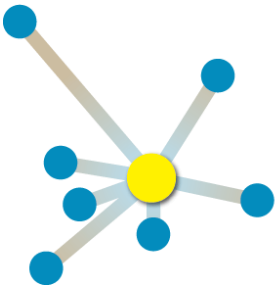
European Commission

ELRC
European Language Resource Coordination



European Language Resource Coordination

Connecting Europe Facility

FINAL REPORT

A study prepared for the European Commission
DG Communications Networks, Content & Technology by:

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

elDa IEA-LSP TILDE

Digital Single Market

This study was carried out for the European Commission by the ELRC Consortium

Internal **identification**

Contract number: 30-CE-0696785/00-64

SMART 2014/1074

DISCLAIMER

By the European Commission, Directorate-General of Communications Networks, Content & Technology.

The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the Commission. The Commission does not guarantee the accuracy of the data included in this study. Neither the Commission nor any person acting on the Commission's behalf may be held responsible for the use which may be made of the information contained therein.

Table of Contents

EXECUTIVE SUMMARY

The European Language Resource Coordination (ELRC) was a service contract under the Connecting Europe Facility's (CEF) Programme[1] stemming from a call for tender SMART 2014/1074 which covered the set-up of a permanent Language Resource Coordination mechanism.

The ELRC targets all CEF-affiliated countries, i.e. the 28 EU Member States plus Norway and Iceland. The overall goal of the ELRC is to collect language resources from and for public service administrations in all CEF-affiliated countries in order to improve the quality, coverage and performance of CEF eTranslation[2] in the context of current and future CEF digital online services (CEF DSIs) such as Online Dispute Resolution (ODR), Electronic Exchange of Social Security Information (EESSI), e-Justice, Safer Internet etc.[3] As such, all data collected by the ELRC will be used by the European Commission (EC) to support the development of CEF eTranslation and its adaptation to the relevant CEF digital services.

The ELRC had a run-time of 24 months (17th of April 2015 until 16th of April 2017). The main results achieved through the ELRC included:

225 Language Resources collected, validated and delivered: Overall, ELRC collected 138 bi-/multi-lingual corpora, 50 terminologies and 37 mono-lingual corpora. As required by the contract, ELRC managed to cover all languages with the required types of language resources for each language. In addition, ELRC performed the required evaluation and validation of the language resources so as to ensure their quality and suitability for machine translation purposes. All language resources collected by the ELRC have been uploaded to the ELRC-SHARE Repository (see below, supporting services).

29 ELRC Workshops : ELRC organised 29 country-specific ELRC workshops with participants such as national or regional/municipal governmental organisations, language competence centres, relevant European institutions and other potential holders of language resources from the respective national public service administration. Bringing ELRC to each country and getting engaged on the national level was key to fostering the local ownership and local responsibility on which ELRC is built. Through the workshops, ELRC managed to identify more than 1.000 potential data sources. Moreover, the ELRC workshops provided the contacts to potential data holders which were key to the subsequent data collection process.

Two ELRC Conferences: The first conference took place in April 2015 in Riga during the Latvian EU presidency and it marked the launch of the ELRC effort. The 2nd ELRC Conference was organised as concluding conference in October 2016 in Brussels, attached to the Translating Europe Forum. At each conference, more than

1 Further information on the Connecting Europe Facility programme: https://ec.europa.eu/digital-single-market/en/connecting-europe-facility

2 Further information on eTranslation: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation

3 Further information on CEF and the different CEF DSIs: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/CEF+Digital+Home

120 key stakeholders of the ELRC network (in particular potential data donors) were present.

The successful set-up and running of the Language Resource Board (LRB) (including corresponding communication structures and processes): The LRB is the governance body of the ELRC. It consists of a total of 60 members - one Technology National Anchor Point (Technology NAP) and one Public Service National Anchor Point (Public Services NAP) from each participating country. Overall, the National Anchor Points play the key role on the national level in effectively mobilizing the public sector and encouraging and facilitating the contribution of language resources among the public authorities and ministries in each country. They are the necessary bridges between the consortium and relevant players in each country, thus ensuring the effectiveness of the project's tasks through local ownership and connection to the national community. While the Technology NAPs are predominantly bridge-building with regard to technical aspects of data contribution and processing, the Public Sector NAPs act as promoters and central liaison to the relevant national public authorities, ministries and services. Moreover, the NAPs also directly contribute to both workshop organisation and data collection.

The successful set-up and operation of the ELRC Secretariat, the ELRC Website, the ELRC Repository and the ELRC Helpdesk: These supporting services proved to be the back-bone of any ELRC operations. While the ELRC Secretariat was in charge of the overall coordination of the action, the ELRC Website was the public face of the ELRC on the World Wide Web: it provided access to the ELRC Repository (ELRC-SHARE) to which all language resources can be uploaded and access to the ELRC Helpdesk which provides support in the case of any technical and legal questions associated with the sharing and provision of language resources from public service administration.

The information provided in this report provides an overview of major challenges faced by ELRC as well as clear indications and recommendations for future actions regarding language resources sharing and collection, in particular with regard to:

- Support channels to be provided to potential data holders
- The organisation of future conferences
- The organisation of workshops (including in particular the insurance of stakeholder involvement and sustainability)
- The work of the Language Resource Board
- Data collection activities to be undertaken

RÉSUMÉ

La Coordination Européenne de Ressources Linguistiques (CERL) était un contrat de services établi dans le cadre du programme « Mécanisme pour l'Interconnexion en Europe (MIE) »[4], résultant de l'appel d'offres SMART 2014/1074, et qui a permis la mise en place d'un mécanisme de coordination permanent pour les ressources linguistiques.

La CERL s'adresse à tous les pays qui participent au programme MIE, à savoir les 28 États membres de l'Union européenne (UE), ainsi que la Norvège et l'Islande. L'objectif principal de la CERL est de collecter des ressources linguistiques issues et à destination des administrations et services publics dans l'ensemble des pays européens qui participent au programme MIE afin d'améliorer la qualité, la couverture et les performances de la plateforme de traduction automatique eTranslation pour les services numériques existants et à venir du MIE tels que l'Échange électronique d'informations sur la sécurité sociale, le Règlement en ligne des litiges, la Justice électronique (e-Justice), les infrastructures de services pour un internet plus sûr, etc.[5] Ainsi, toutes les données collectées par la CERL seront utilisées par la Commission Européenne dans le but d'appuyer le développement de la plateforme eTranslation du MIE ainsi que son adaptation aux services numériques du MIE.

La durée d'exécution de la CERL a été de 24 mois (du 17 avril 2015 au 16 avril 2017). Les principaux résultats obtenus grâce à la CERL comprennent entre autres :

La collecte, la validation et la livraison de 225 ressources linguistiques : En tout, la CERL a collecté 138 corpus bi- et multilingues, 50 terminologies et 37 corpus monolingues. Comme énoncé dans le contrat, la CERL a couvert tous les types de ressources linguistiques requis pour chaque langue. En outre, la CERL a réalisé les travaux nécessaires à l'évaluation et la validation des ressources linguistiques collectées, de façon à garantir leur qualité et leur pertinence aux fins de traduction automatique. Toutes les ressources linguistiques collectées par la CERL ont été téléchargées dans l'outil de dépôt ELRC-SHARE (voir ci-dessous, les services de support).

L'organisation de 29 ateliers CERL : Dans 29 pays du MIE, la CERL a organisé un atelier auquel ont pris part des représentants des organisations gouvernementales ou régionales/municipales, des centres de compétences linguistiques, des institutions européennes concernées, ainsi que d'autres détenteurs potentiels de ressources linguistiques provenant des services publics nationaux. La présence de la CERL dans chaque pays et l'engagement au niveau national ont été essentiels pour favoriser l'appropriation et la responsabilité locales qui sous-tendent toute l'action de la CERL. Par le biais de ces ateliers, la CERL a réussi à identifier plus de 1000 sources de données potentielles. En outre, ces ateliers CERL ont permis de nouer des contacts avec les détenteurs de données potentiels qui sont au cœur du processus de collecte de données.

[4] Pour de plus amples informations sur le programme « Mécanisme l'Interconnexion en Europe » veuillez consulter le site suivant : https://ec.europa.eu/digital-single-market/en/connecting-europe-facility

[5] Pour de plus amples informations sur le mécanisme pour l'interconnexion en Europe MIE numérique et les différents services, veuillez consulter le site https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/CEF+Digital+Home

L'organisation de deux Conférences CERL : La première conférence a marqué le lancement du projet CERL, à Riga en avril 2015, dans le cadre de la présidence lettone de l'UE. La deuxième conférence a été organisée comme évènement de clôture du projet CERL en octobre 2016 à Bruxelles, conjointement au forum « Traduire l'Europe ». Chaque conférence a rassemblé plus de 120 participants, dont les principaux intervenants du réseau CERL (en particulier des fournisseurs potentiels de données linguistiques).

La réussite de la mise en œuvre et du fonctionnement du Conseil des Ressources Linguistiques (CRL) (ainsi que les structures et les procédures de communication afférentes) : Le Conseil des Ressources Linguistiques (CRL) est l'organe de gouvernance de la CERL. Ses membres sont les 60 points d'ancrage nationaux, soit un point d'ancrage national technologique et un point d'ancrage national des services publics par pays participant. Les points d'ancrage nationaux jouent un rôle clé au niveau national : ils mobilisent le secteur public, encouragent et facilitent la contribution de ressources linguistiques par les services publics et les ministères nationaux. Les points d'ancrage nationaux sont les relais indispensables entre le consortium et les acteurs concernés dans chaque pays, et ce afin de garantir le bon déroulement du projet tout en maintenant l'implication locale et le lien avec la communauté nationale. Alors que les points d'ancrage nationaux technologiques établissent des passerelles pour les aspects techniques, pour les contributions et le traitement des données, Les points d'ancrage nationaux des services publics assurent à la fois la promotion du projet et sa coordination vis-à-vis des autorités publiques nationales, ministères et services publics. Enfin, les points d'ancrage nationaux contribuent de manière directe à l'organisation d'ateliers et à la collecte de données.

La mise en place et le fonctionnement efficaces du Secrétariat CERL, du site internet, de l'outil de dépôt de la CERL et du service d'assistance technique et juridique de la CERL: Ces services de support fournis par la CERL se sont révélés être la colonne vertébrale opérationnelle du projet. Alors que le secrétariat CERL est responsable de la coordination générale de l'action, le site internet de la CERL en est la vitrine publique : en effet, il permet l'accès tout à la fois à son outil de dépôt (ELRC-SHARE) sur lequel toutes les ressources linguistiques peuvent être déposées par téléchargement, ainsi qu'au service d'assistance technique et juridique pour toutes les questions liées à l'échange et la mise à disposition de ressources linguistiques par les services d'administration publique.

Les informations fournies dans le présent rapport donnent un aperçu des principaux défis rencontrés par la CERL, ainsi que des indications claires et des recommandations pour des actions futures en vue de la collecte et du partage des ressources linguistiques, en particulier pour :

- Le soutien à apporter aux détenteurs de données potentielles
- L'organisation de futures conférences
- L'organisation d'ateliers (en particulier l'implication des intervenants et à terme, la pérennisation de leur participation)
- le travail du Conseil des Ressources Linguistiques (CRL)
- Les activités de collecte de données à entreprendre à l'avenir

1 METHODOLOGY

The overall objective of all ELRC actions was to manage, maintain and coordinate the collection of language resources in all official languages of the EU and CEF associated countries, in order to improve the quality, coverage and performance of automated translation systems and solutions in the context of current and future CEF digital services (CEF DSIs). The project's progress indicators are available in Annex 6. The following methods were key to achieving this objective:

Local ownership and responsibility

Local ownership and responsibility proved to be the central ingredient to the successful functioning of the ELRC. Significant language resources reside in and are in fact generated every day by national state, governmental, non-governmental and private organisations in EU Member State and CEF associated countries. Unlocking this currently largely untapped potential required the direct involvement of relevant local players particularly in the preparation of the country-specific workshops, but also in other tasks such as data collection and conferences. Therefore, in the course of the ELRC action the consortium collaborated closely and intensely with key local stakeholders and in particular with:

- Local public service administrations,
- Local technical experts,
- Local legal experts (liaison with ePSI Platform initiative[6] and the national experts of the Public Sector Information Group[7]),
- DGT Local Field Officers.

Complementing Grass-Roots Involvement with EU Patronage

An important element for the success of the project was to signal that ELRC is run under the EU Patronage. This pre-empted any perceived self-interest or commercial interest on the part of the ELRC consortium, which could be a considerable obstacle to unlocking local language resources. As already indicated above, ELRC closely worked with the DGT Local Field Offices on the identification of potential data holders/workshop participants; the support of the DGT Local Field Offices proved to be crucial for the establishment of initial contacts with national public service administrations, relevant language competence centres and potential future users of CEF eTranslation services thus signalling the direct ownership of the project by the EC.

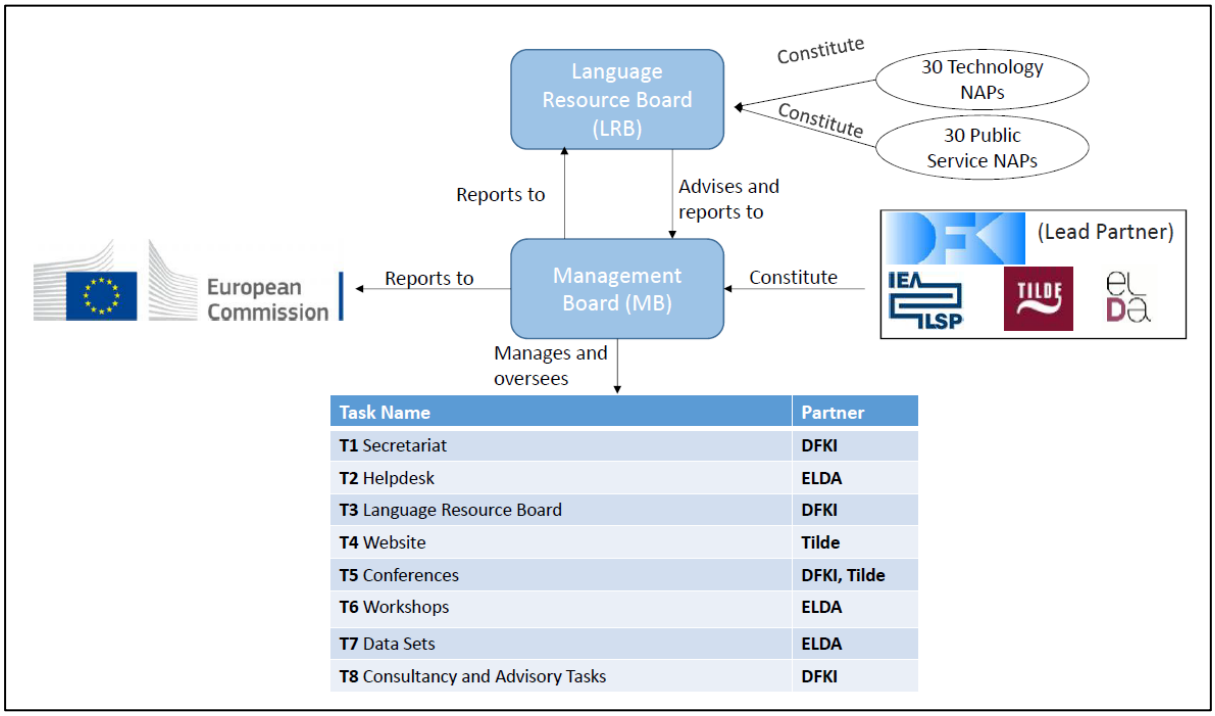
Effective management structures and procedures

Another key to the success of ELRC lied in its simple but accurate and effective communication and management structure illustrated below in Figure 1 which allowed for easy monitoring and early identification of deviations. As illustrated below,

[6] The activities and output of the ePSI Platform which was discontinued in 2016 were to a large extent taken over by the European Data Portal (https://www.europeandataportal.eu/).
[7] https://ec.europa.eu/digital-single-market/en/news/updated-list-psi-members

the main decision making body of ELRC was its Management Board (MB) which was composed of the four ELRC consortium members (DFKI (lead partner), ILSP, TILDE and ELDA). In addition, the Language Resource Board (LRB) reviewed the work of ELRC continuously and formally, within the four face-to-face LRB meetings, but also where applicable as part of the monthly LRB Online Progress and Q&A Sessions, monitoring project progress against targets.

On the level of each individual task, responsibilities and timelines have been carefully streamlined, including in particular carefully designed preparation process with corresponding milestones to be followed. This enabled easy continuous follow-up through the different supervising bodies. Corresponding procedures for quality management and conflict resolution have been established, as well as potential project risks (internal and external) have been identified for each task.



