LT-OBSERVATORY

LT_OBSERVATORY – OBSERVATORY FOR LR ANDMT IN EUROPE

Acronym: LT_OBSERVATORY

COORDINATION AND SUPPORT ACTION

INFORMATION AND COMMUNICATION TECHNOLOGIES

D1.4 Analysis of Coverage Situation

GRANT AGREEMENT	644583
DELIVERABLE NUMBER	D1.4
DELIVERABLE TITLE	Analysis of Coverage Situation
DUE DATE OF DELIVERABLE	1st version 28/02/2016 – update 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		☐

TABLE OF CONTENTS

DOCUMENT INFO

AUTHORS

Name	Company	E-mail
Bente Maegaard Sussi Olsen Lina Henriksen	CLARIN	bmaegaard@hum.ku.dk saolsen@hum.ku.dk linah@hum.ku.dk
Vesna Lusicky	UNIVIE	vesna.lusicky@univie.ac.at
Philippe Wacker Andrew Joscelyne	LTI	phw@ lt-innovate.eu aj@lt-innovate.eu

REVIEWERS/CONTRIBUTORS

Name	Company	E-mail
Margaretha Mazura	EMF	mm@emfs.eu
Blanca Rodriguez / Luz Esparza	ZABALA	brodriguez@zabala.es /mlesparza@zabala.es
Partners	All partners	

DOCUMENT CONTROL

Document version	Date	Change
D1.4.1	15/02/2016	Initial draft version to be completed
D1.4.2	24/02/2016	First internal draft
D1.4.3	26/02/2016	Comments from partners
D1.4.4	28/02/2016	Fine-tuning; typo corrections
D1.4_V1	29/02/2016	Final version by the consortium to be submitted to the EC.
D1.4_V2	23/12/2016	Final updated version

1. SUMMARY

From the project summary:

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.

Report summary:

The purpose of this report is to analyze the coverage situation with respect to parallel corpora and the suitability of terminology resources for MT purposes. An optimal coverage scenario could be that all language pairs were covered for all domains by a minimum set of e.g. 3 language resources that all had a suitable set of metadata and were available for all potential users. This is obviously not the situation yet, and in the following we will analyse the situation with respect to coverage, primarily in relation to language pairs and domains. The size of language resources is also an important feature. The size is taken into account for terminology resources, but for corpora it has not been easy, as size is sometimes expressed in number of words, sometimes in MB. Quality is also of importance, but has been left out in the discussion as only acknowledged sources are used.

We have collected close to 150 relevant resources as reported in D1.1 and D1.2. Relevant resources are those that are (or could at a low price/ work effort become) operationally usable for MT applications. One experience from our collection efforts is that there are much fewer resources than seems to be the case at first sight. As it turned out there is a considerable duplication of resources among the openly available EU corpora – different repositories often hold the same language resources.

More importantly, the analysis shows that there are huge coverage gaps, both in terms of languages and in terms of domains, and this is true for corpora as well as for terminology.

2. RESOURCES

The identified resources can be divided into parallel corpora (56), comparable corpora (9), monolingual corpora (12), thesauri (3), speech corpora (4), glossaries (2), terminological resources (25), tools (15), wordnets (16), treebanks (4), and translation memories (2).

In Figure 1 we show the amount of resources collected for each type. The figure shows that the collection effort has concentrated on parallel and comparable corpora, and on terminology.

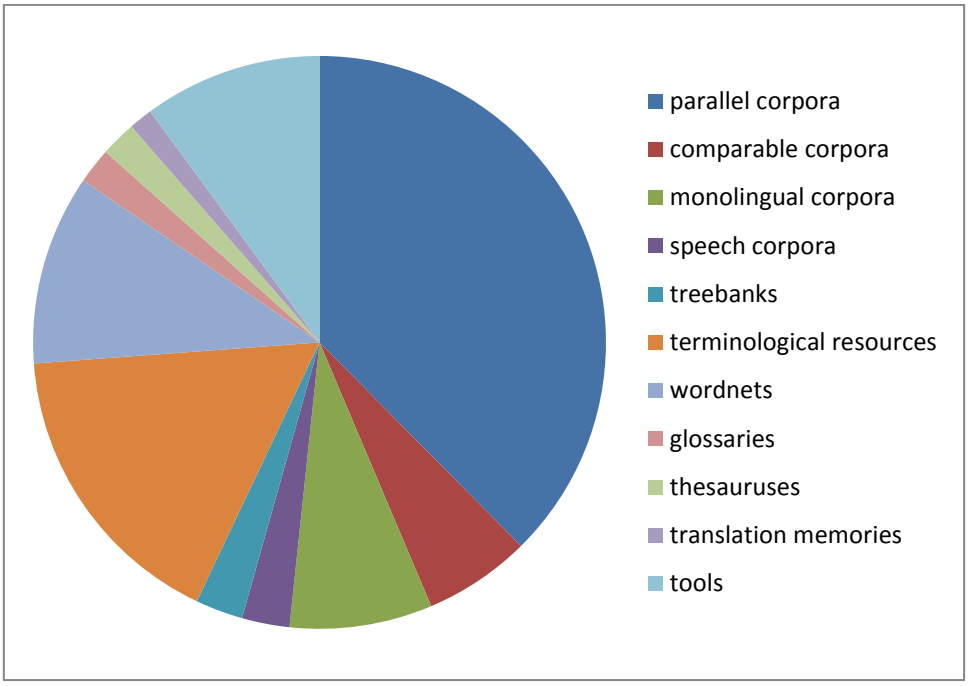

FIGURE 1. AMOUNT OF RESOURCES PER RESOURCE TYPE

In the analyses of the corpora below, we have focused on official and regional EU-languages. Other European languages and languages from other parts of the world form part of some resources.

3. CORPORA

The analyses of corpora focus on languages and domains. The size and format of a corpus are also interesting parameters, but size of the collected corpora is expressed in many different ways (words, sentences, MB), or not given at all. Similarly for format specification: many of the corpora have no format specified in their metadata. Therefore an analysis of format and size is not feasible in this context.

In the following we will focus on parallel corpora, comparable corpora, monolingual corpora and terminological resources.

3.1 PARALLEL CORPORA

Parallel corpora are here corpora where the same text appears in more than one language. In Figure 2 we give for each language the number of corpora in which it appears.

