Providing a catalogue of Language Resources for Commercial Users

Bente Maegaard, Gerhard Budin, Luz Esparza, Lina Henriksen, Andrew Joscelyne, Steven Krauwer, Vesna Lusicky, Margaretha Mazura, Sussi Olsen, Claus Povlsen, Blanca Rodrigues, Philippe Wacker

CLARIN-DK, University of Copenhagen, Centre for Language Technology, NFI
Njalsgade 140, DK-2300 Copenhagen
E-mail: bmaegaard@hum.ku.dk

1. Introduction

Language resources (LR) are indispensable for the development of tools for machine translation (MT) or various kinds of computer-assisted translation (CAT). In particular language corpora, both parallel and monolingual are considered most important for instance for MT. But corpora are expensive and labour-intensive to create or adapt e.g. for MT usability. Furthermore the availability of LRs differs considerably from language to language. It is true that LRs are created by EU and national projects and institutions, but they and information about them are scattered across Europe, and commercial users are not prepared to search for them. In order to remedy this lack of overview for the professional user, it is important to apply a user driven approach towards the identification/mapping of best practices in terms of collecting relevant LT resources.

The Language Technology Observatory will provide easy access to information about LRs deemed to be useful for MT and other tools for translation. In order to determine what aspects of an LR are useful for MT practitioners, a user study has been made, providing a guide to the most relevant metadata and the most relevant quality criteria. At the same time, knowledge and best practice has been extracted from previous studies (LetsMT!, META-SHARE etc.) on collecting relevant LT resources.

2. Related work

We have taken the point of departure in existing catalogues and repositories. These comprise: ELRA Catalogue of Language Resources (http://www.elra.info/en/catalogues/), CLARIN VLO (the Virtual Language Observatory, http://clarin.eu/content/virtual-language-observatory), a search facility of metadata for language resources, META-SHARE (http://www.meta-net.eu/meta-share), OPUS - the open parallel corpus (http://opus.lingfil.uu.se/), TAUS Data, a platform for sharing language data (https://www.taus.net/), LetsMT! (https://www.letsmt.eu/Systems.aspx), LIDER (http://www.lider-project.eu/), FALCON (http://falcon-project.eu/), 'localization web', PANACEA, http://www.panacea-lr.eu/, a factory of Language Resources (LRs), TTC (http://www.ttc-project.eu/) - Terminology extraction, translation tools, CESAR: http://cesar.nytud.hu/about/ EUROTERMBANK: www.eurotermbank.com, JRC – Joint Research Centre (https://ec.europa.eu/jrc/) (https://ec.europa.eu/jrc/en/language-technologies/).

3. User study

Apart from identifying existing catalogues and existing resources, the consortium has conducted a limited user study in the language resource (LR) user base of EU stakeholders through interviews. Dialogue Days in Brussels June 2015 and a workshop in Vienna July 2015 contributed to this purpose as well.

The general concerns and facts about the availability, quality and usability of LRs for the commercial sector are shared among the user base at large – *free (in the sense of freely accessible, not necessarily without costs, but with a reasonable cost), good and usable resources are needed*. And obviously the perspective of identifying and making available LRs for commercial & administrative users will have to start with the *user situation,* i.e. to enable potential end users of LRs to access precisely those LRs that fit their purpose. So, from the existing catalogues only relevant resources should be identified.

The following domains were deemed important by the stakeholders: **Construction**, **Healthcare**, **Media monitoring**, **Procurement**, **Legal and financial**. But all relevant domains are taken into account and the dialogue with users is continuing.

Input from the user study: Obtaining Language resources

We have collected observations from the stakeholder base under a number of headlines which we will further elaborate in the full paper. *LRs for/from research, in commercial contexts, Multilingual or monolingual LRs?*

Input from the user study: perspectives on Evaluating LRs

There is currently no clear answer from industry about any shared method for evaluating the *quality* of LRs. There are many aspects of the quality of language resources, and often quality and usefulness for a specific task are interrelated. So the question is: What *can* be evaluated? What *should* be evaluated?

One of the aspects that came up is the *Time frame* of the LR – old lexicons and lists (5, 10, 15 years old) of out-of-date technical terminology will not be relevant to much MT usage today and tomorrow.

It is widely thought that LR evaluations using some simple list of key parameters for ratings could be crowd-sourced from the user community. But it will inevitably be time-consuming and partial. This is why in commercial contexts LSPs ask for all the data from their customers and then see what works which is not the ideal solution.

It can be noted that ELRA has a full *Validation* procedure for LRs. What is measured is adherence to the standards used, exhaustiveness etc. see e.g. http://elra.info/en/services-around-lrs/validation/standards-best-practices/: This does not address the value of a language resource for a specific purpose, but it addresses the soundness of the resource as such. This seems to be as far as quality can be measured.

As in our work we are also relying on existing and acknowledged catalogues, we have assumed that their quality criteria are acceptable also to our users. [1]

User comments on metadata
Users (e.g. LT developers, LSPs, end users using a MT system) usually say they need very little metadata apart from the source, the date, the languages used, plus any unusual encoding information.

User comments on Clean Data: Users want clean data! No formatting codes etc., just plain text or xml.

User comments on In-domain data: The "more data is good data" approach is now commonplace. But the real question for users is whether or not a given LR is "in-domain" for the translation task in question.

4. Defining LR Usability

As we have seen, usability is a complex but vital criterion for evaluating and using LRs in the context of a one-stop shop catalogue service for the community (one of the objectives of the LT_Observatory project), where relevant information, time and quality are key values for decision-making.