Task Name	Partner
T1 Secretariat	**DFKI**
T2 Helpdesk	**ELDA**
T3 Language Resource Board	**DFKI**
T4 Website	**Tilde**
T5 Conferences	**DFKI, Tilde**
T6 Workshops	**ELDA**
T7 Data Sets	**ELDA**
T8 Consultancy and Advisory Tasks	**DFKI**

Figure 1: ELRC Organisational Chart

Other important elements

The ELRC largely benefited from building on established efforts and expertise in particular with regard to the collaboration with local technical experts as part of the LRB and the data collection process. Because of the expertise covered by the ELRC consortium, it was also possible to create valuable synergies with past and existing related initiatives such as META-NET, CLARIN, ELRA, but also CRACKER and Translating Europe Forum (TEF).

To allow for adequate dealing with all legal tasks and the initial clearing of language resources, legal best practice has been subcontracted in the tasks concerned (legal helpdesk and workshops):

- Legal Helpdesk: professional legal support was key to minimizing legal hurdles in collecting data. The central issues concerned in particular conditions of use but also copyright and intellectual property rights (IPR) as well as privacy classification of the data;
- ELRC Workshops: professional support was sought and provided for the creation of the master presentation on legal issues involved in the sharing of language resources in the public sector (focus: PSI and its practical implications for data donations).

The sustainable, scalable and secure web services (including ELRC website, Helpdesk and Repository) proved indispensable for the permanent availability of all services to potential data donors and the presence and visibility of the ELRC action.

It is important to note that the project publicity proved to be more effective when strongly oriented towards local ELRC events; and finally, the originally assumed value of commercial resources proved minor in comparison to what could be achieved through data collection from the public service administrations.

2 CONFERENCE AND WORKSHOP DESCRIPTIONS

2.1 DESCRIPTIONS OF THE ELRC CONFERENCES

Overall, the ELRC organised two conferences: The first one right at the beginning of the project as kick-off event and the second one in the final year of the ELRC as final boost for data collection. Both conferences were of fundamental importance to the ELRC activities because they provided a useful and much needed forum for exchange of best practice and experiences among all participating countries, i.e. the 28 EU Members States, Norway and Iceland. While the ELRC Workshops were key for enabling local reach and for fostering local ownership and responsibility, the ELRC Conferences were vital for establishing a joint identity, for exchanging experiences and tackling problems, as well as for real encouragement through joint problem solving. Learning from each other was the major outcome of the conferences. This was particularly visible at the 2nd ELRC Conference where all relevant players were brought together for the purpose of intensifying data collection and overcoming potential difficulties experienced during the data collection phase.

General context of the 1st ELRC Conference

The first ELRC conference took place on 29 April 2015 as part of the Riga Summit 2015 on the Multilingual Digital Single Market. The main objective of the first ELRC conference was to create awareness, publicity and momentum for the launch of the ELRC activity as a service contract under the Connecting Europe Facility Programme to support data collection for CEF Automated Translation building block across EU and CEF-affiliated countries.

With the goal of providing a thematically attractive setting while at the same time exploiting synergies, the conference was organized as part of the Riga 2015 Summit on the Multilingual Digital Single Market, in tandem with the 2015 META-Forum, the Summit Plenary and the 2015 Multilingual Web Workshop.

Target audience / participants of the 1st ELRC Conference

The conference targeted language and language technology activists across Europe, evangelists for a multilingual Europe and a multilingual Digital Single Market and, in particular, technology support for multilingualism. In order to identify potential data donors, representatives from national public services, administration and governmental institutions, representatives were targeted as well.

The conference attracted 155 registered participants. 135 of them actually attended the conference. The distribution of participants was as follows:

- 45 participants from industry (including Language Service Providers)
- 34 participants from the public sector
- 51 participants from academia
- 5 other participants

Focus and contents of the 1st ELRC Conference

The welcome and opening session was provided by four highly distinguished individuals from the European Commission, the Republic of Latvia, the Internet Governance Forum and the World Wide Web Consortium: Márta Nagy-Rothengass (DG CNECT), Dace Melbārde (the Mister of Culture of Latvia), Jānis Kārkliņš (Internet Governance Forum and Director of the NATO Strategic Communications Center of Excellence) and Richard Ishida (W3C).

The opening was followed by a keynote on the "Digital Priorities of the 2015 EU Presidency" delivered by Gatis Ozols (Head of the eServices Unit, Ministry of Environmental Protection and Regional Development, the Republic of Latvia).

In order to avoid a monolithic session block and information structure, the CEF, CEF DSIs and CEF Automated Translation information sessions were already interspersed with contributions from the EU and CEF associated countries (Ireland, the Republic of Latvia) highlighting key language technology use and activities under way to support public services and administrations in the EU Member States.

The theme of language technology support for public services and administrations across Europe was strongly reinforced in the next session with diverse contributions from Hungary, Lithuania, Estonia and Portugal. In addition to reporting on data and technologies in use in pioneering applications for public services the session also focused on important requirements and needs of public services currently not supported.

Against the combined background of the Connecting Europe Facility and the CEF Automated Translation platform and taking into account the translation needs of public services across Europe, the next session presented the aims and objectives of the ELRC action and explained the structure of the ELRC work programme for data collection to serve public sector and administration requirements across EU and CEF associated countries. A specific focus of the presentation was on (i) how the Members States are involved in the collection process through the National Anchor Point tandems and take ownership and responsibility for the data collection and (ii) the concrete benefits derived from providing data in the form of CEF Automated Translation services directly geared towards supporting the needs and requirements of the data providers. The presentation also introduced the ELRC motto: "*Supporting our Languages is Supporting Europe and Supporting Europe is Supporting our Languages*".

All video and audio-materials of the conference are available in the "Events" section of the ELRC website.

General context of the 2nd ELRC Conference

The second ELRC Conference took place in Brussels, Charlemagne Building, on 26 October 2016, as a satellite event of the Translating Europe Forum (TEF) conference which focussed on translation tools and technologies. The collocation with TEF proved important as more than half of all ELRC Conference participants expressed their explicit interest to participate in TEF as well. The opportunity to

attend TEF after the ELRC Conference provided an added value to most ELRC Conference participants.

It is also important to note that the 2nd ELRC Conference was supported by travel reimbursements in order to facilitate the participation of the right people (i.e. potential data donors). Overall, 70 conference participants received travel support. The travel support provided by the ELRC proved to be one of the key enablers for the successful conference participation.

Target audience / participants of the 2nd ELRC Conference

The conference targeted public service administrators and representatives of public sector bodies in all EU Members States, Norway and Iceland. This included in particular potential holders and/or users of language resources, such as those responsible for translation services, representatives of information offices and/or public relation offices, as well as people responsible for data and data management. ELRC intended to attract at least three representatives from each country to this invitation-based event.

The rationale for having an invitation-based event as opposed to an open event was to ensure that the right audience (i.e. potential data donors) will be present at the ELRC Conference. Invitations were signed by the EC and administered by the ELRC Secretariat.

Overall, 124 participants attended the 2nd ELRC Conference; only Luxembourg and Spain were not represented. The distribution of participants was as follows:

- 11 Technology NAPs
- 11 Consortium Members
- 15 Speakers
- 87 Representatives of Public Service Administrations

Focus and contents of the 2nd ELRC Conference

The central goal of the conference was to support the sharing of language resources from public service administrations across all EU Member States, Norway and Iceland. In view of this goal the ELRC Conference:

- Demonstrated the benefits and possibilities of machine translation for public service administrations and raised awareness on the importance of corresponding data and language resources: Experts from the different stakeholders (including both translators in the public service administration as well as Language Service Providers (LSPs)) provided clear and hands-on accounts from their day-to-day practice of how MT is being used and the associated benefits. Moreover, there was a direct illustration of the role of data for the quality of MT, showing what can be achieved with the right data.
- Illustrated what kind of data is needed and how public service administrations can provide such data: Best practice examples and practical experiences of data donation from prominent national ministries and public service administrations were given. Moreover, ELRC gave an overview of the status of data collection in all countries. Questions linked to the provision of data were raised and answered by an international panel of experts.

- Gave an insight into the services of the CEF Digital Service Infrastructures (CEF DSIs) that should in the future be available to public service administrations in a multilingual manner. Following a short overview of the general requirements by public sector administrations, an overview was given of existing CEF DSIs (e.g. Open Data Portal, Europeana, Online Dispute Resolution platform, Electronic Exchange of Social Security Information).and their multilingual requirements.

All video and audio-materials are available via the "Events" section on the ELRC website.

2.2 DESCRIPTIONS OF THE ELRC WORKSHOPS

ELRC organised 29 country-specific workshops with participants such as national or regional/municipal governmental organisations, language competence centres, relevant European institutions and other potential holders of language resources from the respective national public service administration. Bringing ELRC to each country and getting engaged on the national level was key to fostering the local ownership and local responsibility on which ELRC is built.

In agreement with the EC, the workshop in the U.K. was omitted due to the overall political circumstances. The details on the workshops including the workshop reports, the workshop presentations and corresponding videos (where available) are available through the "Events" section of the ELRC website.

Workshop contents and concept

The central goal of the ELRC workshops was to raise awareness about the importance of language resources and in doing so, facilitate the collection of language resources from the public sector. The detailed objectives were:

- Raising awareness on the value and importance of data held by public services in overcoming language barriers;
- Engaging the public sector in the identification and sharing of data for CEF Automated Translation;
- Helping with legal and technical issues associated with the collection and/or provision of data by public service administrations;
- Collecting data to adapt CEF eTranslation to the day-to-day needs of public services in all EU Member States, Norway and Iceland.

Following an introduction by local VIPs and EC representatives, the workshops illustrated the impacts of multilingualism in Europe, provided and introduction to the Connecting Europe Facility programme, CEF Automated Translation and MT@EC, illustrated the foundations of automated translation (how does it work?), the data paradigm for machine translation and legal and technical issues involved in the provision or sharing of language resources.

Subsequently the workshops showed what can be achieved with one's data (how to manage it and why) and how participants can engage with ELRC (and why). A number of panels (e.g. on public sector administrations language needs or on public sector administrations language resources) ensured active participation of workshop

attendees, for ELRC to understand where potential data holders or language resources could be found.

Overall timeline

The vast majority of the ELRC Workshops was conducted in the first year of the project, and only four workshops were conducted in the first half of the second project year. The reason for rolling out the workshops as soon as possible was because they represented the first and most important step in making contacts with potential data donors. As such, in most countries, they provided the kick-off for data collection efforts and/or identifying potential data donors in that country.

Rollout and organisational aspects

The organisation of the workshops within ELRC was spread among all consortium partners with DFKI being in charge of the overall coordination (including subcontracting) and the remaining partners ELDA, ILSP and Tilde being in charge of the country-specific organisational management of the workshops in their region:

- **Tilde** was responsible for workshops in Latvia, Lithuania, Estonia, Finland, Sweden, Norway, Denmark and Iceland (Northern Region).
- **ILSP** was responsible for workshops in Greece, Cyprus, Bulgaria, Romania, Croatia, Slovenia, Austria, Czech Republic, Slovakia, Hungary and Poland (South-Eastern Region).
- **ELDA** was responsible for workshops in France, Spain, Portugal, Italy, Malta, Belgium, the Netherlands, Luxembourg, Germany and Ireland (South-Western Region).

The preparation time for each workshop was at least 3 months. The preparation phase typically involved the following activities:

- Step 1: Contacting the local DGT Field Officer 3 months prior to the event;
- Step 2: Getting contacts to the national public administration from the local DGT Field Officer;
- Step 3: Involving the country's experts and practitioners in the area of machine translation and Open Data (e.g. local data.gov.x, EU Open Data Portal anchors, national PSI experts);
- Step 4: Finalising the list of speakers and invitees as soon as possible, at the latest 4-6 weeks before the event;
- Step 5: Signature of invitations by the EC and e-mailing the participants no later than 4 weeks before the event;
- Step 6: Adaptation of the workshop agenda and contents to the local context dynamics 4 weeks prior to the event;
- Step 7: (Before finalizing invitations) Identification of a venue large enough to host all participants of the workshop which also provides the opportunity for interpretation;
- Step 8: Publishing the details of the workshop on the ELRC website.

All workshops were held in the national language and as such, all master slides prepared by the ELRC had been translated and adapted to the particular local context. The adaptation of the contents was of crucial importance for several reasons:

- First of all, the local adaptation and translation of the technical presentation slides allowed the audience not only to better understand the overall process of training MT-systems, but also to directly see the effects of data on the quality of the machine translated output. The latter can typically only be judged by native speakers or speakers who are fluent to maximum extent in a language.
- Second, the local adaptation was also vital to understanding the legal context. While the PSI governs the exchange of public sector information in general, the local implications (i.e. the ones the national public administrations have to deal with) can differ in various Member States. Therefore, it was of utmost important to include the local adaption also in the non-technical parts of the workshop.
- Last but not least, the adaptation of the general context presentations (focus of the particular country) was important for the ELRC audience to build the bridge to the ELRC initiative – why is ELRC important in my particular country with regard to my particular activities?

Interpretation to and from the local language was provided at each workshop. It was indispensable for the conduct of the workshop, in particular to allow the answering of questions from the audience by international experts (e.g. legal advisers) or ELRC representatives as well as to facilitate the open discussion between international experts/ELRC representatives and the audience. Without interpretation, panel discussions and even the presentation of the ELRC activities by the ELRC team or any other contents presented by relevant international experts would not have been possible. The duration of the ELRC workshops was 1 day.

3 COUNTRY PROFILES AND COUNTRY ACTIONS

The following section provides an insight into country-specific profiles and actions that emerged in the course of the ELRC activities. The focus is on the ELRC workshops, resulting collaborations and contacts established as well as major findings from these activities.

3.1 STAKEHOLDERS AND STAKEHOLDER INVOLVEMENT

Overall, the ELRC has managed to reach out to more than 1.100 potential data holders across Europe, Norway and Iceland in the course of the ELRC workshops, resulting in the identification of 1.083 potentially useful data sources.

As shown above, however, the ELRC workshops were mainly tailored to awareness-raising about the Connecting Europe Facility (CEF) and CEF Automated Translation and to attracting potential data owners in order to facilitate data collection for the needs of adapting the EC machine translation service available at the time (MT@EC). The targeted participants were public sector organisations and their representatives dealing with and managing multilingual content (typically heads of departments of language services). Asking a public sector representative to select, record and share data relevant to CEF Automated Translation was, in many cases, considered as an additional imposed burden to the already overloaded internal procedures of public administrations. As a result, even in the case of engaged participants who were convinced of the benefits of the endeavour, the data collection process was hindered by the fact that public administrations across Europe lack the resources to support ELRC activities. Moreover, and closely connected to this issue was the one of authorisation of data donations and collaboration with ELRC: Even when data holders were willing to provide data and when on departmental level permission was given, corresponding authorisation by their superiors and even acting heads of ministries was needed to proceed.

In order to effectively overcome such impediments, a "top-down" approach interleaved with a "bottom-up" is needed with regard to stakeholder engagement, i.e. in future ELRC will need to reach out to high level officials, competent and willing to facilitate internal administrative procedures for the purposes of ELRC. The involvement of policy level and decision-makers becomes even more crucial for ensuring the sustainability of the data supply in the future. While one-spot donations will always be possible, the commitment to a continuous, mid-to-long-term collaboration with the ELRC can only be made with the support and consent of the policy level and top decision makers in each country. Therefore, future actions should clearly focus on targeting key decision-makers to advocate the necessity of a) investing in LR collection and maintenance, and b) supporting national actions related to digital services and multilingualism, in order to secure the presence of their language(s) in a digitally connected Europe.

Another important aspect with regard to the target audience and stakeholder involvement is the inclusion of the end-users, i.e. national representatives of CEF Digital Services that should be empowered through CEF eTranslation. Therefore it is

crucial that future workshops present machine translation as an indispensable solution to providing multilingual online services. In fact, machine translation and hence CEF eTranslation is the only solution for ensuring multilingual functionalities for online services with exponentially increasing amount of content to be translated as well as with dynamic and user generated content (e.g. users feedback, comments, queries, posts, etc.). In this respect machine translation in general, and CEF eTranslation in particular can and will enhance public online services for citizens, businesses and administrations, whether these are CEF Digital Services or other public services. The involvement of end-user representatives is therefore a key to understand what kind of data is needed in order to make CEF eTranslation work for such services.

3.2 KEY ISSUES RAISED BY STAKEHOLDERS AND ACTIONS

In the course of ELRC action, in particular when conducting the ELRC Workshops, a number of often recurring questions and concerns was raised as regards data sharing and contributing data for CEF Automated Translation. One central question raised by the participants was on why one should participate in the ELRC and donate data (i.e. what are the benefits for the data donors/donating institutions?). Closely related to this were questions on why to donate data to the ELRC at all if one does not use machine translation or if machine translation seems unable to deliver good results (e.g. in the case of morphologically rich languages). Other frequently discussed issues were the legal constraints of donating data (i.e. Can data be shared? Which data can be shared?) or the fundamental question on why use machine translation at all and more general questions on how machine translation works, what CEF eTranslation and MT@EC are etc.