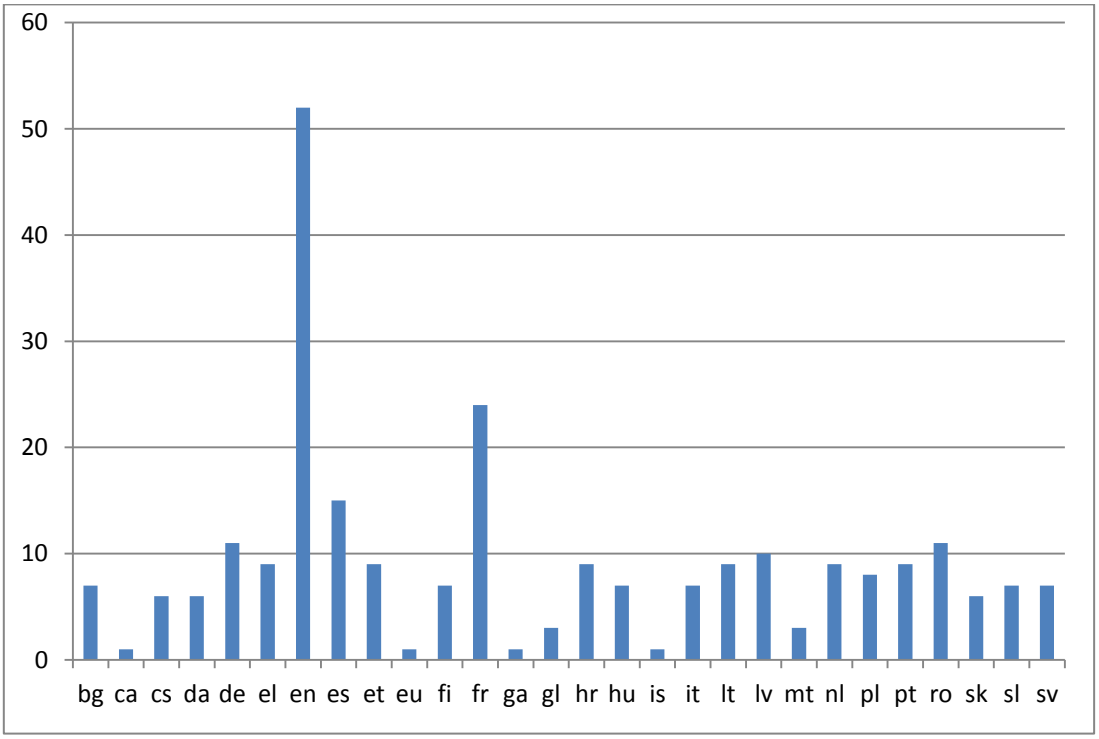

FIGURE 2.NUMBER OF PARALLEL CORPORA PER LANGUAGE

3.1.1 LANGUAGE PAIRS

The most frequent language by far is English as it appears in 45 parallel corpora out of 56 in all. Other frequent languages are French, Spanish, German and Romanian appearing in 24, 15, 11 and 11 corpora respectively. These languages in different combinations also constitute the most frequent language pairs. It should however be noted

that the corpora where these languages appear together are mostly multilingual - and the languages are therefore not language pairs in the sense where a source and a target language can be identified.

Other languages with a medium frequency are Croatian, Dutch, Estonian, Greek, Latvian, and Portuguese and as they appear in 9-10 corpora each.

Out of the 56 parallel corpora 37 are bilingual (the rest are multilingual) and the majority of these have English as one of the languages. French is the language of 9 bilingual corpora with Arabic, and of the remaining bilingual corpora Croatian, Portuguese, Greek, Hungarian and Latvian are among the languages that constitute either the source or target language together with English.

Among the languages with the poorest coverage are Maltese, Danish, Swedish and Czech. Many languages with poor representation are primarily part of multilingual corpora as e.g. *CESAR Multilingual Corpora*, the *Wikipedia Parallel Titles* and *OpenSubtitles 2013*. This means that even if these languages are represented in very few corpora and only together with very few domains, they still combine with many other languages.

Figure 3 is an example showing the languages that appear together with Danish (Danish is represented in only 6 corpora). The figure also shows the number of corpora for each language.

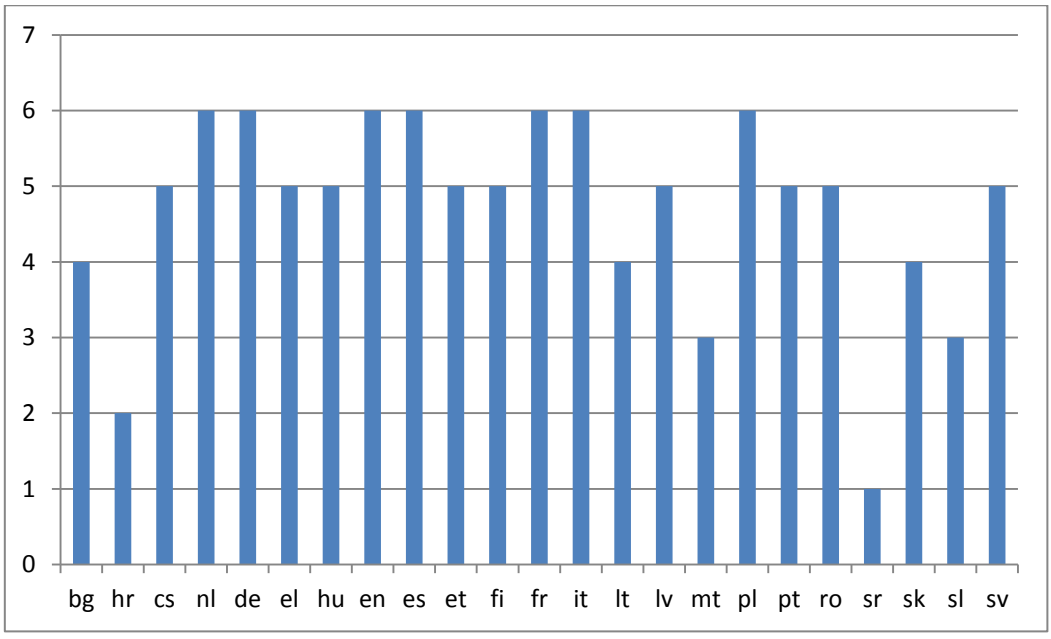

FIGURE 3. LANGUAGES COMBINED WITH DANISH AND THE NUMBER OF CORPORA FOR EACH LANGUAGE

A major language such as German is not very well represented in the collection. German is only represented in 11 corpora, but again mainly in large multilingual corpora and therefore it combines with a lot of languages.
Figure 4 shows the EU and regional languages that German combines with and the number of corpora for each language. Apart from these languages German also combines with a long list of languages from other parts of the world.

