. For a given client, the text corpus to be use to extract terms will be the client’s own document collection.

The most commonly used terminology extraction methods apply the following approaches: linguistic, statistics, and hybrid.

a) In the linguistic approach, terminology is filtered by linguistic features, by using 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.

b) The statistical approach is language-independent, based on examining repeated sequences of lexical items.

c) The hybrid approach is the most common method, combining linguistic rules and statistical filters.

### 3.2 Using term extraction tools

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Languages supported</th>
<th>Supported formats</th>
<th>file</th>
<th>Availability</th>
<th>Available from</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlchemyAPI</td>
<td>EN, FR, DE, ES, IT, PT, RU, SV</td>
<td>HTML; TXT, or url</td>
<td>Commercial</td>
<td><a href="http://www.alchemyapi.com/api/keyword-extraction">http://www.alchemyapi.com/api/keyword-extraction</a></td>
<td></td>
</tr>
<tr>
<td>Fivefilters</td>
<td>Any</td>
<td>Plain text via web interface or url</td>
<td>Free</td>
<td><a href="http://fivefilters.org/term-extraction/">http://fivefilters.org/term-extraction/</a></td>
<td></td>
</tr>
<tr>
<td>Lexterm</td>
<td>Any</td>
<td>TXT, *.csv</td>
<td>Free</td>
<td><a href="https://github.com/LexTerm">https://github.com/LexTerm</a></td>
<td></td>
</tr>
</tbody>
</table>

### 3.3 Validating and creating term records

By validating terminology, you can ensure 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.

a) Manual validation requires 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 methods.

b) Automatic validation is mainly based on checking formal criteria.

For the final 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.

---


6 Specifically developed for the bio-medical area.

7 See also http://termcoord.eu/people-power-crowd-powered-terminologist/ (retrieved 18.4.2016).
4. BEST PRACTICES FOR IPR WHEN CREATING NEW CORPORA

4.1 IPR issues related to acquiring 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* (privacy and autonomy, i.e. confidentiality, anonymity and access arrangements). Below you will find a general summary together with the special conditions on legal issues in relation to automatic retrieval from the Web.

4.1.1 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. The terms *Copyright, Intellectual Property Right* (IPR) and *Data Ownership* basically mean the same. The legislation (also called intellectual property law) is not identical everywhere in Europe or worldwide, but the basic principle of data ownership protection is implemented everywhere. The data owner - the first owner - has, by law, the right to decide who can have lawful access to their content and for what purposes.

Different national legislations may implement *copyright exceptions*. These 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. This is known as “fair use” in the USA. Neither the EU nor any European countries have copyright exceptions and limitations implemented in their legislations that would allow the automatic harvesting, processing and subsequent sharing of the harvested material, either for research purposes or for commercial use. However, there is content that by virtue of its nature is classified as *unprotected work*. This refers will mainly to content made available by public administrations or legislatures, both of which are of public interest and published for that reason.

Web crawling for LT purposes, whether manual or automatic, involves copying content and therefore it cannot be expected to fall under copyright exceptions and limitations (fair use) in most of the EU. This means that web crawling requires separate permissions from all rights holders.

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

In other words, the law does not support web crawling activities in search of useful data. The fact is, however, that web crawling is the most common method of obtaining data for most LT purposes, especially parallel or comparable corpora for translation automation.

4.1.2 Data protection

---

8 “Data Protection Directive” (95/46/EC)
9 This is also true for material published through other media than the internet.
10 See Article 5 of the EU Copyright Directive about exceptions and limitations.
This area covers personal data and the legal provisions in the EU are very strict. They apply to content collected from the web whereby individuals can be identified, including data harvested from social media such as Twitter, Facebook, etc. Data protection provisions require informed consent from the person(s) involved before these 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 Legislation11.

The main features 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,

d) Guarantees regarding transfers of personal data outside the EU12.

For LR purposes, the new regulation will not make it easier to legally crawl, harvest, process, and share data for purposes such as machine translation.

### 4.2 Best practices for lawful data acquisition

#### 4.2.1 Acquiring existing resources

Several European players have extensive experience in compiling catalogues and repositories containing monolingual and/or parallel corpora.

ELRA provides a very comprehensive catalogue13, as do META-SHARE14 and OLAC15. CLARIN16 offers a repository with many different materials, including potential MT resources.

The catalogue of LT resources established on LT Observe represents valuable, up-to-date information on resources that can be sourced in some of the repositories mentioned above. Most of these resources 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 the IPR owner and the provider, e.g. ELRA, have signed an agreement about the conditions regarding the distribution and use of the resource in question. Extensive sets of license types covering a broad range of allowed or prohibited types of usage are presented on the web sites of CLARIN and META-SHARE, for example.

#### 4.2.2 Lawfully acquiring data from web-crawling

Web data is restricted by copyright legislation unless otherwise stated, but these data are extremely useful for many kinds of language technology. How can we overcome this barrier?

Unfortunately, there are no easy answers or shortcuts. A very useful case study was carried out in the PANACEA17 project, and we summarize it here.

---

11 www.eudataprotectionregulation.com
14 http://www.elra.info/en/catalogues/meta-share/
15 http://www.language-archives.org/
16 http://www.clarin.eu/
1. Obtaining permission to use web data is source dependent and therefore potentially very costly, depending on the number of sources that must be approached.
2. When you have managed to identify a contact point (often difficult), you need to negotiate about usage or licensing conditions.
3. The data owner will typically want to know what their content will be used for as they are afraid it will be misused.
4. The owner will therefore only accept usage that is identical to their own intended usage.
5. Many data owners are unfamiliar with concepts such as LT and MT, machine learning etc.
6. As a result, negotiations can be both lengthy and complicated.