Table 1 below gives a summary of the key concerns of the participants along with the corresponding answers. All questions and answers can also be found – and are regularly updated – in the "Helpdesk" section of the ELRC website (see "Full list of FAQs"). It is important to note that the issues raised by the stakeholders represent the key questions that had to be tackled and addressed by the ELRC. ELRC addressed these issues in the workshops and also in the day-to-day communication with the stakeholders.

However, potential data donors require and deserve a continued re-assurance that collaboration with the ELRC is indeed to their benefit and that data donations will not have any negative effects (e.g. legal consequences). The aforementioned issues cannot be overcome in a single workshop but they require extensive and continued communication / lobbying on the national level. The communication efforts should involve all relevant stakeholders and data providers in each country and most importantly, the key decision-makers / policy level. Future efforts may take the form of exhaustive national roadshows and local presence instead of single workshops, to win the support of the decision-makers, fully address and overcome the existing concerns and allow for sustainability of data donations.

Question:	**What is going to happen to the data we provide?**
Answer:	The data will go to the EC (Directorate-General for Translation (DGT)) to support the improvement of the EC machine translation system MT@EC. Open datasets will be made available through the open data portals, e.g. EU Open Data Portal.
Question:	**Why should we (public institutions) actually provide data?**
Answer:	Supporting your own language is supporting Europe and vice versa. The data you will provide will help improve the performance of CEF Automated Translation services, so the more language resources the better. Within the CEF programme, CEF Automated Translation services are free, secure and accessible to public administrations in all EU Member States and CEF affiliated countries (Iceland and Norway).
Question:	**We (public institutions) don't have any data for you! We work only paper-based. We outsource our translations.**
Answer:	If translations are outsourced, you should ask for the translated data to be delivered with the corresponding translation memories. Make sure to negotiate the provision of the translation memories with the language service provider ahead.
Question:	**We cannot just share our data with you – they are confidential!**
Answer:	Most data held by the public sector is public data. Administrations provide various types of information online to the citizens (e.g. news, legal texts, official communications, interviews, brochures, background information, etc.). This information can also be available in a foreign language. For example, on the website of the German national government, all information is provided in German, English, and French.
Question:	**How can I upload my data to the repository?**
Answer:	You can upload data to the ELRC-SHARE Repository in three simple steps: 1. Register (new user) or login (returning user) 2. Provide a basic description for the language resource (title, short description, language(s)) 3. Upload the .zip file For further instructions, please read the Walkthrough for Contributors and/or contact the ELRC Helpdesk.

Question:	**What is MT@EC? What is CEF eTranslation?**
Answer:	MT@EC is the current EC Machine Translation system used since June 26th, 2013. It is an online service with a web user interface in 24 languages for human use. It can be used as a web service in a machine-to-machine scenario. Using a highly secured protocol (sTESTA) coupled with the European identification (EU-Login) MT@EC guarantees confidentiality of data. MT@EC can be used by staff working for public administrations in EU countries, Iceland and Norway free of charge. eTranslation is a service developed under the Connecting Europe Facility (CEF) Programme which provides automated translation services with the goal of making CEF DSIs accessible to any EU citizen in his/her own language. European public online services such as Europeana, the European Data Portal, the Online Dispute Resolution Platform, etc. should benefit from eTranslation.
Question:	**Why would we need MT@EC/eTranslation? We have human translators!**
Answer:	MT@EC can substantially help make the translation process more productive and more efficient. EC translators are responsible for translating content into all official EU languages. In total, more than 7,000 translators working for DG Translation and EU institutions have translated more than 2.3 million pages in 2014. MT@EC is used daily by French, Spanish, Portuguese and Italian translators to produce initial translations that are post edited in a very efficient way. For other languages (e.g. German) the quality level of the output is still too low. The eTranslation service which will be available towards the end of 2017 is going to see further progress in the quality for German due to the deployment of neural machine translation engines. Furthermore, in the last year, significant progress has been achieved through domain-specific engines. In particular reports and texts in the area of economics can be successfully translated using MT@EC. In other cases, MT@EC can be used to rapidly scan long texts in a foreign language and point out passages to be translated by humans. Overall, the translation quality is directly related to the availability of good quality training data in the language concerned: if the data for training MT is good, then the MT system will be good.

Question:	**How can we access MT@EC?**

Answer:	MT@EC can be used by any Member State administration free of charge at least until the end of 2020. It can be accessed as follows:
	• Staff working for **EU institutions or agencies** can use MT@EC with their (EU login) credentials.
	• Staff working for a **public administration** in an EU country should follow these steps:
	  ○ Sign up for your personal EU Login account and password (using only your professional email address).
	  ○ Send an email to DGT-MT@ec.europa.eu stating that you have an EU-Login account, indicating what your job involves and which public administrative body you work for. Don't forget to include your full signature with your contact details
	  ○ DGT will create your MT@EC account and notify you.
	Apart from individual users, the MT service is also available to EC information systems and online services through an API.
	Details on obtaining access to MT@EC are also available here.

Question:	**Why should we support MT@EC / eTranslation – we can have our own national solution?**
Answer:	Typically, national solutions are targeted for a particular range of topics or languages. Hence, the scope of MT@EC/eTranslation is broader and more comprehensive. By supporting MT@EC/eTranslation, participants can expect to have access to a broader service.
Question:	**Machine translation is directly opposed to our national policy that young people should learn foreign languages.**
Answer:	Not necessarily. Machine translation can actually provide a good basis for learning languages. Initially, it can be used to bridge the gap for people who cannot speak a particular language until they acquire initial language skills. For instance, at university level, machine translation can be used to provide automatic and simultaneous translations of lectures for foreign students who do not master the language.

Question:	**Machine translation will never work for our languages (e.g. Estonian, Finnish, Hungarian and other morphologically rich languages).**
Answer:	Processing certain languages with the current MT technologies is more difficult because of e.g. their free morphology or their free constituent order. MT experts are working on new MT solutions based on neural networks more adapted to these languages. Moreover, the European Commission funds several research and innovation actions within the H2020 programme, like QT21 to investigate MT solutions for languages which currently receive only sub-optimal MT support. Within CEF eTranslation neural machine translation engines are being built for Hungarian and German. They will be available by the end of 2017. What is important is that – regardless of the methodology – huge amounts of parallel resources are needed for the implementation of the MT systems, since these systems rely on machine learning.
Question:	**Why should I care about translations and get hold of/keep corresponding language data?**
Answer:	Whether you translate your material internally or outsource it, your process can benefit from the re-use of language data from previous translations in a cost-effective way while improving the quality of the output. For instance, if you outsource your translations, you can negotiate with your language service providers a better price for the translation if you are able to provide them with previously translated texts in your area (e.g. earlier versions of leaflets, reports, etc.).
Question:	**How should I manage my data and why? We don't have any infrastructures or resources for this!**
Answer:	In the public sector there is a great diversity in translation management: from paper-based to digitized workflows with term lists and translation memories storage. From an organizational point of view, much benefit can arise even from small changes in dealing with language data. Suggested actions can be taken without major effort, including: • Analysis of all phases of data development • Based on this, creation of a corresponding "data management plan" (DMP), even a very basic one, covering key questions such as: ✓ Which data is important? ✓ Where is it stored? ✓ Can it be further processed? • Documentation of all relevant data • If possible, using the web as additional publication channel and reap the benefits of linked data. For further details on best practices for data management please visit the ELRC website or this workshop presentation.

Table 1: Key Issues raised by the ELRC stakeholders

4 REPORT ON THE WORK OF THE LANGUAGE RESOURCE BOARD

4.1 COMPOSITION OF THE LANGUAGE RESOURCE BOARD

The Language Resource Board (LRB) was set up as the governance body for European Language Resource Coordination effort. For each CEF language, the LRB includes one technological representative (**Technology National Anchor Point, Technology NAP**) and one representative from the public administration (**Public Services National Anchor Point, Public Services NAP**).

The **Technology NAPs** are highly regarded language or language technology experts. They often have a distinguished academic or research background, and/or represent a national language institution. The Technology NAPs were vital to the work of the LRB and to the success of the ELRC action, as they included relevant local expertise with regard the processing of language resources, such as knowledge of the particular language as well as tools and technologies relevant for the processing of such data. They played a key role in the conduct of the ELRC Workshops and in the subsequent data collection phase.

The **Public Services NAP** are representatives of national public services, public administrations or ministries. They act as a liaison contact persons to the national, regional and local administrations, and are able to effectively mobilize and spread the word about the importance of language resources and the ELRC effort among the public authorities/ministries in each country. They played a key role in identifying potential data holders, establishing contacts with them and continuously promoting the ELRC activities.

The composition of the ELRC Language Resource Board is provided in Annex 1. In April 2017, the LRB comprised 54 National Anchor Points. It is important to note that in the case of Lithuania, Ms. Daiva Vaisniene, the Head of the State Commission for Lithuanian language, holds both positions (Public Service and Technology NAP).

In the two years of its existence, the LRB has been undergoing several individual changes, mainly due to the retirement of existing Technology and Public Service NAPs, due to their leave for other functions or positions etc. Even though in each case, a successor could be identified, it is clear that the LRB composition will require constant attention and will always be subject to changes.

The "Anchor Points" section on the ELRC website presents the up-to-date list of National Anchor Points.

4.2 LANGUAGE RESOURCE BOARD MEMBER SELECTION AND APPOINTMENT PROCESS

The Technology National Anchor Points (Technology NAP) had already been mostly identified at the beginning of the project. Changes occurred only in the case of Belgium, France, and Poland.

The Public Services National Anchor Points (Public Services NAPs) were appointed either through the CEF Expert Group or they were identified in the course of the ELRC action. Each Technology NAP has been asked to provide 2-3 suggestions for Public Service NAP in his/her country, taking into account the profile of a public service representative who is likely to be effective given the CEF Automated Translation and ELRC objectives at stake, together with short justifications of why/how the candidate will be effective.

After a first round of Public Sector NAP appointments through the CEF Expert Group which took place until October 2015, the remaining Public Service NAPs have then been nominated as part of the workshop follow-up and the ongoing data collection efforts. The rationale behind this strategy was to ensure that the right candidate for NAP was chosen: The actual collaboration and engagement of different members of the public service administration in the frame of the ELRC workshop often revealed the most suitable candidate for future collaboration with ELRC.

4.3 ACTIVITIES OF THE LANGUAGE RESOURCE BOARD

Because of its distribution across all participating countries, the day-to-day work of the LRB had to take place remotely over the phone, by email and through web conferences. The main tool for continued information exchange and collaboration were the Regional Q&A Online Sessions which were organised on a monthly basis and to which all NAPs of the particular region were invited. The Regional Q&A Online Sessions were organised for the following regions:

- Northern Region: Latvia, Lithuania, Estonia, Finland, Sweden, Norway, Denmark and Iceland (Regional ELRC Representative: Tilde)
- South-Eastern Region: Greece, Cyprus, Bulgaria, Romania, Croatia, Slovenia, Austria, Czech Republic, Slovakia, Hungary and Poland (Regional ELRC Representative: ILSP)
- South-Western Region: France, Spain, Portugal, Italy, Malta, Belgium, the Netherlands, Luxembourg, Germany, United Kingdom and Ireland (Regional ELRC Representative: ELDA)

In the monthly Regional Q&A Online Sessions all NAPs were informed on topics of interest, upcoming tasks and events, supporting documents etc. Moreover, the regional Q&A Online Sessions provided the major tool for steering and controlling the progress in different ELRC tasks (workshops, data collection, conference preparation etc.) and thus have been a vital instrument in the collaboration with the National Anchor Points. Overall, there were more than 2.000 information requests by LRB members in the last project year alone.

During the first two years of the LRB activity there were four face-to-face LRB meetings:

- the first LRB and Kick-off meeting in Riga in April 2015;
- the extended LRB meeting in Berlin in November 2015 with the goal to prepare all NAPs for the conduct of the ELRC workshops;
- the LRB Meeting in Lisbon in July 2016 with the goal to kick-off the data collection phase and to provide guidance to the NAPs with regard to the collection of language resources from public service administrations;

- the final LRB Meeting in Berlin in March 2017 with the goal to give an insight into the continued efforts of ELRC and to provide the best possible support and guidance to the successor projects.

All relevant documents, contents and guidelines necessary for the work of the LRB have been provided through the corresponding Dropboxes: The ELRC Workshop Information Package Dropbox (which contained all relevant documents, master contents, subcontracts, templates etc. for the ELRC workshops) and the ELRC Data Collection Information Package Dropbox (which contained all relevant information on data specifications, data collection process and corresponding subcontracts).

Overall, the LRB proved to be one key ingredient for the successful collection of language resources. As shown in Annex 2, 20 out of 54 NAPs signed data collection subcontracts and delivered the corresponding number of language resources. In addition, 12 NAPs provided data without the need for a subcontract. 60% of the LRB members made direct data contributions to the ELRC. Most importantly, 51 out of the 54 NAPs provided contacts to the potential data holders in their country and/or corresponding sources for data, facilitating and extending the ELRC efforts in their countries.

5 LANGUAGE RESOURCE COLLECTION

The central goal of all ELRC activities was to identify and collect language resource data sets (typically originating from institutions of workshop participants), which are readily usable to train and optimize machine translation systems for the CEF languages. These data sets could be aligned parallel corpora, translation memories, language models, comparable corpora, monolingual corpora, terminologies, grammars, etc.

Each data set was clearly identified either as Open Data (to be published on the EU Open Data portal and any other appropriate place), or as restricted/confidential language resources, specifying the licensing conditions and the right-holder(s), in view of obtaining the right to use such restricted resources for setting up and adapting automated translation services for the CEF DSIs.

ELRC managed to collect a total of 225 language resources. The list of language resources delivered by the ELRC, including all these details is provided in Annex 3.

5.1 GENERAL APPROACH TO DATA COLLECTION

Figure 2 below illustrates the process for data collection followed by ELRC.

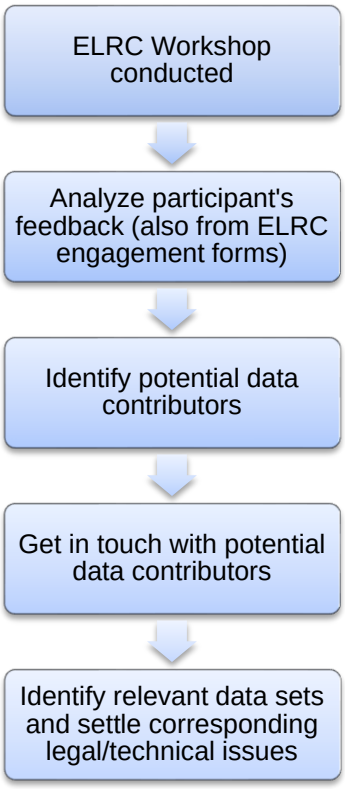


Figure 2: Process for data collection

The standards for the ELRC Data Collection were set in the Resource Collection Guidelines available through the "Helpdesk" Section of the ELRC Website or directly

at the Info Point. The Guidelines were shared with all NAPs and people involved in the data collection process.

The task of data collection was split between the ELRC Consortium and the ELRC NAP network. The work of the NAPs was often supported by subcontracts. Overall, 23 subcontracts were signed amounting to 93.000 EUR in total. A total of 125 Language Resources could be collected through subcontracts.

The responsibility for data collection was shared among all ELRC consortium members in the following way:

- **DFKI (6 countries)**: Germany, Austria, Luxembourg, Netherlands, Hungary, Czech Republic
- **ELDA (8 countries)**: Ireland, Spain, Portugal, Belgium, Italy, Malta, France, (U.K.)
- **Tilde (8 countries)**: Latvia, Estonia, Lithuania, Finland, Sweden, Denmark, Iceland, Norway
- **ILSP (8 countries)**: Greece, Cyprus, Slovakia, Slovenia, Bulgaria, Poland, Romania, Croatia

One important tool for monitoring the data collection progress, encouraging data collection activities, exchanging best practice and finding solutions were the monthly Q&A Online Sessions that were organised for the Northern Region, South-Eastern Region and South-Western Region (see 5.3 on the activities of LRB). Every month, the progress in data collection was monitored and illustrated, and in doing so, the "friendly competition" between the ELRC regions was encouraged. With 102 language resources contributed, the South-Eastern Region was the largest contributor, closely followed by the South-Western Region (67 language resource contributions) and the Northern Region (56 language resource contributions). It is important to note that the data collection process in the South-Eastern Region started almost two months earlier than the data collection in the other regions due to the early finalisation of all workshops. It is also important to note that the number of language resources is not the only measurement for the success of the data collection efforts: The suggested contribution size for corpora was 100.000 words. In some cases, this threshold was exceeded by far and large corpora of more than 1 million words were contributed.

5.2 DATA VALIDATION PROCESS

ELRC carried out a validation of the datasets to estimate their quality. In the context of ELRC, validation is understood as the quality control of a Language Resource against a list of relevant criteria. It is important to note that the validation of donated data differed from the validation of crawled/processed data.