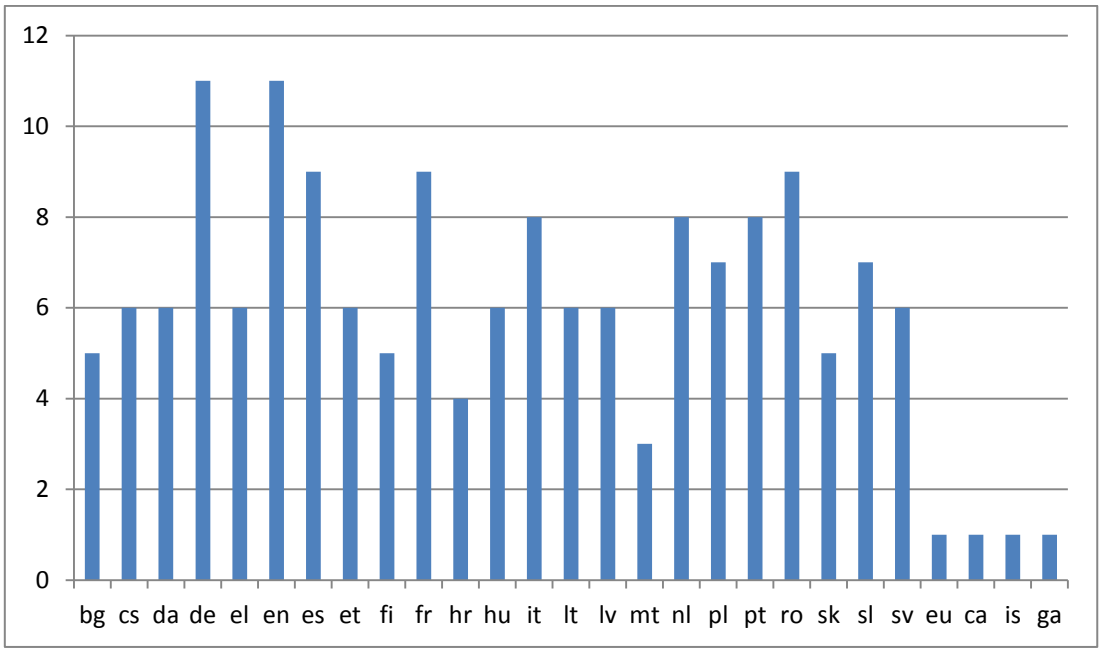

FIGURE 4. LANGUAGES COMBINED WITH GERMAN AND THE NUMBER OF CORPORA FOR EACH LANGUAGE

3.2 COMPARABLE CORPORA

In the collections that were examined in search for usable LRs we found 9 comparable corpora. 4 of these are bilingual[1] while 5 are multilingual including between 5 and 9 languages.

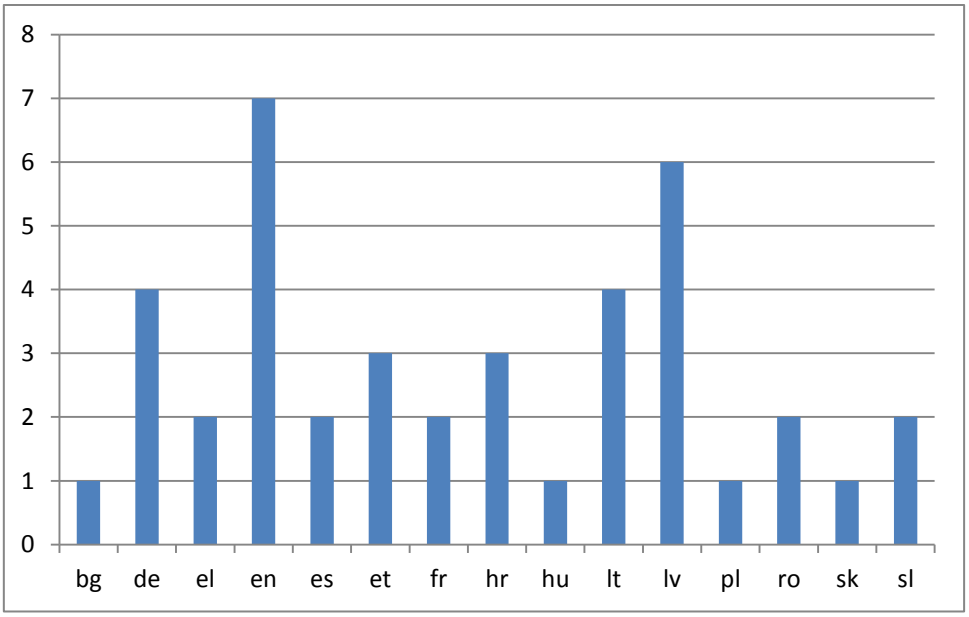

FIGURE 5. DISTRIBUTION OF COMPARABLE CORPORA PER LANGUAGE

As can be seen in Figure 5, many of the 15 languages in the comparable corpora are East European languages. Only two of these resources do not have English as one of the languages.

[1] Aligned comparable sentences with alignment confidence score for each sentence pair.

3.3 MONOLINGUAL CORPORA

The 12 monolingual corpora that form part of the resource collection are distributed as can be seen below.

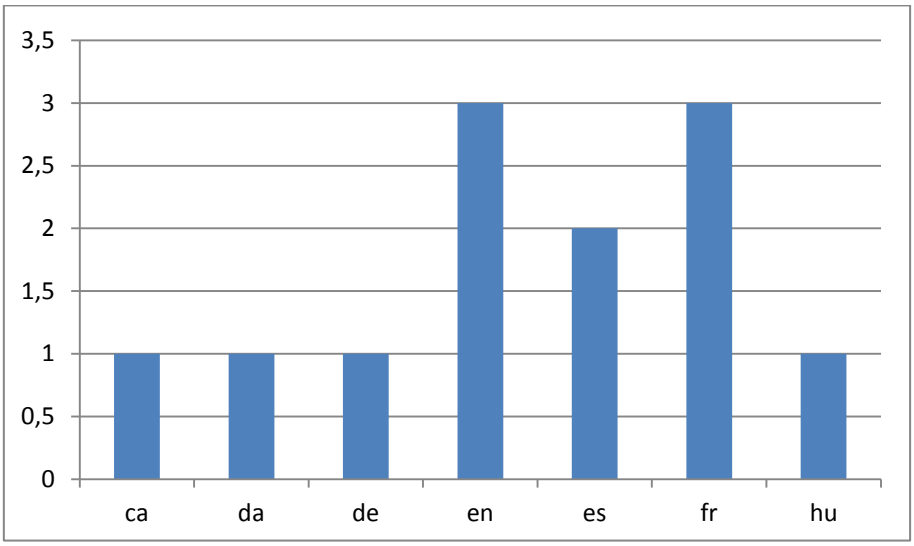

FIGURE 6. MONOLINGUAL CORPORA

As mentioned in D1.1, the project decided to focus on collecting information about parallel corpora as these are very important for MT, and are normally quite difficult to have access to. The small number of monolingual corpora is therefore not representative of the situation with respect to existing monolingual corpora. Consequently we are not analyzing the situation.