The aspects of usability on which users are focusing, are: ease of download, domain relevance, language pair, availability/cost, time to implementation. Many of these aspects have already been covered above.

It is important that usability in this context is understood as being a criterion involved in specifically human decision making. It is entirely possible that language resources will in some future development be selected automatically by digital services operating as part of a broader and deeper language and translation infrastructure, as indicated by some of the EC-funded projects (LIDER and FALCON) listed above.

Here we mention a few important aspects of usability. Language/language pair and domain are of a different nature. **Ease of download:** This refers to the simplicity (number of clicks) of finding the relevant LR on some dedicated LR website. This is a very important aspect of usability. **Availability/cost:** The cost of an LR is obviously an important factor. However, it has to be taken into account that quality comes with a cost, so if a resource has been validated, e.g. by the private sector, there may be a fee to pay, and this may be a good investment for the user.

Above we have given a rather open description of the user needs as seen in our discussions with stakeholders. All of the main aspects mentioned have been taken into account in the methodology chosen for the resource collection.

5. Metadata discussion

It is essential that the LTO catalogue provides users with easy access to the resources they need. This means that the search options of the LTO catalogue must accommodate exactly the information types (metadata) that users are interested in. Therefore the metadata categories selected for the LTO are based on the user study, the previous usability check list as well as on experiences from similar projects.

Similarly, it is important that the language resources selected from various existing catalogues for inclusion in the LTO catalogue are exactly the resources that users need for MT-purposes. They must therefore be chosen by means of carefully prepared selection criteria.

We have compared the use of metadata in three major projects: LetsMT!, CLARIN VLO and META-SHARE.

6. LTO metadata categories

Building on the knowledge described in section 5, combined with the user input, we have agreed on a minimal list of metadata: Title (of the resource), Type of resource (corpus, terminology, lexicon etc.), Creator, Language(s), Availability (available for commercial use, price) , Modality (written for the time being), URL, Domain, Format (e.g. plain text), Size (in words, or any other measure), Production date, Comment (here additional information can be stored).

7. Selection criteria for collection of language resources

Based on user input and on experience from previous projects, the below selection criteria have been used as a framework for extraction of useful language resources.
The list below shows the most important selection criteria with a rating from 0-2 where 0 is the least

[1] For further investigation of the evaluation question we also collaborate with sister projects where applicable. But in the first instance we see that users want to evaluate if a resource is useful for them.

accepted/desired value. (ratings deleted in this short version)

Availability,
Languages covered
Longevity/date:
Validation
Modality
Ease of download

8. The collection process

The methodology for collection of language resources relies on three main features: 1) the identification of existing relevant language resource catalogues and projects as described in *2 Related work* and 2) the list of selection criteria which is based on the user study as well as experiences from related projects and 3) input from partners' network.

9. Preliminary Results

During this phase and using the methodology described above, information about around 100 resources have been collected. According to the selection criteria (building on users' expressed needs) only information about resources that are available for commercial use (or general research use) and that are either free or with a moderate licence fee will go into the LTO portal. The collection is continuing, and we hope that it will bring renewed focus on those useful resources that exist across the various catalogues and repositories.

The most important conclusion is that many resources exist which are useful for MT and similar work, but the majority are for (academic) research or educational use only, and as such **not available for commercial use**.

If companies have collected useful language resources for their own purposes, this is an asset that they do not easily share with their competitors. During the user study companies expressed that if they need LR they search the web and use what they find, this is often the fastest and cheapest way. It should be noted that it may happen that some of the resources collected this way are actually not available for commercial or any other use because of IPR problems. And anyway, we believe the existing resources should be brought to use.

10. Languages covered

An investigation of the languages covered by the language resources identified until present was made:

Language list	No of resources	Language list	No of resources
Bg	11	**it**	17
Bs	1	**lt**	16
Ca	1	**lv**	16
Cs	12	**mk**	1
Da	13	**mt**	7
De	32	**nl**	13
El	20	**no**	4
En	87	**pl**	14
Es	36	**pt**	20
Et	16	**ro**	15
Fi	15	**ru**	1
Fr	52	**sk**	10
Ga	6	**sl**	14
Gl	4	**sq**	1
Hr	15	**sr**	1
Hu	13	**sv**	15
		tr	1

TABLE 1 COVERAGE OF LANGUAGES IN THE RESOURCES IDENTIFIED.

11. Future work

In the next 4 months the following tasks will be undertaken
Methodology update: The methodology was made building on user studies and previous experience. Still to be modified.
More resources will be featured.
Valorisation: One of the next steps in the process is to select LRs that can be of better value if modified. One of the actions will be to add or modify domain information.
Discussion with users will continue, next workshop will be held November/December, and will focus on how users rate quality of resources.
Creation and population of the portal: first version available December.

12. References

A proper list of references will be given, this is just a list of the catalogues and projects we have been using
http://www.elra.info/en/catalogues/
http://clarin.eu/content/virtual-language-observatory),
http://www.meta-net.eu/meta-share
http://opus.lingfil.uu.se/
https://www.taus.net/
https://www.letsmt.eu/Systems.aspx
http://www.lider-project.eu/
http://falcon-project.eu/
http://www.panacea-lr.eu/,
http://www.ttc-project.eu/
http://cesar.nytud.hu/about/
www.eurotermbank.com
https://ec.europa.eu/jrc/
https://ec.europa.eu/jrc/en/language-technologies/.