The donated data provided by the public service administrations typically already consisted of qualitative data in terms of contents (in particular asserted translations for multilingual data, data produced by human experts). Thus, their validation consisted of:

- checking the compliance of data with ELRC scope
- checking the format of provided data

- checking if legal information provided is compliant with ELRC scope

The crawled data required a much deeper content validation whereas the technical part was already included in the language resource production requirements. The ELRC Data Validation Guidelines available through the "Helpdesk" Section of the ELRC Website or directly at the Info Point provide all details on the validation process including in particular manual and automated validation employed. As illustrated in the report, automated evaluation was mainly used to:

- Identify potentially relevant bi-texts in a particular domain / identify whether or not a particular text belongs to a particular domain or not;
- Estimate the alignment quality by calculating the so-called alignment score;
- Calculate the length ratio (note: segments with a length ratio close to 1 have similar length whereas segments having a ratio far from 1 have a big difference in terms of length, which could reveal segments are not well aligned).
- Identify different numbers in TUVs, indicating when numbers in the target segment are different compared to the source segment (for segments containing numbers in the text).

All these scores (automatically calculated by the ILSP-FC crawler, see 6.5 "Tools used for LR processing") were taken into consideration to detect potential alignment or translation issues.

To produce a final resource, filtered and potentially problematic translation units were discarded or flagged: first, translation units assigned with errors were automatically removed; then the remaining translation units were annotated based on the probability of finding the same errors. The definition of error thresholds for the 5 main error types (wrong language identification, alignment, tokenization, MT translation and translation errors) allowed for clear standardized evaluation of resources.

5.3 FINAL DATA SETS

Details of the 225 language resources that were delivered to CEF eTranslation through the ELRC effort are available in Annex 3. Overall, ELRC collected 138 bi-/multi-lingual corpora, 50 terminologies and 37 mono-lingual corpora[8]. ELRC managed to cover all CEF languages and to provide the required types of language resources for each language.

For the training of machine translation systems, bi-/multi-lingual corpora present the most important input. Figure 3 below shows the number of bi-/multi-lingual corpora collected by the ELRC for each of the 26 CEF languages. As can be seen from the diagramme, only for 8 languages (English, French, German, Italian, Modern Greek, Polish, Romanian, and Spanish) more than 10 bi-/multi-lingual corpora could be collected. All other languages could hence, at least as an initial result of the ELRC

[8] ELRC committed to submitting at least one bi/multi-lingual corpus, one mono-lingual corpus and one terminology for each language. However, in agreement with EC bi-lingual corpora were preferred over mono-lingual corpora because of their value for MT training. As such, when there was a choice to provide either a bi/multi-lingual or monolingual corpus, ELRC provided the bi-/multi-lingual corpus.

action, count as “under-resourced” languages for which data collection efforts should be intensified in future. A more detailed view on the number of resources available for each language by type is provided in the Tables of Annex 4 alongside with an overview of LR available by domain for each country.

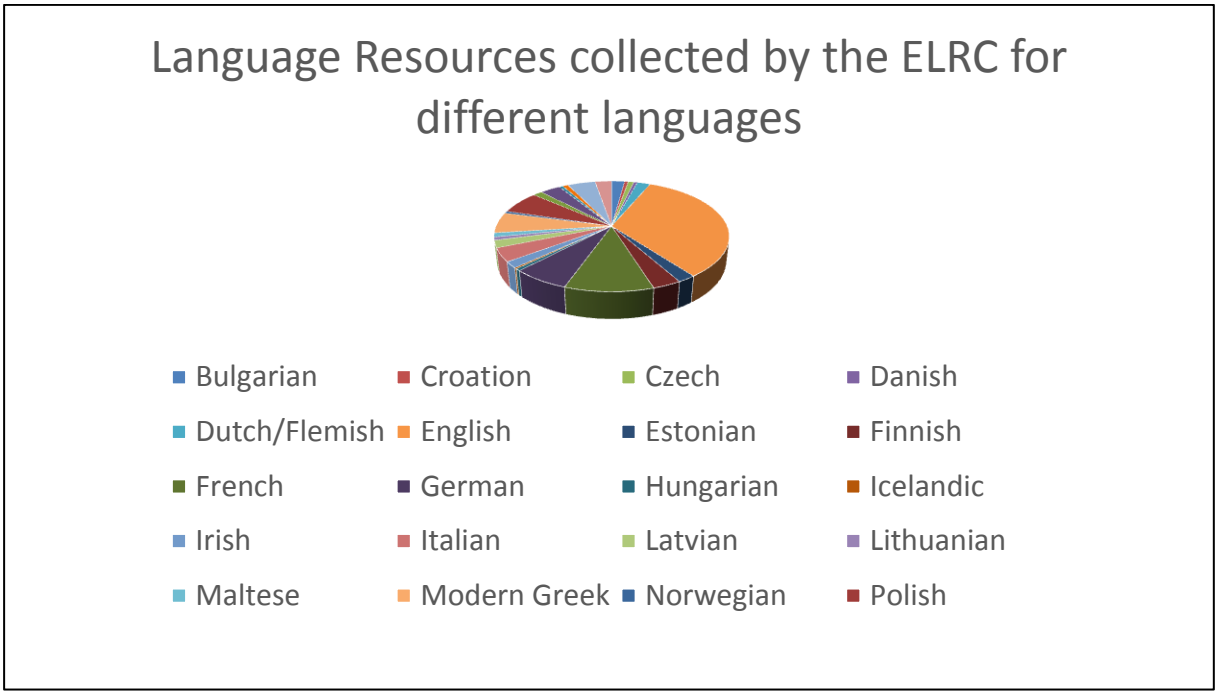


Figure 3: Language coverage for bi-/multi-lingual corpora

ELRC analysed to what extent the collected language resources are relevant to the different CEF DSIs. Figure 4 below provides an overview of the language resources collected by ELRC with regard to their direct relevance for the CEF DSI. As shown in the diagramme, many language resources were of general nature and could not (without a detailed analysis) be directly linked to a particular CEF DSI. The best covered CEF DSIs are e-Justice (32 language resources) and Europeana (17 language resources).

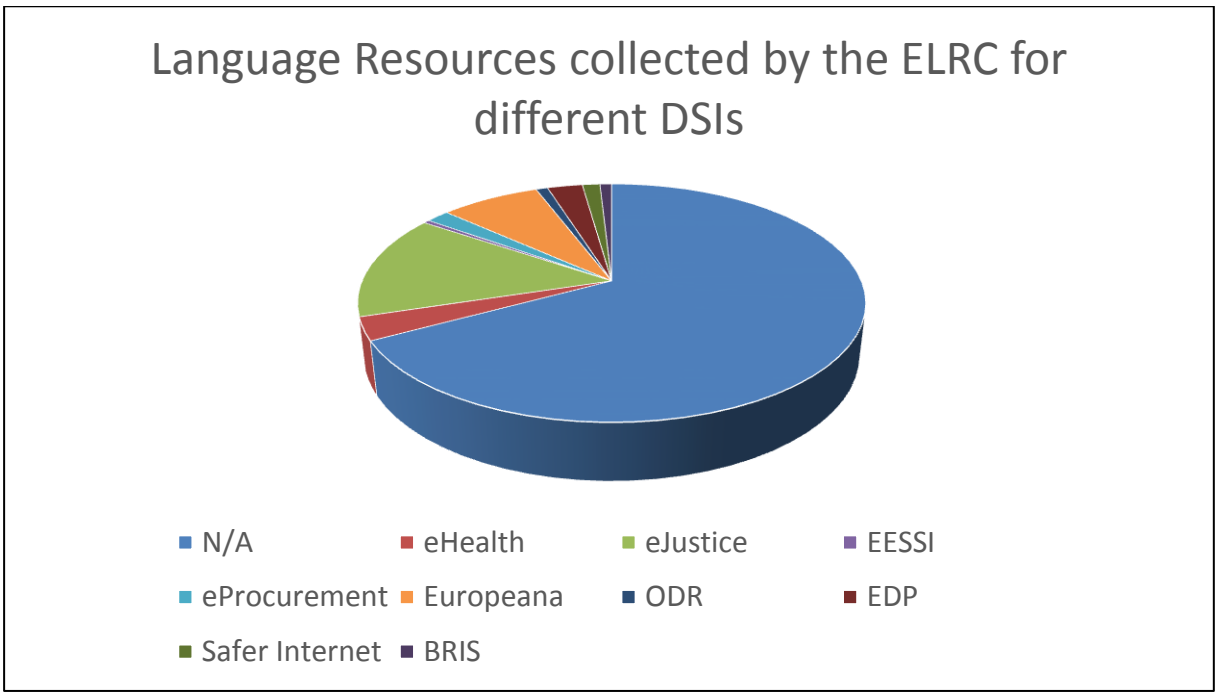


Figure 4: Coverage of CEF DSI by the language resources contributed by the ELRC

As a consequence, future data collection efforts should be focussed on the following CEF DSIs and areas:

- Online Dispute Resolution ODR (Area: Consumer Rights)
- Business Registration BRIS (Area: Business Registry)
- Electronic Exchange of Social Security Information (EESSI) (Area: Social Security)
- European Data Portal EDP (Area: Open Data)
- Safer Internet (Areas: Security/Cybersecurity)

As regards the legal status of the language resources collected by the ELRC, only 39 data sets are covered by a non-standard-/other licensing terms and 10 were still under review at the ELRC contract end in April 2017 (see Annex 3 for details). The remaining largest part of the collected language resources was open under CC licences or PSI. The ELRC effort focused on identifying and collecting data that can easily be shared. The restrictions with regard to the content and size of a language resource were minimal. It is likely that future data collection efforts (which will have to be more focussed and targeted towards particular domains and languages) will require a much greater efforts to free and make available such data. Future data collection activities will likely involve significantly more efforts with legal (licensing) and technical (anonymization, processing, development) activities.

5.4 ELRC-SHARE Repository

The ELRC-SHARE Repository is the solution provided by the ELRC for documenting, uploading, storing and accessing language resources. All resources identified and made available as a result of the ELRC need to be:

- documented with the appropriate information describing the resource (aka metadata)

- easily uploaded and stored
- updated as necessary, both metadata & data
- indexed and, as a result, searched, accessed and downloaded (as necessary) according to the terms and conditions of their use.

The ELRC-SHARE Repository supports all the aforementioned tasks. It was developed and maintained by ILSP as a tool to support the whole lifecycle of the data collection task in ELRC. Through the ELRC-SHARE, people interested in language resources can also browse and – depending on the licensing situation – access and download the corresponding language resources.

ELRC-SHARE is based on a META-SHARE software instance. The software has been adapted the operational needs of ELRC and it has been evolving to respond to specific requirements of its stakeholders.

Using the ELRC-SHARE Repository

The ELRC-SHARE Repository is hosted by ILSP/Athena RC and is accessible through the "Resources" section on the ELRC website or directly.

From a user perspective, ELRC-SHARE offers[9]:

- **For data owners and contributors**: basic functionalities for registering/login and contributing language resources through a simple web form. A detailed Walkthrough for data owners and contributors is available in the "Help" section of the ELRC-SHARE Repository website.
- **For metadata editors** (ELRC members): a user-friendly metadata editing environment for the description of resources. A detailed Walkthrough for metadata editors is available in the "Help" section of the ELRC-SHARE Repository website.
- **For the general public**: browsing, simple and faceted search of the resources inventory.

At the backend, ELRC-SHARE supports:

- storage, upload, download of language resources, i.e. the primary datasets, in zip format. The size of the zipped file to be uploaded currently cannot exceed 50MB. For larger datasets, the contributors should contact the ELRC team at: elrc-share@ilsp.gr.
- storage, import, export of the following for each language resource:
  - the metadata record (db record, export in xml format)
  - legal documentation (in .zip format)
  - licence text (e.g. if standard licence, then the respective official text)
  - language resource Validation report.
- notification and reporting mechanisms for the efficient monitoring of updates of the hosted language resources

[9] Status in April 2017.

Access rights are regulated through the **ELRC-SHARE user management module**, through which specific access rights to resources and repository operations are assigned to users, depending on their role(s) and the publication status of a language resource.

Documenting language resources with metadata

ELRC language resources are (formally) documented using the ELRC-SHARE schema. In essence, the ELRC-SHARE metadata schema[10] is an application profile of the META-SHARE schema, appropriately modified for the requirements of ELRC, e.g. pruned resource types (e.g. audio), adapted legal component (e.g. PSI compliance element, licence values added), additional elements, e.g. appropriateness for CEF DSI.

The schema includes the following **mandatory metadata categories**:

- Administrative information
  - Identification (e.g. resource name, description)
  - Distribution (e.g. licensing, access form of the resource)
  - Contact Person
  - Metadata (e.g. creator and creation date of the metadata record)
- Technical information for all resource types
  - Linguality (i.e. number of languages included in the resource)
  - Languages
  - Sizes
  - Text formats (e.g. plain text, PDF, XML, TMX etc.)
- Technical information for specific resource types
  - for lexical/conceptual resources
    - Lexical conceptual resource type (e.g. terminological glossary, bilingual dictionary etc.)
  - for language descriptions
    - Language description type (e.g. grammar).

The following metadata categories are **optional**:

- Administrative information
  - Resource Documentation
  - Resource Creation
- Technical information
  - Character Encodings
  - Annotations (only for corpora)
  - Domains
  - Text Classifications
  - Creation
  - Encoding Information for Language Description (e.g. linguistic level of encoding)
  - Encoding Information for Lexical Conceptual Resource (e.g. types of encoding for its contents, such as lemmas, grammatical information, translation equivalents etc.)

[10] Full description and documentation of the metadata schema is available online at https://elrc-share.ilsp.gr/documentation/ELRC-SHARE_schema.html

Conclusions from the ELRC-SHARE use

The ELRC-SHARE Repository has proved a very useful tool for managing both

- the interactions between ELRC and data contributors when the latter share their data, and
- the whole lifecycle of ELRC language resources, from the (initial) contribution to the final delivery of resources to the EC.

The process through which data contributors can upload data to ELRC-SHARE has been kept as simple as possible in order to encourage people to upload their data. Data contributors only have to fill in a simple web form with three fields: resource title, resource short description, and language(s). All other required meta-data are filled in by the ELRC team, after contacting the contributor/data owner, if needed. The work of the metadata editor is also made simpler through an easy to use metadata editor, endowed with new possibilities to attach supporting information for each contribution, like legal or technical documentation of a resource.

The flexible architecture of the user management module has also proved extremely helpful, as it made it quite easy to securely accommodate new user roles with carefully designed access and operation rights on resources.

5.5 TOOLS USED FOR LR PROCESSING

To support the provision of the appropriate data for training the MT@EC/eTranslation engines, ELRC examined the use of automatic methods for discovering, acquiring and rendering in the appropriate form generic or domain-specific, monolingual and/or bilingual language resources (LRs) emerging from web content.

A significant number of processing steps is needed before such resources can be used for training today's natural language processing and machine translation engines. If one targets the acquisition of resources from the web, a web crawler like Apache Nutch[11] can be used for the fast acquisition of large text collections. The text and metadata of documents in these collections have to be extracted with the use of tools (e.g. Apache Tika[12]), or in the case of PDF documents with text extractors targeting this specific format (e.g. PDFBox[13]). Advertisements and repetitive text (e.g. disclaimers and menus), have to be identified with tools like Boilerpipe[14]. Language identification at document and paragraph level can be performed with tools covering almost all EU languages (e.g. lang-detection[15]), which nevertheless may have to be further adapted to recognize text in specific language dialects or writing systems (e.g. Norwegian Bokmål and Nynorsk). Clean text and metadata have to be exported to an easily processable collection of XML or JSON documents. Other components that can prove useful are topic modelling tools (e.g. Mallet[16]) that in an unsupervised way

[11] http://nutch.apache.org/
[12] https://tika.apache.org/
[13] https://pdfbox.apache.org/
[14] https://github.com/kohlschutter/boilerpipe
[15] https://github.com/shuyo/
[16] http://mallet.cs.umass.edu/

cluster documents together; or topic classifiers which, like JEX[17], assign IDs from predefined ontologies (e.g. the EU's multilingual thesaurus[18]) to particular documents.

In case of bilingual/multilingual LR acquisition, the set of necessary modules should include online discovery and prioritization of translation links during crawling, as well as a solution for document pairing (a.k.a document alignment) like the ones recently evaluated in the Bilingual Document Alignment Shared Task in the First Conference on Machine Translation (WMT16)[19]. Sentence aligners (e.g. Hunalign[20], Maligna[21] and Gargantua[22]) are then used to extract sentence alignments from bitexts. The results can be evaluated with tools like c-eval[23] and should be exported in standard industry formats including Moses and TMX. Comprehensive toolkits like Bitextor[24] and the ILSP Focused Crawler (ILSP-FC)[25], which integrate almost all of the task-specific functionalities mentioned above and cover all EU languages are also available.