3.4 TREEBANKS

The 4 treebanks collected comprise the languages English, French, Spanish and Italian, one of these is bilingual en-es.

3.5 DOMAINS

The subject domains covered by the parallel and comparable corpora comprise 19 different domains as given in:

LTO Domains
administration
building foam and sealant
computer science
economy
education
environment
fiction
health/medicine
labour legislation
law/legal
law_politics
mobile technology
newswire
politics
popular science
renewable energy
tourism/ travelling
wind power
wikipedia

FIGURE 7. LIST OF DOMAINS FOUND IN PARALLEL AND COMPARABLE CORPORA

The domains are of very different character, some are very broad e.g. health/medicine and some are very narrow e.g. building foam and sealant, and a domain like 'popular science' is a mixture of communication level 'popular' and domain 'science'. This difference is due to the fact that the descriptions of the resources in their home repositories are not based on a classification standard (or if they are, not the same standard for different resources). The lack of classification standard is also the reason why not all domains are listed here. Some domains that occur in only once have not been included in this overview.

Figure 8 shows the distribution of domains in the parallel corpora for each language. The first column (Language code) specifies the language and the second (Number of parallel corpora per domain) lists the domains of each

language with the number of corpora for each domain. It should be noted that some corpora have no domain information at all - and some corpora are denoted by many different domain names. Therefore the total number of resources for each language (Fig. 2) does not correspond with the number of resources per domain in the below table.

Lang code	Number of parallel corpora per domain
bg	administration 1, health/medicine 1, law_politics 1, newswire 2, politics 1
hr	newswire 4, renewable energy 5
cs	administration 1, health/medicine 1, law_politics 1, politics 1
da	administration 1, health/medicine 1, law_politics 1, politics 1
nl	administration 1, health/medicine 1, law_politics 2, politics 1
en	administration 4, computer science 3, economy 1, education 2, environment 3, fiction 2, health/medicine 4, labour legislation 2, law/legal 6, law_politics 1, newswire 4, politics 2, popular science 3, renewable energy 1, tourism/travelling 4
es	administration 4, computer science 3, economy 1, environment 1, fiction 2, health/medicine 2, law/legal 3, law_politics 1, politics 2, popular science 2
et	administration 1, building foam and sealant 1, health/medicine 1, law/legal 1, law_politics 1, politics 1
de	administration 1, health/medicine 1, law_politics 1, politics 1
el	administration 1, education 1, environment 1, health/medicine 3, labour legislation 1, law/legal 1, law_politics 1, newswire 2, politics 1, tourism/travelling 1
fi	administration 1, health/medicine 1, law/legal 1, law_politics 1, politics 1
fr	administration 3, computer science 1, fiction 1, health/medicine 1, labour legislation 1, law/legal 1, law_politics 2, politics 2, popular science 1
hu	administration 1, health/medicine 1, law_politics 1, politics 1
it	administration 1, health/medicine 1, law_politics 1, politics 1
lv	administration 1, building foam and sealant 1, health/medicine 1, law_politics 1, politics 1, renewable energy 1
lt	administration 1, health/medicine 1, law_politics 1, politics 1, renewable energy 1
mt	health/medicine 1, law_politics 1
pl	administration 1, health/medicine 1, law_politics 1, politics 1
pt	administration 2, health/medicine 1, law_politics 1, politics 2, tourism/travelling 2
ro	administration 1, health/medicine 1, law_politics 1, newswire 2, politics 1, renewable energy 1
sk	administration 1, health/medicine 1, law_politics 1, politics 1
sl	administration 1, health/medicine 1, law_politics 1, politics 1
sv	administration 1, health/medicine 1, law/legal 1, law_politics 1, politics 1
ga	
gl	administration 3, computer science 3, fiction 3, law/legal 3, popular science 3

Lang code	Number of parallel corpora per domain
ca	
is	
eu	

FIGURE 8. NUMBER OF PARALLEL CORPORA FOR EACH DOMAIN FOR THE OFFICIAL AND REGIONAL EU LANGUAGES

English is the only language with resources in all 17 domains, with more than one resource for the majority of the domains. The other two most frequent languages Spanish and French also have a broad coverage of domains. Other languages concentrate on few domains, e.g. the 9 Croatian resources are all in the domains newswire and renewable energy.

Health/medicine is represented for all languages except for Croatian; politics, administration and law are also covered for many languages, whereas building foam and sealant is found in only one resource, Estonian-Latvian.
In absence of a domain classification system that would be used in the home repositories of the resources collected, we decided to use the EuroVoc (Multilingual Thesaurus of the European Union) for this report. By mapping the provided domain descriptors to the top domain categories in EuroVoc, we could establish a basis for comparative analysis of the domain coverage. EuroVoc is a multilingual, multidisciplinary thesaurus, containing terms in 23 EU languages (plus Albanian, Macedonian and Serbian). Although it is mainly geared towards the activities of the EU, the top categories cover main private and public endeavors (e.g. law, transport, production, agriculture, social questions etc.). The advantages of EuroVoc are multilingualism (language equivalences between identical concepts are expressed in different languages), and conformance with W3C recommendations. EuroVoc users include the European Parliament, the Publications Office, national and regional parliaments in Europe, national governments and private users around the world.

A comparison of the domains in the LTO corpora with Eurovoc clearly shows the difference in the character of the domains, cf. Figure 9. The administration domain spreads over at least 4 different subdomains or microthesauri of Euorovoc, other domains such as politics are equivalent to a single top domain in Eurovoc while others again are equivalent to a subdomain e.g. health/medicine -> 2841 Health.