Overall, the use of the ILSP-Focused Crawler (ILSP-FC)[26] proved to be most valuable as it is a modular system that includes components and methods for all tasks required to acquire monolingual and bilingual domain-specific corpora from the Web: link classification, text normalization, document clean-up, boilerplate removal, language identification, metadata extraction, identification of bitexts (i.e. documents that are translations of each other), alignment of segments, and filtering of segment pairs.

In the course of the ELRC project and based on feedback by all ELRC consortium members, the tool was continuously tested and enhanced at ILSP in order to provide more accurate results. It was eventually deployed at all four ELRC partner sites for acquiring language resources for specific (EN-X) language pairs. The crawler uses open source language identification libraries that perform at over 99% precision at document level for more than 50 languages. In order to meet ELRC needs to cover all CEF languages, missing resources were constructed and integrated in the tool (for example language profiles for both Norwegian written standards, Bokmål and Nynorsk). The accuracy of the language identifier, when examining Norwegian texts is 98%/90% for text chunks of at least 500/100 characters, respectively.

Turning to the identification of bitexts, the crawler employs a combination of methods that are language-pair agnostic, i.e. they do not use bilingual lexica or MT results that are often difficult to generate. Instead, the methods are based on shallow features that two web documents may have, including translation links to each other, similar URLs, high rate of common digits in their content, links to the same images, similar HTML structure etc. For evaluation purposes, the bitext identification module was

[17] https://ec.europa.eu/jrc/en/language-technologies/jrc-eurovoc-indexer
[18] http://eurovoc.europa.eu
[19] http://www.statmt.org/wmt16/bilingual-task.html
[20] https://github.com/danielvarga/hunalign
[21] https://github.com/loomchild/maligna
[22] https://github.com/braunefe/Gargantua
[23] https://github.com/tilde-nlp/c-eval
[24] https://sourceforge.net/p/bitextor/wiki/Home/
[25] http://nlp.ilsp.gr/redmine/projects/ilsp-fc/
[26] http://aclweb.org/anthology/W/W13/W13-2506.pdf

submitted in the WMT 2016 Bilingual Document Alignment Shared Task and scored a high recall of 91%. It was 7th among 21 participations, 1st among those not using language- or language-pair specific information.[27]

ELRC LR -ID	Language Resource Name	Results	
115	Parallel corpus (Greek - English) in the public administration domain	good=12332 bad = 177	98.59% 1.41%
379	Parallel corpus (Bulgarian - English) in the public administration domain	good = 11094 bad = 168	98.51% 1.49%

Table 2: Results of test on segment alignment of the ILSP-FC

For segment alignment, the crawler uses open source aligners to construct collections of candidate parallel segments. A battery of criteria are applied on these candidates with the purpose of filtering out specific types of TUs and of generating precision-high multilingual LRs for training MT systems. In order to test the parallelness of the LRs created with the tool, we trained the C-Eval parallel corpora cleaning and evaluation tool[28] on the DGT-TM 2016 release and applied it on two datasets delivered as ELRC LRs. Results in Table 2 indicate that the two LRs include a high percentage of useful translation segments. ILSP-FC[29] is available under a GPL v3.0-icense. Licensing and support for commercial uses and applications is also available.

As an alternative to the pipeline use of the tool, specific modules in the post-crawling process can be called as standalone modules for all tasks mentioned above. To this end, they could be used for the processing of resources residing in the ELRC-SHARE Repository.

Finally, ELRC also developed a toolkit (the ELDA Crawled Data Management Toolkit (ELDA CMTK)) that allows to exploit the output from the ILSP-FC. The tookit contains 14 distinct tools and is publicly available under a GPL v3 Licence. It exploits in a dependant way other external tools, in particular the GNU aspell in order to do filtering based on "spell checking", as well as SGBD PostgreSQL and/or SQLite. Aspell is publicly available (http://aspell.net/) under a LGPL v2.1 licence through GitHub (https://github.com/GNUAspell/aspell). PostgreSQL and SQLite are both SGBD open-source, free (BSD-like for PostgreSQL and "Public Domain" for SQLite). The ELDA Crawled Data Management Toolkit is available on Github (https://github.com/ELDAELRA/elda_cmtk) along with a detailed description.

[27] http://www.aclweb.org/anthology/W16-2375.pdf
[28] https://github.com/tilde-nlp/c-eval
[29] http://nlp.ilsp.gr/redmine/projects/ilsp-fc/

5.6 CHALLENGES FACED DURING DATA COLLECTION

There were several challenges faced in the different countries that impacted on the collaboration with the ELRC. One problem concerned the **less spoken languages**[30]: There is already a general lack of available resources for such "small" languages for translation into English, but even more a lack of language resources for language pairs of these small languages in a language other than English. The final data sets provided by the ELRC clearly illustrate this problem (see 5.3 Final Data Sets, Figure 3: Language coverage for bi-/multi-lingual corpora).

Moreover, even if data contributors are identified and data exists, several problems emerged during the data collection phase:

- Authorisation by superiors: Potential contributors require the official authorisation – targeted actions and official support were necessary;
- Resource-related problems: data had to be assessed by the public service administrations to see what kind of data could be contributed without any need for processing;
- Legal and licensing issues: legal concerns, copyright problems and privacy/confidentiality issues (anonymization in cases of personal data), security-related concerns for security-relevant data etc. made potential contributors reluctant to share their data; In future significant efforts will need to be made to work with each individual data holder in order to overcome these issues; inclusion of key decision-makers will be necessary to put a sustainable data pipeline in place through which data can be shared continuously;
- Translation procedures: The translation procedures were found to be differing not only between different countries, but also between different institutions/ministries (centralized translation services (e.g. in Finland, across the entire country) or translations under the responsibility of individual institutions (e.g. in Germany or France), in-house or outsourcing). Especially, in case of outsourced translations, there is an absence of standardized procedures for quality control: In several cases, digital copies of the translated texts were not made available to the public sector organization because it was not agreed so in the contract. Consequently, the organisations received the actual translation but not the underlying tmx-files or source-files.
- Technical problems (digitization / level of required processing):
    - The level of digitisation and the availability of language resources vary in the different participating countries. Therefore, the range of language resources that could be contributed to CEF Automated Translation varied greatly in terms of the resources' technical readiness for training automated translation systems (identified resources ranged from scanned paper copies/pdfs through .doc/xls files to ready-to-use tmx-files).
    - In many cases data could not be shared as it would need to be anonymized first. Anonymisation, however, requires processing of the data; since data cannot be just handed to externals for this purpose,

[30] The six most spoken EU official languages are German (16%), Italian (13%), English (13%), French (12%), Spanish (8%) and Polish (8%).

corresponding contracts will need to be negotiated and put in place between institutions willing to annonymise and share their data and individual experts to allow for such assistance.

- Technology affinity and lack of awareness: Due to limited affinity to technology by many translators and a lack of awareness about the value of their data and data management in public administration in many cases ELRC had to take the first step in awareness raising and help tp set-up a corresponding supply pipeline. The ELRC workshops were an opportunity to reach out to translators and the subsequent day-to-day follow-up with potential data holders helped in overcoming such barriers. Nonetheless, significant further efforts with targeted outreach will have to be undertaken, particularly for the languages and in the domains presented above (see 6.3 Final Data Sets) in order to free the language resources needed to train CEF eTranslation in the required CEF DSI domains. Outreach activities going beyond the scope of the ELRC and the ELRC Workshops could take the form of presence at relevant national events and the organisation of corresponding “Technology experience Cafés” that give potential data donors the opportunity to see and feel what can be achieved with regard to machine translation with the help of the right training data. Corresponding demonstrations will need to be prepared for each case.

6 REPORT ON CONSULTANCY

Overall, only one official but fairly comprehensive consultancy request was submitted to ELRC from the EC. The request focused on the costs for creating a parallel corpus of sufficient size for a given (arbitrary) topical domain and for each of the EU languages. It is important to note that a general solution to the consultancy task in fact is an open research problem at PhD level. Nonetheless, a corresponding investigation was undertaken by the ELRC and the requested solution was elaborated and developed.

When approaching the problem, the following elements were assumed as given:

- a general description of the differences/distance between GD[31] and D[32] (perhaps in terms of an information theoretic measure such as perplexity, entropy, cross-entropy, Kullback-Leibler divergence etc.),
- a description of the size of the data GD and D
- an MT technology (e.g. PB-SMT)
- a description of the specificity of the GD and D domains and the MT models computed on GD and D (are these narrow or wide/diverse domains), e.g. in terms of translation table entropy
- a targeted average quality level Q (e.g. BLUE score) of the tuned MT output on D,
- a domain adaptation strategy "•" (e.g. difference in cross-entropy),

Consequently, the size (D) should be predicted (i.e. the size of in domain data ID of type D required to tune MT (GD • ID) to achieve quality level Q on data from domain D). The difficulty of the task was that there is currently no analytic formula that could compute the desired prediction from the given information.

Nonetheless, the request was processed by DFKI and timely response was given. The corresponding report included also per word translation cost estimates and domain adaptation best practice (covering domain adaptation through supplementary lexical resources, model selection, supplementary data selection, incremental updates, size of in-domain data required to tune the system to a specified quality level on a specific domain) and the final estimation of the size (D). All details are available through the corresponding ELRC Advisory Report (see Annex 5).

[31] GD = General Domain Data

[32] D = domain which is substantially different from GD (so that the output of a general domain MT system MT(DG) for source side data from D is not of sufficient quality

7 SUPPORT SERVICES PROVIDED BY ELRC

Overall, the ELRC provided several additional supporting services that were vital for the implementation of all actions and for the successful collection of language resources: (i) the ELRC Secretariat, (ii) the ELRC Helpdesk, and (iii) the ELRC Website. The support services continue to be provided under the successor ELRC projects (SMART 2015/1091) running until December 2019.

7.1 ELRC SECRETARIAT

The ELRC secretariat serves as the single point of contact for governance operations as well as for the organisation of events and meetings. It keeps track of all activities, language resources and stakeholders. The secretariat is available and reachable by phone and email during DFKI working hours (phone number: +49 681 85775 5285, email: elrc-secretariat@dfki.de). During the first two years of the ELRC project the main activities and functions of the secretariat included:

- Preparation and submission of reports;
- Organisation and conduct of the ELRC Conferences;
- Organisation of the LRB Meetings;
- Support to the organisation of the ELRC workshops (overall coordination);
- Preparation of any ELRC consortium meetings and meetings/conferences with the EC (virtual and physical);
- Preparation of general dissemination materials and contents (in particular ELRC Brochures) and organisation and conduct of all dissemination activities;
- Support to contracting (in particular workshops and data collection subcontracts).

The day-to-day handling of enquiries about ELRC and about ELRC events proved to be one major function of the ELRC Secretariat (see Table 3 below for details). With a total of 5.179 information requests in the last year, the overwhelming number of enquiries to the ELRC secretariat was made by email. Most support was provided to enquiries by the Language Resource Board (1.169 for general enquiries, 652 for meeting-related enquiries of the LRB). The number and distribution of requests underlines again an observation made already in the first year of ELRC: people involved in ELRC prefer personal contact and feedback rather than submitting their questions and queries through the anonymous helpdesk (see also 8.2 ELRC Helpdesk, for further details).

Type of Request	Received by email	Received by Phone
T1 - ELRC General	93	1
T1 – Communication with the EC	334	-
T1 - Other	2	8
T2 – ELRC Helpdesk	7	-
T3 – LRB Meetings (incl. Reimbursements)	652	3
T3 – LRB Other	1.169	12
T4 - Website	0	-
T5 – Conference Reimbursements	377	4
T5 – Conference Other	1.196	9
T6 – Workshops subcontracts	219	3
T6 – Workshops other	98	-
T7 – Data Collection Subcontracts	598	7
T7 – Data Collection Other	433	-
T7 - Repository	1	-
T8 – Consultancy and Advisory	0	-
TOTAL	5.179	47

Table 3: Enquiries to the ELRC Secretariat in the Final Year

With regard to officially presenting ELRC, the ELRC secretariat prepared corresponding brochures: The first brochure being a general one and information about the ELRC, its background, its activities. The second brochure was a flyer targeted on the question "*Why your data matters*." While both flyers fundamentally answered all key questions around the ELRC and why people should get engaged, it also became obvious that this may not be the best way to communicate and illustrate the reasons "*Why…*?".

In the framework of the successor project under SMART 2015/1091 the ELRC secretariat has prepared an animated video to promote the cause of the ELRC. It is very friendly, easy to understand and straight to the point. The video is permanently available through the ELRC website.

Last but not least, as indicated earlier, the ELRC Secretariat was responsible for the creation of two guidance documents which supported the work of the ELRC NAP and external stakeholder network with regard to workshop organisation and data collection: The ELRC Workshop FAQ and The ELRC Resource Collection Guidelines available through the "Helpdesk" Section of the ELRC Website or directly at the Info Point.

7.2 ELRC HELPDESK

A legal and technical helpdesk for Language Resources provision has been in operation since June 2015. It was available over the phone, reachable by email and web interface. It provided answers to questions related to the preparation and provision of language resources. The helpdesk was permanently manned by (junior) experts who were supervised by senior experts. All questions that concerned technical and legal questions with regard to data sharing, data contributions and machine translation which are relevant to the ELRC stakeholders are published and maintained in the "Helpdesk" section of the ELRC website (see FAQ section and Webforum).

The page statistics of the Helpdesk site show that there were more than 1.100 page views and 228 unique visitors in total (i.e. 10,4 unique visitors per month on average) across the 22 months of Helpdesk operations. Most notably, in several months parallel to the workshop conduct, the number of unique visitors exceeded 20 per month with more than 100 page views for one month. The detailed statistics on the number of visits (unique visitors) to the Helpdesk and the number of viewed pages of the Helpdesk are shown in Figure 6 below.

Despite the frequent visits of the Helpdesk website, on average only 2 queries per month were submitted as actual requests. Most incoming requests were submitted either to the ELRC Secretariat (see section 8.1) or directly to the regional ELRC Representatives (Table 4 below). Overall, this pattern shows that participants heavily relied on the personalized “support structure” that the ELRC team offered rather than using the “anonymous helpdesk”. In general, it was observed that the Helpdesk is only used by newcomers to the project or “outsiders”. Potential data donors, identified at the various workshops, visited the Helpdesk as it was announced to them during the workshops, LRB meeting, Q&A Online Sessions, bi- or multi-lateral web conferences, however those who donated data preferred to contact the ELRC representatives whom they knew from the workshops and Q&A Online Sessions directly.

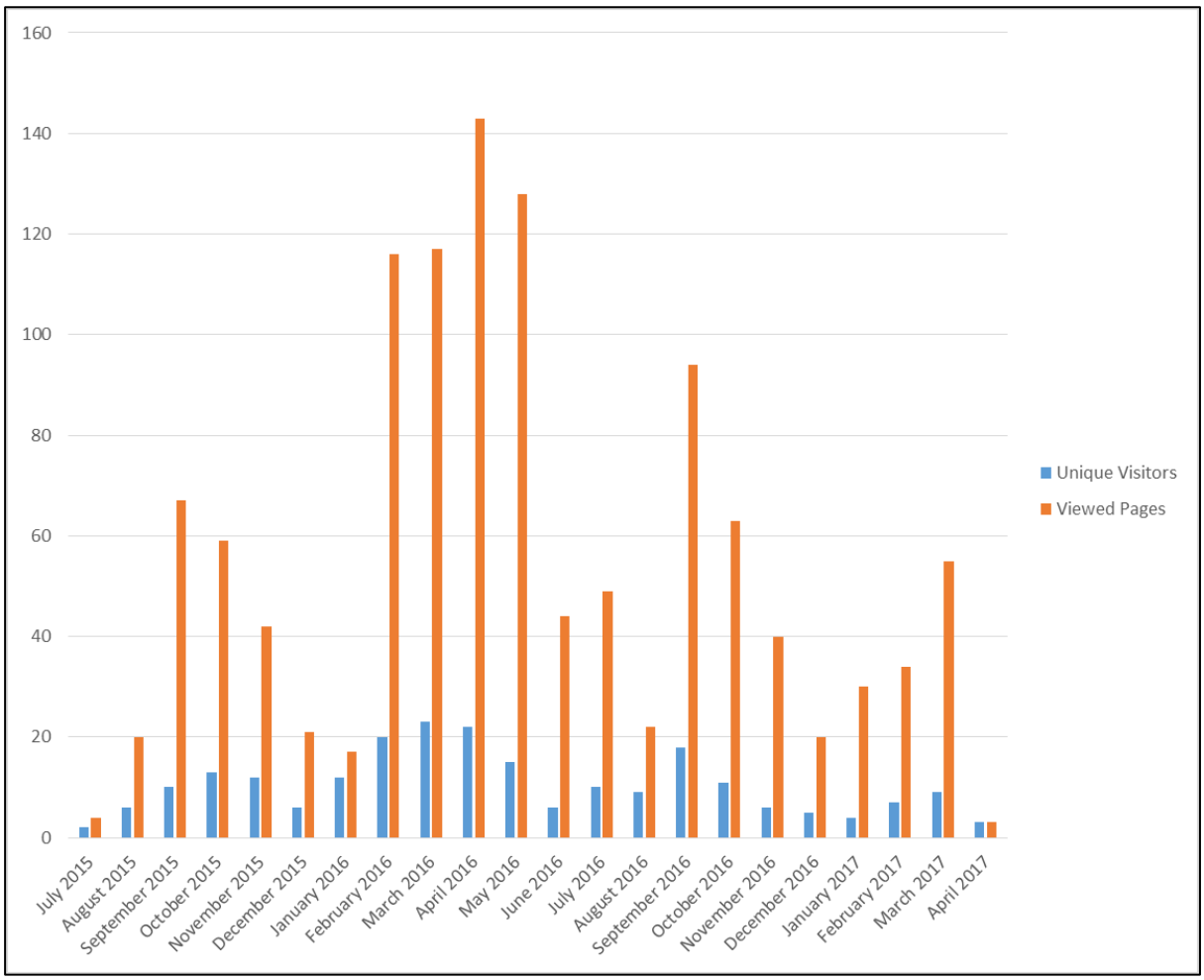


Figure 5: Helpdesk visits

ELRC Representative	Country	Emails (incoming)	Emails (outgoing)	Web conferences, calls etc.
DFKI	Germany	11	7	6
	Austria	19	19	0
	Luxembourg	3	4	0
	The Netherlands	39	39	4
	Hungary	17	19	1
	Czech Republic	5	11	3
ELDA	Ireland	48	40	5
	Spain	64	60	7
	Portugal	16	16	4
	Belgium	24	20	5
	Italy	56	48	6
	Malta	62	62	6
	France	64	64	12
Tilde	Latvia	-	-	-
	Estonia	45	56	16
	Lithuania	61	56	19
	Finland	35	30	14
	Sweden	49	47	11
	Denmark	79	85	12
	Iceland	93	107	16
	Norway	83	74	14
ILSP	Greece	35	44	18
	Cyprus	-	-	3
	Slovakia	7	7	-
	Slovenia	12	13	-
	Bulgaria	21+1	23+1	-
	Poland	33	39	-
	Romania	25	31	-
	Croatia	10	12	-
	Additonal generic support across all countries	60	90	36[33]

Table 4: Enquiries handled by ELRC Representatives with regard to data collection

[33] Redmine tickets

7.3 ELRC WEBSITE

The ELRC website (www.lr-coordination.eu) was set up as the face of the ELRC in the World Wide Web. It provides information on ELRC and on all ELRC events and also is the access point to all support services provided by the ELRC (i.e. the ELRC Helpdesk, the ELRC-SHARE Repository, registration facilities for workshops, information on the NAP network etc.). The website can serve up to 50 concurrent users.

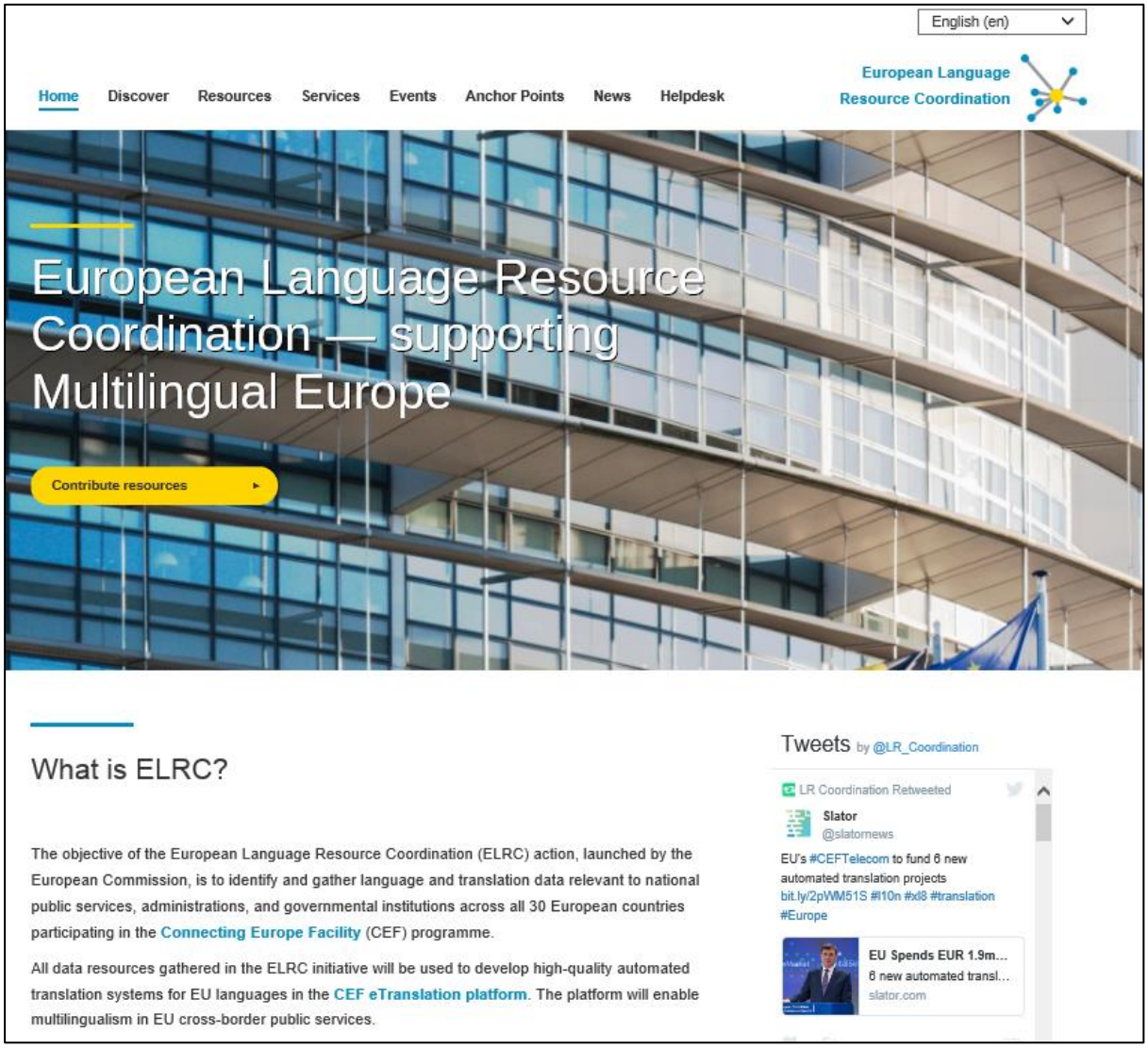


Figure 6: Home page of www.lr-coordination.eu

The web page was set up in June 2015. Throughout the ELRC action the website content was regularly updated and additional content was added. The website continues to evolve within the successor ELRC projects under SMART 2015/1091, (e.g. change of contents on “Home”, omission of the section “About” and addition of new sections “Discover” and “Services”). The ELRC website currently has the following sections:

- The **Home** page presents the key information about the ELRC project and the relation to CEF eTranslation, including link to the CEF Digital portal. On the right hand side are social media channels, including News ticker and corresponding tweets and News headlines are displayed. The Home page has been localized to all CEF languages.
- The **Discover** section provides more detailed information on Automated Translation, MT@EC, CEF eTranslation and the broader context of multilingualism. In 2017 this section replaced the earlier **About** section which provided information on ELRC's background and activities.
- The **Resources** site provides all information on what are language resources, how to identify resources and how to contribute to ELRC. From this site, visitors can access the **ELRC-SHARE Repository** and **Data sources** submission form.
- The **Services** section, which was added in 2017, provides details on language processing tools and services as well as on on-site assistance opportunities.
- The **Events** section provides all event related information on ELRC Conferences, Workshops and LRB Meetings. All upcoming and past events are listed. For all upcoming events, visitors can access the event registration facility, view the agenda, time, place and focus of the particular event. For past events, visitors can also access the corresponding presentations, videos (where available) and reports.
- The section on **Anchor Points** provides information on ELRC's National Anchor Points in all participating countries, i.e. who is the NAP in each CEF country and who can be consulted if any information on ELRC activities is necessary in the respective country. This site is provided in all CEF languages.
- The **News** section is updated on a regular basis and provides a general summary of actions undertaken in the ELRC project and other related information about ELRC and CEF related activities. This part is not localised in all CEF languages and is provided in English only.
- The **Helpdesk** site includes most important information for identification and preparation of language resources. It provides information on channels through which any interested person can get answers to any ELRC and Resource collection related questions. It provides a simple to use web-form for asking questions to the technical and legal helpdesk for Language Resources provision. Last but not least, it is the access page to the FAQs.

8 CONCLUSIONS

Based on the information presented in this report, the conclusions arising from the ELRC activities with regard to future action can be summarized as follows:

Support channels provided by the ELRC

With 5.179 email enquiries handled in the 24 months of the ELRC service contract, the ELRC Secretariat proved to be of fundamental importance to the ELRC effort in general. It played a key role in the coordination of ELRC events, the management of the LRB and the coordination of all data collection activities.

The workshop organisation and the organisation of the data collection by regions proved successful: Each region had their dedicated regional ELRC Representative available to provide help and support with these tasks (Northern Region with regional ELRC Representative Tilde, South-Eastern Region with regional ELRC Representative ILSP, South-Western Region with regional ELRC Representative ELDA).

Through its local activities and engagements ELRC has managed to set-up a functional, personalized and institutionalized network of collaboration. Most emerging issues and requests were submitted either to the ELRC Secretariat or to the ELRC Regional Representatives. The ELRC Helpdesk received only 2 official requests on average per month. LRB members as well as potential data donors preferred to rely on the personal support structures provided by the ELRC rather than on the "impersonal" ELRC Helpdesk. The ELRC Helpdesk was mainly used by newcomers who had not been integrated in the ELRC network yet.

Nonetheless, the ELRC Helpdesk remains a necessary channel for newcomers across all EU and CEF-affiliated countries who are not yet aware of or engaged in the current network to submit their queries, to receive support and to get engaged with the ELRC. As such, it must be maintained in the future work.

ELRC Conferences

As indicated above, the 2nd ELRC Conference managed to attract many representatives from public service administrations. The reasons for achieving this goal were:

- the strong involvement of all National Anchor Points in identifying relevant invitees;
- the fact that the conference was organised as invitation-only event;
- the collocation with the Translating Europe Forum which was an added benefit for all participants of the ELRC Conference;
- the financial contribution by ELRC towards covering the travel costs of participants.

For future ELRC conferences it would be important to find national and European events attractive to potential data holders with which the ELRC Conference could be

collocated. In this case, the registration process must be streamlined and coordinated with the event to which the conference is attached. Most importantly, the event should be invitation-based to gather the most relevant stakeholders to the ELRC activities. Last but not least, the offer of contributing towards participants' travel expenses proved to be important to ensure participation of key stakeholders.

ELRC Workshops - Ensuring stakeholder involvement and collaboration with the ELRC

As indicated earlier, the participants targeted at the ELRC Workshops were public sector organisations and their representatives dealing with and managing multilingual content (typically heads of departments of language services). It appeared, however, that even when the targeted data holders were willing to provide data and when on departmental level permission was given, an authorisation by their superiors and even acting heads of ministries was needed to proceed. In general, the involvement in the ELRC activities clearly required authorisation from the top.

A "top-down" approach interleaved with a "bottom-up" is needed with regard to stakeholder engagement in all future workshops which means that ELRC will need to reach out to high level officials, competent and willing to facilitate internal administrative procedures for the purposes of ELRC. The involvement of policy level and decision-makers becomes even more crucial for ensuring the sustainability of the data supply in future. While one-spot donations will always be possible, the commitment to a continuous, mid-to-long-term collaboration with the ELRC can only be made with the support and consent of the policy level and top decision makers in each country. Consequently, future actions should clearly target key decision-makers and advocate to them the necessity of a) investing in LR collection and maintenance, and b) supporting national actions related to digital services and multilingualism, in order to secure the presence of their language(s) in a digitally connected Europe.

The workshops, and ELRC activities in general, must also be targeted towards responding to the key concerns raised by potential data donors. Some of the key question posed by potential supporters and newcomers to the ELRC were: "Why should we get involved? Why should we donate data? What is the benefit for us?"

Potential data donors require and deserve a continued re-assurance that collaboration with the ELRC is indeed to their benefit and that data donations will not have any negative effects (e.g. legal consequences). The aforementioned issues cannot be overcome in a single workshop but they require extensive and continued communication / lobbying on the national level including all relevant stakeholders and data providers in that country and most importantly, the key decision-makers / policy level. Future efforts may hence take the form of exhaustive national roadshows and local presence instead of single workshops, to win the support of the decision-makers, fully address and overcome the existing concerns and allow for sustainability of data donations.

Work of the Language Resource Board (LRB)

The work with the LRB has been extremely constructive and productive. The regional Q&A Online Sessions helped in organizing the work of the Board and provided a forum for monitoring progress, for discussing and solving any emerging issues. In

future, to further support the work of the LRB, additional country-specific reach-out activities would be advisable. On top of the LRB events such as the aforementioned monthly Q&A Online Sessions or the regular face-to-face meetings of the Language Resource Board which NAPs attend, it would be beneficial for the ELRC to be present in the relavant events in individual countries. The ELRC Experience Cafés launched in the beginning of 2017 pursue this approach. It is planned to organise the ELRC Experience Cafés at events central to the NAP and public administration/services community (e.g. eGov/Digital Conferences of the EUPAN, Week of Innovative Regions, meetings of the CEF Expert Working Group, European Day of Languages etc.) in order to provide an opportunity for NAPs to ask their questions, foster closer collaboration and have personal contact with ELRC in a less formal setting.

Moreover, the promotion and presence of ELRC at such events is expected to increase overall visibility and promotion of the European Language Resource Coordination effort, thus encouraging interest in potential collaboration from public sector participants. The first ELRC Experience Café held on 2-3 March 2017 at the e-Sens conference in Brussels and the ELRC attendance of the EULITA Conference in Vienna on 30-31 March 2017 have already shown a first impact. Links to relevant DSI stakeholders (in particular in the legal domain) could be fostered.

ELRC Data collection

As described in Section 6 of this report only for 8 languages (English, French, German, Italian, Modern Greek, Polish, Romanian, and Spanish) more than 10 bi-/mutli-lingual corpora could be collected during the first two years of the ELRC activity. All other languages could hence, at least as an initial result of the ELRC action, count as “under-resourced” languages for which data collection efforts should be intensified in the future.

Many language resources were of general nature and could not (without detailed analysis) be directly linked to a particular CEF Digital Service Infrastructire (CEF DSI). The DSIs best covered are e-Justice (32 language resources) and Europeana (17 language resources). As a consequence, future data collection efforts should be focussed on the following CEF DSIs and areas:

- Online Dispute Resolution ODR (Area: Consumer Rights)
- Business Registration BRIS (Area: Business Registry)
- Electronic Exchange of Social Security Information (EESSI) (Area: Social Security)
- European Data Portal EDP (Area: Open Data)
- Safer Internet (Areas: Security/Cybersecurity)

Currently, only three CEF DSIs use MT@EC/eTranslation. In general, as of May 2017 four DSIs have committed to using CEF eTranslation (ODR, e-Justice, European Data Portal and EESSI), while three DSIs have committed to analyse the reuse of eTranslation (BRIS, Europeana, Safer Internet)[34]. The DSIs intending to use

[34] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/Reuse+watch

CEF eTranslation largely vary in their maturity: while ODR, e-Justice and European Data Portal are the most mature DSIs in terms of identifying their multilingual requirements, more work is needed for Safer Internet, BRIS and EESSI. For the future ELRC activities and any language resource collection activities, understanding and specifying what kind of data should be sought for each DSI is of utmost importance.

In general, there were several challenges faced by the ELRC with regard to the collection of language resources including in particular:

- Authorisation by superiors: Potential contributors require the official permission; targeted actions and official support is necessary
- Resource-related problems: data will need to be assessed by the public service administrations to see what kind of data can be contributed without any processing necessary
- Legal and licensing issues: general legal concerns and actual legal issues (mainly privacy/confidentiality issues as well as anonymization of personal data)
- Translation procedures: great variety of translation procedures which impacts on the availablility and format of language resources
- Technical problems (digitization / level of required processing):depending on the data formats available (ranging from scanned papercopies/pdfs over .doc/xls files to ready-to-use tmx-files) and contents of the LR (public contents vs. personal or confidential contents) the required processing is substantial;
- Technology affinity and lack of awareness: Limited affinity to technology by many translators.

For more details see Section 6 of this report.

While for most technical issues such as formatting or anonymization, and for legal issues such as IPR issues, confidentiality, inclusion of personal data, solutions are available or can be found (e.g. licensing and/or processing of data), for the more fundamental challenges such as authorisation, resource-related problems, general legal concerns, existing translation procedures, minimal technology affinity and lack of awareness no direct solutions are at hand. Such fundamental issues can only be overcome in close collaboration with the relevant stakeholders at all levels and through continued lobbying and dialogue. Therefore, significant further efforts with targeted outreach will have to be undertaken, particularly for the languages and in the domains of CEF DSIs in order to free the language resources needed to train CEF eTranslation in the required domains.