Domains LTO	Eurovoc
administration	0436, executive power and public service, public administration
	MT 0806 international affairs, international organisation, UF international administration
	1006 EU institutions and European civil service, administration of the Institutions
	4006 business organisation, business policy, NT1 business administration

Domains LTO	Eurovoc
building foam and sealant	MT 6831 building and public works, building materials
computer science	MT 3236 information technology and data processing
economy	16 Economics
education	MT 3206 education
environment	52 Environment
fiction	MT 2831 culture and religion
health/medicine	2841 health
labour legislation	MT 4426 labour law and labour relations
law/legal	12 Law
law_politics	
mobile technology	MT 3226 communications
newswire	MT 3226 communications
politics	04 Politics
popular science	36 Science
renewable energy	MT 6626 soft energy, renewable energy
tourism/travelling	2826 social affairs, leisure, NT1 tourism
wikipedia	
wind energy	6626 soft energy, NT1 wind energy

FIGURE 9. LTO DOMAINS WITH EUROVOC CLASSIFICATION

If we look at the 21 top categories of Eurovoc, the domains covered in LTO belong to 12 of these top categories. The top categories not mentioned in the resource metadata are 20 Trade, 24 Finance, 48 Transport, 56 Agriculture, Forestry and Fisheries, 60 Agri-foodstuffs, 64 Production, Technology and Research, 72 Geography, 76 International Organisations.

The monolingual corpora only comprise 3 domains: newswire, environment and labour legislation. Furthermore a corpus of general language and one of children language have been collected.
The treebanks are all from the newswire domain except one for which domain is not specified.

4. TERMINOLOGICAL RESOURCES AND THESAURI

4.1 LANGUAGE COVERAGE

The terminological resources and thesauri collected in the scope of LTO are bilingual and multilingual. There are no monolingual resources. Roughly a third of the resources are bilingual (as depicted in Figure 10), all but one of them covering either the language combination French-English or Greek-English. More than half of the LTO resources are multilingual (55 percent).

English is the only language represented in all of the terminological resources, closely followed by French (91 percent of the terminological resources). German, Greek, Spanish, Italian, Portuguese, and Swedish are represented in more than 10 terminological resources. Finnish, and Polish are covered in 10 resources. Bulgarian, Czech, Danish, Estonian, Irish, Croatian, Hungarian, Latvian, Lithuanian, Maltese, Dutch, Romanian, Slovak and Slovenian are represented in less than 10 terminological resources in the LTO collection (see Figure 11).

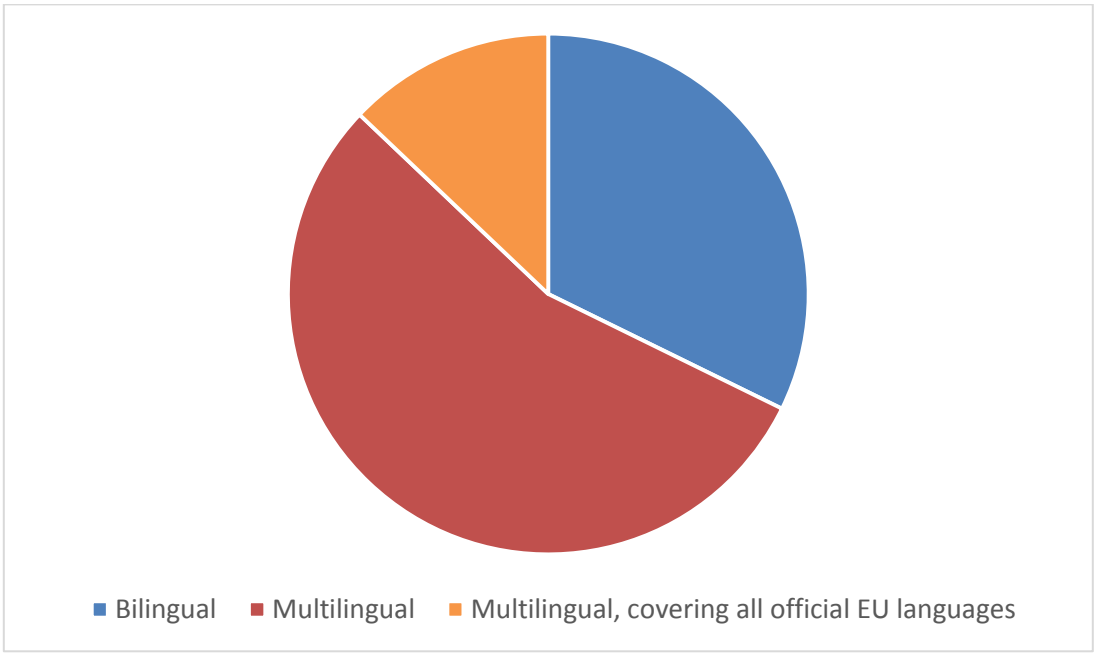

FIGURE 10: DISTRIBUTION BY THE TYPE OF RESOURCE (BI-, MULTILINGUAL RESOURCE)

In addition to the official EU languages, the languages in the multilingual terminological resources are languages that are not the official languages of the EU (Norwegian, Russian, Chinese, Japanese, Latin), and minority languages (Basque, Catalan).