Outreach activities going beyond the scope of the ELRC and the ELRC Workshops could take the form of presence at relevant national events and the organisation of corresponding “Technology experience Cafés” as part of overall national roadshows that give potential data donors the opportunity to see and feel what can be achieved with regard to machine translation with the help of the right training data. Corresponding demonstrations will need to be prepared for each case.

The work and support of the National Anchor Points is and should be the corner stone of all future language resource coordination activities: 60% of the Language Resource Board members made direct data contributions to the ELRC (Annex 2). 51 out of the 54 NAPs provided the contacts to the data holders in their country and/or

corresponding sources for data so that the ELRC could proceed and extend the data collection in these countries. Therefore, the LRB seems to be a key component for a successful data collection. The network established in the course of the ELRC project is now well-established and can effectively support the collection of language resources in the participating countries.

As regards the self-sustainability of the European Language Resource Coordination, it can be concluded that self-sustainability will only be feasible if there is a clear and tangible benefit for all participating institutions, i.e. for public services administrations contributing language resources. As indicated in section 4.2 Key issues raised by stakeholders and actions, they key question of most data contributors is "*Why should I get/stay involved?*" and the answer to that can only be: "*Because it is beneficial for you*."

The contribution of language resources costs time and resources in each participating institution – so there needs to be a clear justification and "return on investment" for all institutions undertaking this activity. This situation is complicated by the fact that data holders, i.e. institutions who are in possession of relevant language resources, are not necessarily the institutions involved in the CEF Digital Service Infrastructures (DSIs). Therefore, operating models that could ensure the self-sustainability of the ELRC would need to take a broader scope: this could involve access to MT services for data donors or creating incentives for sharing language resources among institutions (i.e. if institutions contribute, they also have access to the language resources contributed by others). The exact model of operation involving both the improvement of and access to MT services and access to / mutual sharing and exchange of language resources would need to be investigated and defined as part of a corresponding concept ("business plan"). In this respect, further coordinative support would be needed to (i) maintain and extend the current ELRC network of data contributors and (ii) to develop, in collaboration with all relevant stakeholders, a corresponding concept. Finally, it also became evident from the existing activities and results of the ELRC, that with regard to the domains where no language resources are available, support will need to be provided also in the future for the generation and processing of such resources.

ANNEX

ANNEX 1: COMPOSITION OF THE LANGUAGE RESOURCE BOARD (APRIL 2017)

Last name	First name	Gender	Country	Organisational affiliation	Type
Budin	Gerhard	M	AT	Zentrum für Translationswissenschaft	Tech NAP
Nestler	Bruno	M	AT	Language Institute of the Bundesheer, Landesverteidigungsakademie	Public Services NAP
Hoste	Veronique	F	BE	Computational Linguistics and Psycholinguistics Research Centre, University of Antwerp	Tech NAP
De Smeytere	Stijn	M	BE	Prime Minister's Office	Public Services NAP
Koeva	Svetla	F	BG	Bulgarian Academy of Sciences	Tech NAP
Dobreva	Hristina	F	BG	Ministry of Transport, Information Technology and Communications, Information Technology Directorate	Public Services NAP
Hajic	Jan	M	CZ	Institute of Formal and Applied Linguistics, Charles University in Prague	Tech NAP
Cerníková	Marie	F	CZ	Ministry of education, youth and sports	Public Services NAP
Witt	Andreas	M	DE	Institut für Deutsche Sprache Mannheim	Tech NAP
Soska	Alexandra	F	DE	Federal Ministry of Inner Affairs	Public Services NAP
Sandford Pedersen	Bolette	F	DK	Centre for Language Technology, Department of Nordic Research	Tech NAP
Kirchmeier-Andersen	Sabine	F	DK	Danish Language Council	Public Services NAP
Gylden Houmann	Peter	M	DK	Agency for Digitisation	Public Services NAP
Vider	Kadri	F	EE	Estonian Language Resources Center and the University of Tartu	Tech NAP
Eessalu	Martin	M	EE	Ministry of Education and Research	Public Services NAP
Bel	Núria	F	ES	Institut Universitari de Lingüistica Aplicada, University Pompeu Fabra	Tech NAP
Pérez Fernández	David	M	ES	Gabinete del Secretario de Estado de Telecomunicaciones.	Public Services NAP

Last name	First name	Gender	Country	Organisational affiliation	Type
Linden	Krister	M	FI	Department of Modern Languages, University of Helsinki	Tech NAP
Virtanen	Taru	F	FI	Prime Minister's Office	Public Services NAP
Yvon	François	M	FR	CNRS-LIMSI	Tech NAP
Hongniat-Lange	Martine	F	FR	Ministry of Finances, Chef de traduction	Public Services NAP
Gavriilidou	Maria	F	GR	ILPS	Tech NAP
Routzouni	Nancy	F	GR	Hellenic Ministry of Interior and Administrative Reform	Public Services NAP
Tadic	Marko	M	HR	Institute of Linguistics, Faculty of Humanities and Social Science, University of Zagreb	Tech NAP
Várárdi	Tamás	M	HU	Research Institute for Linguistics, Hungarian Academy of Sciences	Tech NAP
Szenthe	Janos	M	HU	Department of EU Law at the Ministry of Justice	Public Services NAP
Way	Andy	M	IE	School of Computing, Dublin City University	Tech NAP
Ó Conaire	Micheál	M	IE	Department of Arts Heritage and Gaeltacht	Public Services NAP
Rögnvaldsson	Eirikur	M	IS	School of Humanities, University of Iceland	Tech NAP
Hauksdóttir	Auður	F	IS	Vigdís Finnbogadóttir Institute of Foreign Languages	Public Services NAP
Montemagni	Simonetta	F	IT	Consiglio Nazionale Ricerche, Istituto di Linguistica Computazionale "Antonio Zampolli"	Tech NAP
Foti	Clauda	F	IT	Ministry of Justice	Public Services NAP
Vaisniene	Daiva	F	LT	Head of The State Commission of Lithuanian Language	Tech NAP
Ras	Eric	M	LU	Luxembourg Institute of Science and Technology	Public Services NAP
Magone	Armands	M	LV	Director of Cultural Information System Agency	Public Services NAP
Gruzitis	Normunds	M	LV	Artificial Intelligence Laboratory, The Institute of Mathematics and Computer Science, University of Latvia	Tech NAP
Rosner	Michael	M	MT	Department Intelligent Computer Systems, University of Malta	Tech NAP

Last name	First name	Gender	Country	Organisational affiliation	Type
Sant	David	M	MT	Malta Information Technology Agency	Public Services NAP
Odijk	Jan	M	NL	Utrecht Institute of Linguistics, Universiteit Utrecht	Tech NAP
De Smedt	Koenraad	M	NO	Department of Linguistic, Literary and Aesthetic Studies, University of Bergen	Tech NAP
Hails Gjelsten	Sarah Jane	F	NO	Agency for Public Management and eGovernment	Public Services NAP
Ogrodniczuk	Maciej	M	PL	Institute of Computer Science, Polish Academy of Sciences	Tech NAP
Kotarska	Anna	F	PL	National Health Fund, Department of Analysis and Strategy	Public Services NAP
Branco	Antonio	M	PT	Department of Informatics, University of Lisbon	Tech NAP
Vale	Paulo	M	PT	AMA - Agência para a Modernização Administrativa, I.P., Presidência do Conselho de Ministros	Public Services NAP
Tufis	Dan	M	RO	Research Institute for Artificial Intelligence, Romanian Academy of Sciences	Tech NAP
Mihailescu	Laura	F	RO	Head of the Department Translations coordination at the European Institute in România	Public Services NAP
Borin	Lars	M	SE	Department of Swedish Language, University of Gothenburg	Tech NAP
Domeij	Rickard	M	SE	Institute for Language and Folklore / Swedish Language Council	Public Services NAP
Krek	Simon	M	SI	Jozef Stefan Institute	Tech NAP
Novljan Lovrincic	Marcela	F	SI	Secretariat - General of the Government of the Republic of Slovenia	Public Services NAP
Zumrík	Miroslav	M	SK	Ludovit Stur Institute of Linguistics, Slovak Academy of Sciences	Tech NAP
Snircova	Diana	F	SK	Department of the National Language, Ministry of Culture of the Slovak Republic	Public Services NAP
Ananiadou	Sophia	F	UK	School of Computer Science, University of Manchester	Tech NAP

ANNEX 2: NATIONAL ANCHOR POINT (NAP) CONTRIBUTION TO DATA COLLECTION

Last name	First name	Country	Type	Data collection sub-contract	Data contri-butions without sub-contract	Provision of contacts to data holders
Budin	Gerhard	AT	Tech NAP	yes		yes
Nestler	Bruno	AT	Public Services NAP	yes		yes
Hoste	Veronique	BE	Tech NAP	yes		yes
De Smeytere	Stijn	BE	Public Services NAP			yes
Koeva	Svetla	BG	Tech NAP	yes		yes
Dobreva	Hristina	BG	Public Services NAP	yes		yes
Hajic	Jan	CZ	Tech NAP		yes	
Cerníková	Marie	CZ	Public Services NAP		yes	
Witt	Andreas	DE	Tech NAP			
Soska	Alexandra	DE	Public Services NAP		yes	
Sandford Pedersen	Bolette	DK	Tech NAP	yes		
Kirchmeier-Andersen	Sabine	DK	Public Services NAP			yes
Gylden Houmann	Peter	DK	Public Services NAP			yes
Vider	Kadri	EE	Tech NAP			yes
Eessalu	Martin	EE	Public Services NAP			yes
Bel	Núria	ES	Tech NAP	yes		yes
Pérez Fernández	David	ES	Public Services NAP			yes
Linden	Krister	FI	Tech NAP	yes		yes
Virtanen	Taru	FI	Public Services NAP		yes	yes
Yvon	François	FR	Tech NAP			
Hongniat-Lange	Martine	FR	Public Services NAP	yes		yes
Gavriilidou	Maria	GR	Tech NAP		yes	
Routzouni	Nancy	GR	Public Services NAP			yes
Tadic	Marko	HR	Tech NAP	yes		yes
Várárdi	Tamás	HU	Tech NAP	yes		yes
Szenthe	Janos	HU	Public Services NAP			yes
Way	Andy	IE	Tech NAP	yes		
Ó Conaire	Micheál	IE	Public Services NAP		yes	yes
Rögnvaldsson	Eirikur	IS	Tech NAP			yes

Last name	First name	Country	Type	Data collection sub-contract	Data contri-butions without sub-contract	Provision of contacts to data holders
Hauksdóttir	Auður	IS	Public Services NAP		yes	
Montemagni	Simonetta	IT	Tech NAP	yes		yes
Foti	Claudia	IT	Public Services NAP		yes	yes
Vaisniene	Daiva	LT	Tech NAP		yes	yes
Ras	Eric	LU	Public Services NAP		yes	
Magone	Armands	LV	Public Services NAP			yes
Gruzitis	Normunds	LV	Tech NAP			yes
Rosner	Michael	MT	Tech NAP	yes		yes
Sant	David	MT	Public Services NAP			yes
Odijk	Jan	NL	Tech NAP	yes		yes
De Smedt	Koenraad	NO	Tech NAP			yes
Hails Gjelsten	Sarah Jane	NO	Public Services NAP		yes	yes
Ogrodniczuk	Maciej	PL	Tech NAP	yes		yes
Kotarska	Anna	PL	Public Services NAP			yes
Branco	Antonio	PT	Tech NAP			yes
Vale	Paulo	PT	Public Services NAP		yes	
Tufis	Dan	RO	Tech NAP	yes		yes
Mihailescu	Laura	RO	Public Services NAP			yes
Borin	Lars	SE	Tech NAP			yes
Domeij	Rickard	SE	Public Services NAP	yes		yes
Krek	Simon	SI	Tech NAP	yes		
Novljan Lovrincic	Marcela	SI	Public Services NAP			yes
Zumrík	Miroslav	SK	Tech NAP	yes		yes
Snircova	Diana	SK	Public Services NAP			yes
Ananiadou	Sophia	UK	Tech NAP			

ANNEX 3: LIST OF LANGUAGE RESOURCES DELIVERED BY THE ELRC

A total of 225 language resources were delivered by the ELRC to CEF Automated Translation. They are listed below, including also details on the type of language resource, the language(s) covered, and the legal status (licensing type).

ID#	Resource Name	Directory Name	Type	Language(s)	Legal Status
15	International Agreements	ELRC_15_International Agreements	Multilingual Corpus	English | Latvian	CC-BY-SA_4.0
16	Corpus of State-related content from the Latvian Web	ELRC_16_Corpus of State-related content from the	Multilingual Corpus	English | Latvian	CC-BY-SA_4.0
25	Website of the President of the Republic of Lithuania	ELRC_25_Website of the President of the	Multilingual Corpus	English | Lithuanian	CC-BY-SA_4.0
26	Verbatim reports of Saeima of the Republic of Latvia	ELRC_26_Verbatim reports of Saeima of the	Monolingual Corpus	Latvian	CC-BY-SA_4.0
27	POLSIS - Database of policy planning documents	ELRC_27_POLSIS - Database of policy planning	Monolingual Corpus	Latvian	CC-BY-SA_4.0
30	Audioguide for the Military History Museum in Vienna	ELRC_30_Audioguide for the Military History Museum	Multilingual Corpus	German | Italian	openUnder_PSI
31	BMI Brochures 2011-2015	ELRC_31_BMI Brochures 2011-2015	Multilingual Corpus	English | German	openUnder_PSI
36	Glossary City of Vienna	ELRC_36_Glossary City of Vienna	Terminology	English | German	openUnder_PSI
41	German-English website parallel corpus from the Federal Foreign Office Berlin	ELRC_41_German-English website parallel corpus from the	Multilingual Corpus	English | German	openUnder_PSI
42	German-French website parallel corpus from the Federal Foreign Office Berlin	ELRC_42_German-French website parallel corpus from the	Multilingual Corpus	French | German	openUnder_PSI
43	German-Portuguese website parallel corpus from the Federal Foreign Office Berlin	ELRC_43_German-Portuguese website parallel corpus from the	Multilingual Corpus	German | Portuguese	openUnder_PSI
49	OROSSIMO Corpus - Economics	ELRC_49_OROSSIMO Corpus - Economics	Monolingual Corpus	Modern Greek (1453-)	CC-BY_4.0
58	OROSSIMO Corpus - Medicine & health	ELRC_58_OROSSIMO Corpus - Medicine & health	Monolingual Corpus	Modern Greek (1453-)	CC-BY_4.0
61	Parallel Global Voices (Greek - English)	ELRC_61_Parallel Global Voices (Greek - English)	Multilingual Corpus	English | Modern Greek (1453-)	CC-BY_4.0
64	Orossimo Terminological Resource - Photography, film & video	ELRC_64_Orossimo Terminological Resource - Photography, film	Terminology	English | Modern Greek (1453-)	CC-BY_4.0

ID#	Resource Name	Directory Name	Type	Language(s)	Legal Status
65	Orossimo Terminological Resource - Economics	ELRC_65_Orossimo Terminological Resource - Economics	Terminology	English | Modern Greek (1453-)	CC-BY_4.0
66	Orossimo Terminological Resource - Computer Science	ELRC_66_Orossimo Terminological Resource - Computer Science	Terminology	English | Modern Greek (1453-)	CC-BY_4.0
68	Orossimo Terminological Resource - Law	ELRC_68_Orossimo Terminological Resource - Law	Terminology	English | Modern Greek (1453-)	CC-BY_4.0
88	Documents concerning Federal Constitutional Law in Austria	ELRC_88_Documents concerning Federal Constitutional Law in	Multilingual Corpus	English | German	openUnder_PSI
89	Austrian Armed Forces Military Dictionaries	ELRC_89_Austrian Armed Forces Military Dictionaries	Terminology	English | French | German | Hungarian | Italian	openUnder_PSI
90	ANR translation memory containing major publications, as well as several administrative documents and news	ELRC_90_ANR translation memory containing major publications,	Multilingual Corpus	English | French	openUnder_PSI
93	Austrian Criminal Office Police Glossary	ELRC_93_Austrian Criminal Office Police Glossary	Terminology	English | German	openUnder_PSI
108	Trilingual Documents related to International Judicial Cooperation in Civil Matters (Greek-English-French)	ELRC_108_Trilingual Documents related to International Judicial	Multilingual Corpus	English | French | Modern Greek (1453-)	CC-BY_4.0
110	Press and Information Office (PIO) Publication: "CYPRUS still occupied still divided 1974-2016"	ELRC_110_Press and Information Office (PIO) Publication	Multilingual Corpus	English | Modern Greek (1453-)	CC-BY_4.0
111	Cyprus at a glance	ELRC_111_Cyprus at a glance	Multilingual Corpus	English | French | Italian | Modern Greek (1453-) | Spanish; Castilian	CC-BY_4.0
112	Letter of rights for persons arrested on the basis of a European Arrest Warrant	ELRC_112_Letter of rights for persons arrested	Multilingual Corpus	Bulgarian | Dutch; Flemish | English | French | German | Italian | Latvian | Modern Greek (1453-) | Polish | Romanian; Moldavian; Moldovan	CC-BY_4.0
113	Letter of rights for persons arrested and or detained	ELRC_113_Letter of rights for persons arrested	Multilingual Corpus	Bulgarian | English | French | Latvian | Modern Greek (1453-) | Polish | Romanian; Moldavian; Moldovan	CC-BY_4.0