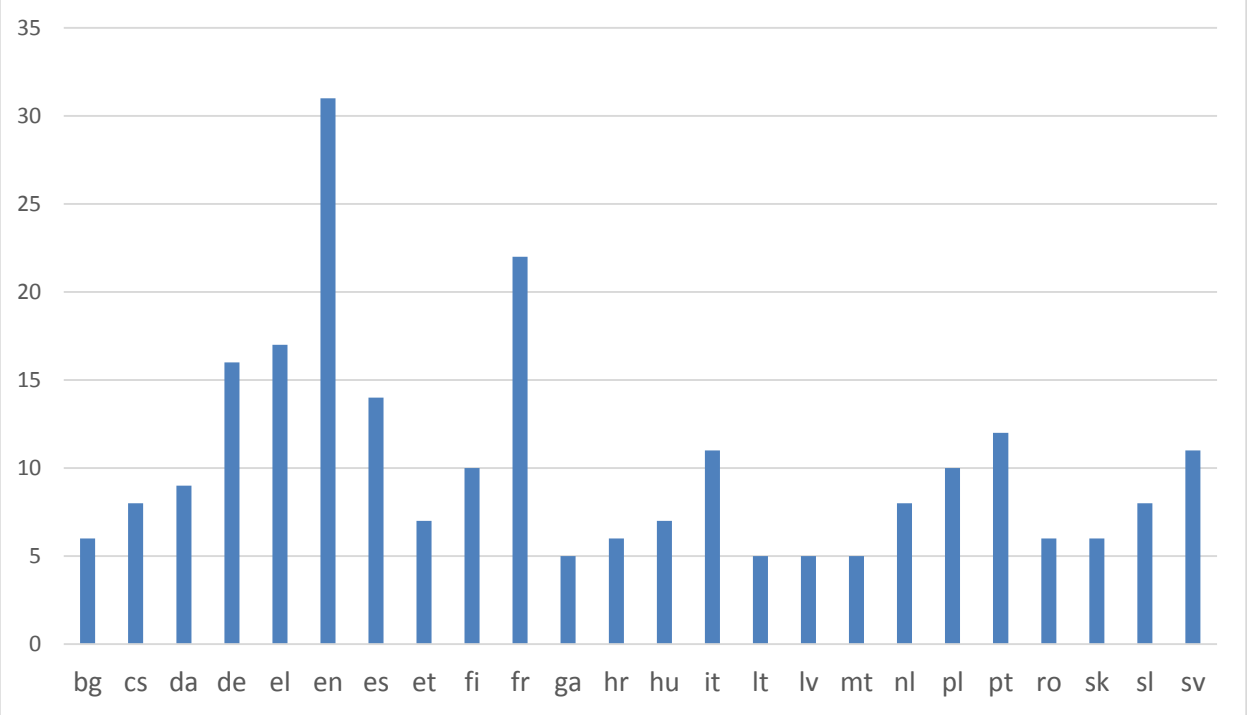

FIGURE 11: DISTRIBUTION BY LANGUAGE

4.2 DOMAINS

Approximately one quarter of the terminological resources in the LTO collection cover only one domain (environment, trade, labour legislation, biotechnology). When terminological resources cover more than one domain, additional steps to filter out the individual domains may be needed. The feasibility and complexity of this processing depends on the termbase definition model and the standard of the terminological resource.

The descriptions of the terminological resources in their home repositories are either not based on one single classification standard or they are not based on any classification standard at all. Thus the descriptions show great differences in granularity of the domain description (law vs. labour law). When mapping the provided descriptions of the terminological resources against the classification in the Eurovoc, all of the EU official languages are represented by at least one of the 21 top domain categories in Eurovoc. This is mainly ensured by large resources (e.g. IATE), containing a great number of languages and domains. In such cases, additional steps would be needed to filter out the languages and domain in question. In order to map the domains down to subcategories in Eurovoc, an extensive systematic analysis and mapping of the resources would be needed.

As depicted in Figure 12, Environment and Production, technology and research are the most widely covered domains (12 resources each), followed by Agriculture, forestry, fisheries (11), Science (11), Law (10), Politics (10),

and Education and communication (10). The least represented domains are International organization (5 resources), International relations (4), and European Union (4). However, in cases in which one resource covers more than one domain, it is not clear if the domains and languages are evenly represented in the resource in question without further deeper analysis of each resource.

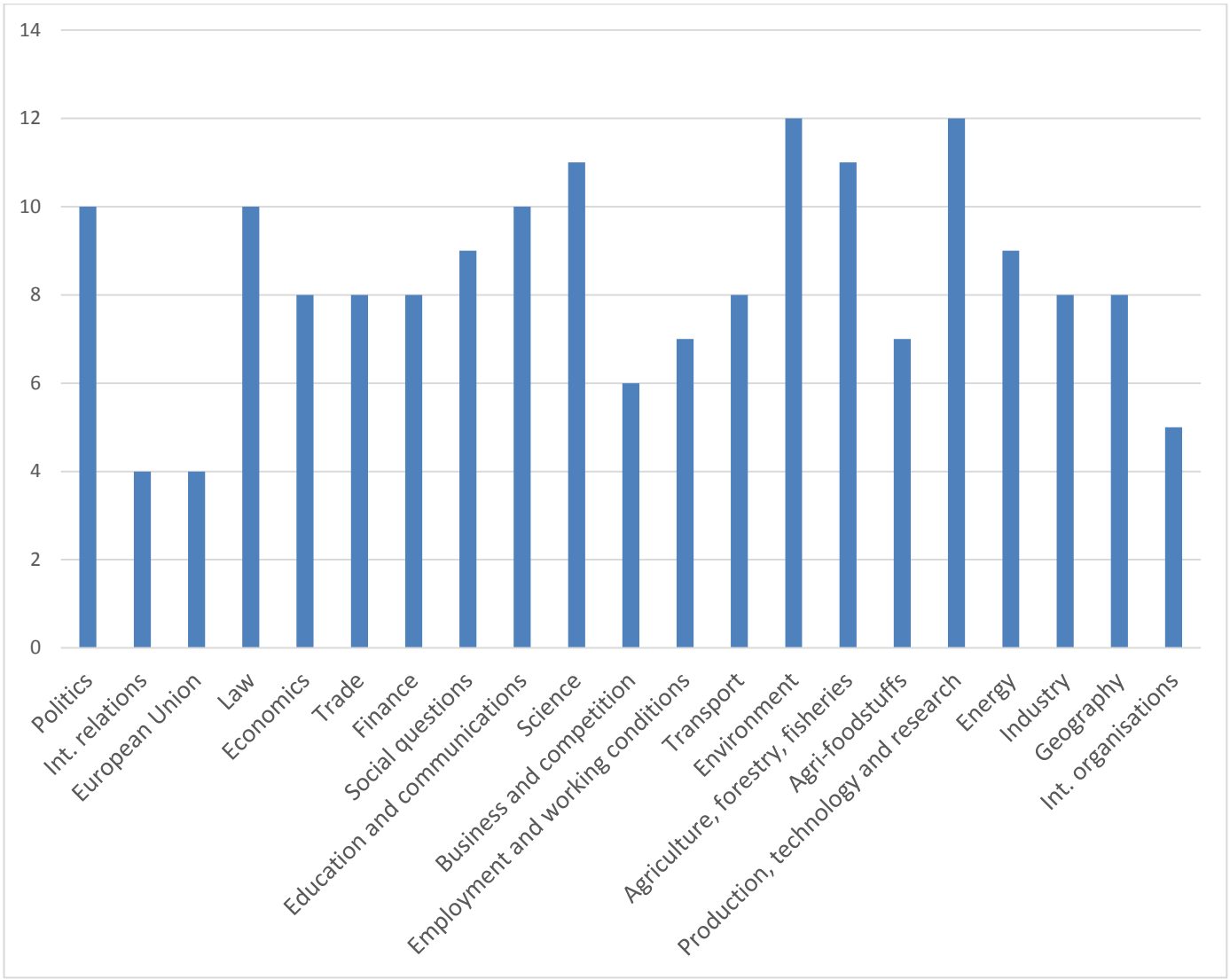

FIGURE 12: DISTRIBUTION BY DOMAIN

4.3 METADATA

Apart from language and domain, the resources in the LTO collection are described by additional metadata harvested from their home repositories (compare D1.1 and D1.2). The size of the terminological resources may be

an important factor to decide which resource to use. The descriptions of the size of the terminological resources in their home repositories are either not based on one standard or the size of the resource is not provided at all (see Figure 13). The size of the terminological resources across the collection is thus not comparable.

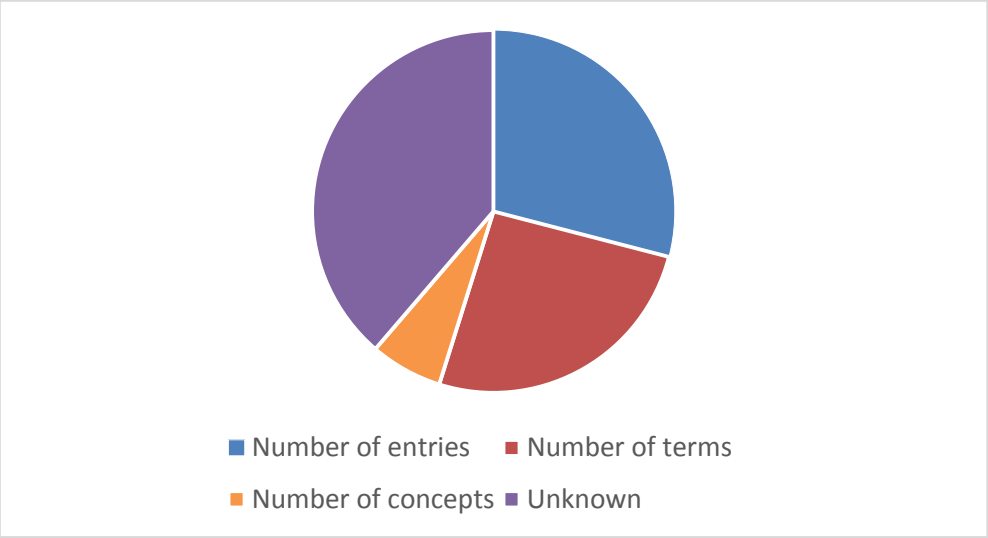

FIGURE 13: TYPE OF VALUE GIVEN FOR THE SIZE OF RESOURCE

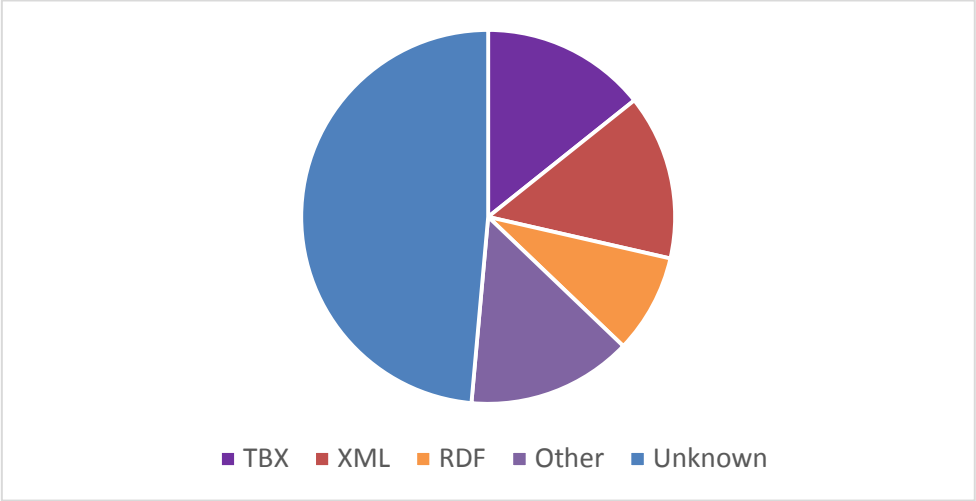

FIGURE 14: FORMAT TYPE

In order to ensure re-use, further processing or integration in tools, the information on the format type of the resource is important. The distribution of the most commonly used standard formats in the collected resources falls as expected fairly even between TBX (TermBase eXchange) and XML, followed by RDF. However, almost a

fifth of the collected resources use other, not so commonly used formats for terminological resources. The format type is unknown in almost half of the terminological resources in the collection (see Figure 14). This information, as well as any further required information, may be obtained directly from the resource provider if contact information is provided, which is not the case in 10 percent of the collected resources (see Figure 15).

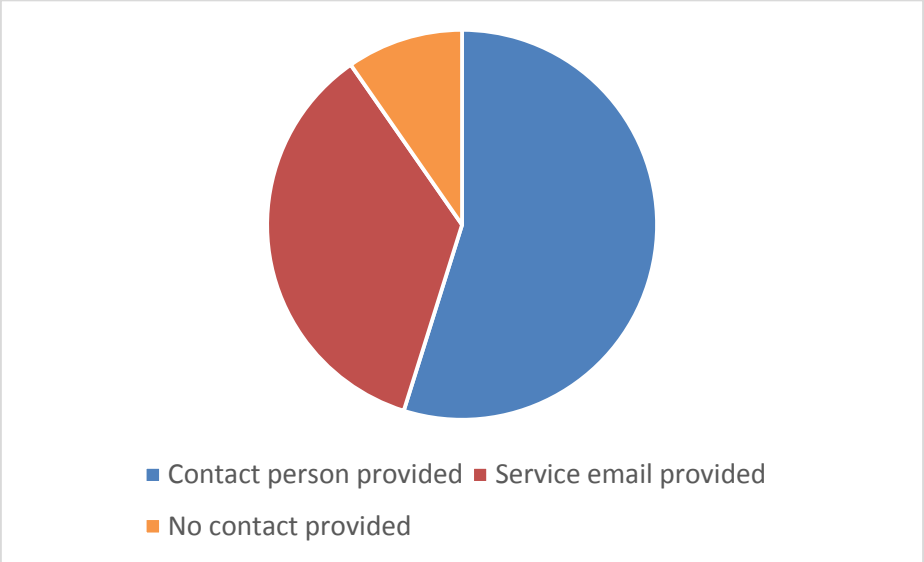

FIGURE 15: CONTACT TYPE

5. CONCLUDING REMARKS AND NEXT STEPS

The work done in the Language Technology Observatory project has revealed a number of gaps. Below we are going to draw some conclusions in terms of coverage of languages and domains.

Corpora

Only English has a good coverage in terms of combinations with other languages. Spanish, German, French, Latvian, Romanian, Croatian, Polish and Lithuanian are moderately covered in relation to other languages. Maltese, Danish, Czech and Slovak are poorly covered. All languages, including English have gaps in relation to domains. Eurovoc top categories not mentioned in the resource metadata for any language are: Trade, Finance, Transport, Agriculture Forestry and Fisheries, Production, Technology and Research, Geography, International Organizations. Besides, many subdomains within many other top categories are not represented in any or only in very few languages.

Below we compare the META White Papers' support level with the support level we have arrived at in the LTO project.

The META-NET White Papers (2012) provide the overview given below in Figure 16[2].

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech	Basque	Icelandic
		Dutch	Bulgarian	Irish
		French	Catalan	Latvian
		German	Croatian	Lithuanian
		Hungarian	Danish	Maltese
		Italian	Estonian	Welsh
		Polish	Finnish	
		Spanish	Galician	
		Swedish	Greek	
			Norwegian (Bokmål,Nynorsk)	
			Portuguese	
			Romanian	
			Serbian	
			Slovak	
			Slovene	

FIGURE 16. META-NET: SPEECH AND TEXT RESOURCES. STATE OF SUPPORT FOR 30 EUROPEAN LANGUAGES

[2] http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison

In Figure 16 we cannot distinguish between spoken and written resources, but nevertheless we can see that even if the overall picture is similar - English is still reasonably well served as the only language - there are also some developments: The moderately supported group is different today as Latvian, Romanian, Estonian and Lithuanian are included whereas Czech and Swedish are in the fragmentarily supported group.

Terminology

In terms of the number of resources in which a language is represented, English, followed by French, is the most represented in terminology resources, as both languages are often used as pivot languages. German, Greek, Spanish, Italian, Portuguese, and Swedish are moderately represented in terms of the number of resources in which they can be found. Finnish and Polish are less respresented in terminological resources. Bulgarian, Czech, Danish, Estonian, Irish, Croatian, Hungarian, Latvian, Lithuanian, Maltese, Dutch, Romanian, Slovak and Slovenian are the least represented languages in the terminological resources.

Nevertheless, the number of the resources covering a language should not be the sole criterion for coverage. The size of the resource in relation to the domains covered should also be considered. In this regards, smaller languages with national terminology infrastructures or terminology centres (e.g. Swedish, Catalan, Irish) have better coverage than only the number of available resources would suggest.

All of the languages have coverage in the domains analysed in this report, although the actual in-domain coverage varies greatly in size and granularity.

In order to sum up we suggest to consider gaps along several dimensions:

Awareness gap

There is clear evidence that a large part of the demand side is unaware of the offer. When the project made the first 30 pre-selected LRs available, it appeared that a large segment of potential users was not aware of their existence. Our initiative was therefore welcomed by these actors. This was also amply confirmed by the very interested and positive reception of Ralf Steinberger's presentation of the JRC's LRs at the LT-Accelerate conference on 23-24 November 2015. Clearly, there is a need to reach out to the demand side and to "market" the offer to it in a more proactive fashion.

Usability gap

Already at pre-selection stage, it became clear that very few LRs that are presently on offer in existing repositories correspond to minimum quality requirements on metadata. Moreover, it appeared that only VERY FEW LRs are made available by repositories in a way that enables their straightforward commercial use. The latter is often restricted in the first place; licensing conditions are not clearly spelled out; contact persons to obtain additional

information are not identified; where LRs are made available for a price, the usability-price relationship is often considered inadequate; etc. Hence, there is a need to improve the usability of existing LRs and to reflect seriously on the conditions at which they can/should be made available for commercial use (particularly when they have been compiled with the support of public money).

Quantity gap

For LRs to be useful in an operational context they need to be available in large quantities. It is obvious that the quantities available today (at the required quality levels) are largely insufficient to have a positive impact on the quality of MT in a commercial context. A large combined effort should be launched to produce new LRs across the board (and in all languages) that correspond to a set of agreed usability criteria. This would also (and URGENTLY) require a clarification of their copyright status in a commercial context. Furthermore, there is demand for in-domain resources, i.e. LRs that are clearly customised for use in specific domains (healthcare, finance, security, tourism, etc.). Whether the compilation of such in-domain resources should be left to private initiative as is currently the case or whether they should become part of a European Language Cloud should be, at the very least, debated seriously. There is evidence that only very few European commercial players will be able to compile such in-domain resources in sufficient quantities and qualities over the long-run to allow them to offer specialised domain-specific language clouds. Furthermore, the question whether such specialised language clouds should be left to private appropriation in the first place should also be debated.

Coverage gap

For corpora only English has a good coverage in terms of combinations with other languages. Spanish, German, French, Latvian, Romanian, Croatian, Polish and Lithuanian are moderately covered in relation to other languages. Maltese, Danish, Czech and Slovak are poorly covered. All languages, including English have gaps in relation to domains. Eurovoc top categories that are not represented in metadata of any language are: Trade, Finance, Transport, Agriculture Forestry and Fisheries, Production, Technology and Research, Geography, International Organizations. Besides, many subdomains within many other top categories are not represented in any or only in very few languages. For corpora we have consequently seen gaps for most of the languages, and for nearly all domains.

For terminological resources, English, French and German have a good coverage of the domains. Italian, Dutch, Spanish, Danish, Portuguese, Finnish, and Swedish show a moderate coverage of domains. The granularity of the domains covered varies greatly, and all languages have gaps in this regard.

Some future steps

As we have seen above, only one language (English) has a reasonable coverage in relation to volume as well as in relation to domains, and a very limited number of languages have a moderate support. The reason for this may be

that EU language resource identification and management has been a pretty random process, unsupervised for several years (decades) despite best intentions.

From now on, LR identification and operational management needs to be organized by means of a clear **strategy** of identifying, usability-checking and promoting all those LRs that can contribute to better MT productivity in the years ahead[3].

The technology that can help enable this provision of LRs is itself developing by automatic methods of creating parallel corpora e.g. from crawling the web. Best Practice methods for creating language resources is one of the themes of deliverable D1.3.

New methods of categorizing the meaning of words and sentences across multiple languages are opening up new opportunities for more effective resources for MT, so a bit further into the future, we need to explore how these results can also be used for improving the usability and the quantity of LRs for MT. And if possible, clearly align a given tranche of technology R&I with LR usability needs.

It will take several cycles for MT selection, use, and refinement to make the most of what exists. Tools will need to be developed that can

- Help the quality-checking of existing LRs.
- Boost LR creation (automatic methods, semantic categorization etc.)
- Help identifying the usability (domain relevance and quality) of any given resource/language pair etc.

LT Observe is taking the first step: simplifying access to usable (and usually free) translation data from public repositories in the EU via a one-stop access point. Once the LT Observatory catalogue is open for business, there should be a pilot study of usability, with feedback from users to improve the service for a second round of LR collection/invitations/pooling/crowdsourcing opinion.

[3] Later a plan will be needed for better productivity in other important areas, apart from MT