ID#	Resource Name	Directory Name	Type	Language(s)	Legal Status
114	Greek anti-corruption legislation and National Anti-Corruption Plan (greek-english)	ELRC_114_Greek anti-corruption legislation and National Anti-Corruption	Multilingual Corpus	English | Modern Greek (1453-)	CC-BY_4.0
115	Parallel corpus (Greek - English) in the public administration domain	ELRC_115_Parallel corpus (Greek - English) in	Multilingual Corpus	English | Modern Greek (1453-)	openUnder_PSI
124	OECD Anti - Bribery Convention (English - Greek)	ELRC_124_OECD Anti - Bribery Convention (English	Multilingual Corpus	English | Modern Greek (1453-)	CC-BY_4.0
127	Central Statistical Office Dataset	ELRC_127_Central Statistical Office Dataset	Multilingual Corpus	English | Polish	CC-BY_4.0
128	PKN Orlen Dataset	ELRC_128_PKN Orlen Dataset	Multilingual Corpus	English | Polish	CC-BY_4.0
129	Natolin European Centre Dataset	ELRC_129_Natolin European Centre Dataset	Multilingual Corpus	English | Polish	CC-BY_4.0
133	Monolingual Bulgarian corpus in the culture domain	ELRC_133_Monolingual Bulgarian corpus in the culture	Monolingual Corpus	Bulgarian	underReview
135	Monolingual Polish corpus in the culture domain	ELRC_135_Monolingual Polish corpus in the culture	Monolingual Corpus	Polish	underReview
136	Monolingual Romanian corpus in the culture domain	ELRC_136_Monolingual Romanian corpus in the culture	Monolingual Corpus	Romanian; Moldavian; Moldovan	underReview
137	Monolingual Romanian corpus in the public administration domain	ELRC_137_Monolingual Romanian corpus in the public	Monolingual Corpus	Romanian; Moldavian; Moldovan	openUnder_PSI
140	Monolingual Polish corpus in the public administration domain	ELRC_140_Monolingual Polish corpus in the public	Monolingual Corpus	Polish	openUnder_PSI
141	Monolingual Greek corpus in the public administration domain	ELRC_141_Monolingual Greek corpus in the public	Monolingual Corpus	Modern Greek (1453-)	openUnder_PSI
142	Monolingual Bulgarian corpus in the public administration domain	ELRC_142_Monolingual Bulgarian corpus in the public	Monolingual Corpus	Bulgarian	openUnder_PSI
148	Bilingual Croatian-English Parallel Corpus	ELRC_148_Bilingual Croatian-English Parallel Corpus	Multilingual Corpus	Croatian | English	openUnder_PSI
150	Monolingual Polish corpus in the law domain	ELRC_150_Monolingual Polish corpus in the law	Monolingual Corpus	Polish	underReview
151	Monolingual Romanian corpus in the law domain	ELRC_151_Monolingual Romanian corpus in the law	Monolingual Corpus	Romanian; Moldavian; Moldovan	underReview

ID#	Resource Name	Directory Name	Type	Language(s)	Legal Status
152	Parallel corpus (Greek - English) in the law domain	ELRC_152_Parallel corpus (Greek - English) in	Multilingual Corpus	English | Modern Greek (1453-)	underReview
156	Parallel Corpus from the Web Site of the the MFA of Latvia	ELRC_156_Parallel Corpus from the Web Site	Multilingual Corpus	English | Latvian	CC-BY-SA_4.0
157	Translation of the Luxembourg.lu web site	ELRC_157_Translation of the Luxembourg.lu web site	Multilingual Corpus	English | French | German	openUnder_PSI
158	SIP Publications	ELRC_158_SIP Publications	Multilingual Corpus	English | French | German	openUnder_PSI
159	SIP Internal dictionary	ELRC_159_SIP Internal dictionary	Terminology	English | French | German	openUnder_PSI
160	SIP Dictionary of places and people (Luxembourg)	ELRC_160_SIP Dictionary of places and people	Terminology	English | French | German	openUnder_PSI
161	Polish Food Dataset	ELRC_161_Polish Food Dataset	Multilingual Corpus	English | Polish	CC-BY_4.0
162	PAH_Oxfam Dataset	ELRC_162_PAH_Oxfam Dataset	Multilingual Corpus	English | Polish	non-standard/Other_Licence/Terms
163	National Health Fund Dataset	ELRC_163_National Health Fund Dataset	Multilingual Corpus	English | Polish	openUnder_PSI
174	GRECO (Council of Europe) Reports on Greece (English - French - Greek)	ELRC_174_GRECO (Council of Europe) Reports on	Multilingual Corpus	English | French | Modern Greek (1453-)	openUnder_PSI
175	Portuguese-French bilingual corpus from Portuguese law on referendum	ELRC_175_Portuguese-French bilingual corpus from Portuguese law	Multilingual Corpus	French | Portuguese	openUnder_PSI
176	Portuguese-English bilingual corpus from Legislation concerning the Portuguese Parliament	ELRC_176_Portuguese-English bilingual corpus from Legislation concerning	Multilingual Corpus	English | Portuguese	openUnder_PSI
177	Portuguese-English bilingual corpus from the Portuguese Constitution	ELRC_177_Portuguese-English bilingual corpus from the Portuguese	Multilingual Corpus	English | Portuguese	openUnder_PSI
178	Legislation PT	ELRC_178_Legislation PT	Monolingual Corpus	Portuguese	openUnder_PSI
179	Portuguese legislation in EN	ELRC_179_Portuguese legislation in EN	Monolingual Corpus	English	openUnder_PSI
180	Portuguese legislation in FR	ELRC_180_Portuguese legislation in FR	Monolingual Corpus	French	openUnder_PSI
182	Bilingual documents Bulgarian-English in the field of transport	ELRC_182_Bilingual documents Bulgarian-English in the field	Multilingual Corpus	Bulgarian | English	openUnder_PSI

ID#	Resource Name	Directory Name	Type	Language(s)	Legal Status
183	Romanian Ombudsman archive	ELRC_183_Romanian Ombudsman archive	Multilingual Corpus	English | Romanian; Moldavian; Moldovan	CC-BY_4.0
190	Bilingual resource with Bulgarian strategic documents in the field of telecommunications and broadband (Bulgarian - English)	ELRC_190_Bilingual resource with Bulgarian strategic documents	Multilingual Corpus	Bulgarian | English	openUnder_PSI
191	Bilingual resource with Bulgarian strategic documents in the field of innovations and digital growth (Bulgarian - English)	ELRC_191_Bilingual resource with Bulgarian strategic documents	Multilingual Corpus	Bulgarian | English	openUnder_PSI
192	2015 Calls for Tenders for Translation	ELRC_192_2015 Calls for Tenders for Translation	Monolingual Corpus	Dutch; Flemish	openUnder_PSI
193	Romanian – English literature corpus	ELRC_193_Romanian - English literature corpus	Multilingual Corpus	English | Romanian; Moldavian; Moldovan	CC-BY_4.0
194	General Romanian-English bilingual corpus	ELRC_194_General Romanian-English bilingual corpus	Multilingual Corpus	English | Romanian; Moldavian; Moldovan	CC-BY-SA_3.0
195	Romanian – English New Criminal Procedure Code	ELRC_195_Romanian - English New Criminal Procedure	Multilingual Corpus	English | Romanian; Moldavian; Moldovan	CC-BY_4.0
196	Romanian New Civil Procedure Code	ELRC_196_Romanian New Civil Procedure Code	Monolingual Corpus	Romanian; Moldavian; Moldovan	CC-BY_4.0
197	Macroeconomic Developments	ELRC_197_Macroeconomic Developments	Multilingual Corpus	English | Modern Greek (1453-)	CC-BY_4.0
198	Methodological Reconciliation	ELRC_198_Methodological Reconciliation	Multilingual Corpus	English | Modern Greek (1453-)	CC-BY_4.0
199	Expression of interest	ELRC_199_Expression of interest	Multilingual Corpus	English | Modern Greek (1453-)	underReview
200	Romanian – English news corpus	ELRC_200_Romanian - English news corpus	Multilingual Corpus	English | Romanian; Moldavian; Moldovan	CC-BY_4.0
204	OROSSIMO Corpus - Computer Science	ELRC_204_OROSSIMO Corpus - Computer Science	Monolingual Corpus	Modern Greek (1453-)	CC-BY_4.0
205	OROSSIMO Corpus - Law	ELRC_205_OROSSIMO Corpus - Law	Monolingual Corpus	Modern Greek (1453-)	CC-BY_4.0
210	OROSSIMO Corpus - Photography, film & video	ELRC_210_OROSSIMO Corpus - Photography, film &	Monolingual Corpus	Modern Greek (1453-)	CC-BY_4.0

ID#	Resource Name	Directory Name	Type	Language(s)	Legal Status
213	Hallituskausi 2007-2011 fi-en	ELRC_213_Hallituskausi 2007-2011 fi-en	Multilingual Corpus	English | Finnish	CC-BY_4.0
214	Hallituskausi 2011-2015 fi-en	ELRC_214_Hallituskausi 2011-2015 fi-en	Multilingual Corpus	English | Finnish	CC-BY_4.0
219	Orossimo Terminological Resource - Medicine & health	ELRC_219_Orossimo Terminological Resource - Medicine &	Terminology	English | Modern Greek (1453-)	CC-BY_4.0
221	National Bank of Belgium Terminology	ELRC_221_National Bank of Belgium Terminology	Terminology	Dutch; Flemish | English | French | German	openUnder_PSI
223	Belgian government bilingual parallel corpus	ELRC_223_Belgian government bilingual parallel corpus	Multilingual Corpus	Dutch; Flemish | French	openUnder_PSI
224	Memorandum for a ESM programme	ELRC_224_Memorandum for a ESM programme	Multilingual Corpus	English | Modern Greek (1453-)	CC-BY_4.0
225	Convention on the transfer of sentenced persons (English - Greek)	ELRC_225_Convention on the transfer of sentenced	Multilingual Corpus	English | Modern Greek (1453-)	CC-BY_4.0
226	Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment - United Nations (French-English-Greek)	ELRC_226_Convention against Torture and Other Cruel,	Multilingual Corpus	English | French | Modern Greek (1453-)	CC-BY_4.0
227	Collection of Greek National Spatial Plans	ELRC_227_Collection of Greek National Spatial Plans	Monolingual Corpus	Modern Greek (1453-)	CC-BY_4.0
229	Romanian – English parallel wordlists	ELRC_229_Romanian - English parallel wordlists	Terminology	English | Romanian; Moldavian; Moldovan	CC-BY_4.0
230	Term dictionary (Business and Competition) from Swedish Authorities	ELRC_230_Term dictionary (Business and Competition) from	Terminology	English | Swedish	CC-ZERO
231	Term dictionary (Agriculture) from Swedish Authorities	ELRC_231_Term dictionary (Agriculture) from Swedish Authorities	Terminology	English | Swedish	CC-ZERO
232	Term dictionary (Law) from Swedish Authorities	ELRC_232_Term dictionary (Law) from Swedish Authorities	Terminology	English | Swedish	CC-ZERO
233	BMI Brochures and Website 2016	ELRC_233_BMI Brochures and Website 2016	Multilingual Corpus	English | German	openUnder_PSI
234	BMI Brochure Civil Protection	ELRC_234_BMI Brochure Civil Protection	Multilingual Corpus	English | German	openUnder_PSI

ID#	Resource Name	Directory Name	Type	Language(s)	Legal Status
235	Parallel corpus from Social Insurance Agency - Socialstyrelsen (Sweden)	ELRC_235_Parallel corpus from Social Insurance Agency	Terminology	Swedish	CC-BY_4.0
239	Term dictionary (Police) from Swedish Authorities	ELRC_239_Term dictionary (Police) from Swedish Authorities	Terminology	English | Swedish	CC-ZERO
240	Parallel corpus from Social Insurance Agency -- Försäkringskassan (Sweden)	ELRC_240_Parallel corpus from Social Insurance Agency	Multilingual Corpus	English | Swedish	CC-ZERO
241	Medicines descriptions in English and Icelandic from the European Medicines Agency	ELRC_241_Medicines descriptions in English and Icelandic	Multilingual Corpus	English | Icelandic	non-standard/Other_Licence/Terms
242	Medicines descriptions in English and Norwegian from the European Medicines Agency	ELRC_242_Medicines descriptions in English and Norwegian	Multilingual Corpus	English | Norwegian	non-standard/Other_Licence/Terms
243	English-Danish Parallel corpus from Tatoeba project	ELRC_243_English-Danish Parallel corpus from Tatoeba project	Multilingual Corpus	Danish | English	CC-BY_4.0
244	Parallel corpus from Parliament of Estonia	ELRC_244_Parallel corpus from Parliament of Estonia	Multilingual Corpus	English | Estonian	CC-BY_4.0
245	Parallel corpus from Estonian Cabinet of Ministers	ELRC_245_Parallel corpus from Estonian Cabinet of	Multilingual Corpus	English | Estonian	CC-BY_4.0
246	Parallel corpus from Estonian Ministry of Foreign Affairs	ELRC_246_Parallel corpus from Estonian Ministry of	Multilingual Corpus	English | Estonian	CC-BY_4.0
247	Parallel corpus from Bank of Estonia	ELRC_247_Parallel corpus from Bank of Estonia	Multilingual Corpus	English | Estonian	CC-BY_4.0
248	Corpus on Finance and Economics from Bank of Latvia	ELRC_248_Corpus on Finance and Economics from	Multilingual Corpus	English | Latvian	CC-BY-SA_4.0
249	Opinions of the CONSULTATIVE COUNCIL OF EUROPEAN JUDGES in Bulgarian	ELRC_249_Opinions of the CONSULTATIVE COUNCIL OF	Monolingual Corpus	Bulgarian	openUnder_PSI
250	Priorities and schedule of the Dutch Presidency of the EU (in Bulgarian)	ELRC_250_Priorities and schedule of the Dutch	Monolingual Corpus	Bulgarian	openUnder_PSI
251	Guidelines - Judicial maps in Bulgarian	ELRC_251_Guidelines - Judicial maps in Bulgarian	Monolingual Corpus	Bulgarian	openUnder_PSI
252	Statistics Finland translation memory Finnish-English	ELRC_252_Statistics Finland translation memory Finnish-English	Multilingual Corpus	English | Finnish	non-standard/Other_Licence/Terms

ID#	Resource Name	Directory Name	Type	Language(s)	Legal Status
253	Statistics Finland's Finnish to Swedish translation memory	ELRC_253_Statistics Finland's Finnish to Swedish translation	Multilingual Corpus	Finnish | Swedish	non-standard/Other_Licence/Terms
254	Term lists and Dictionaries from Swedish Authorities	ELRC_254_Term lists and Dictionaries from Swedish	Terminology	English | Finnish | French | German | Modern Greek (1453-) | Spanish; Castilian | Swedish	non-standard/Other_Licence/Terms
255	BMVI Website	ELRC_255_BMVI Website	Multilingual Corpus	English | German	openUnder_PSI
256	BMVI Publications	ELRC_256_BMVI Publications	Multilingual Corpus	English | German	openUnder_PSI
257	Parallel texts from Swedish Labour market agency. Part 2	ELRC_257_Parallel texts from Swedish Labour market	Multilingual Corpus	English | Finnish | French | German | Polish | Romanian; Moldavian; Moldovan | Spanish; Castilian | Swedish	non-standard/Other_Licence/Terms
258	Parallel texts from Swedish Social Security Authority	ELRC_258_Parallel texts from Swedish Social Security	Multilingual Corpus	Croatian | English | Finnish | French | German | Italian | Polish | Romanian; Moldavian; Moldovan | Spanish; Castilian | Swedish	non-standard/Other_Licence/Terms
259	Parallel texts from Swedish Labour market agency	ELRC_259_Parallel texts from Swedish Labour market	Multilingual Corpus	English | Finnish | French | German | Romanian; Moldavian; Moldovan | Spanish; Castilian | Swedish	underReview
260	Parallel texts from Swedish National Food Agency	ELRC_260_Parallel texts from Swedish National Food	Multilingual Corpus	English | Finnish | French | Polish | Spanish; Castilian | Swedish	non-standard/Other_Licence/Terms
261	Parallel texts from Swedish Work environment Authority	ELRC_261_Parallel texts from Swedish Work environment	Multilingual Corpus	Bulgarian | Czech | English | Estonian | Finnish | French | German | Hungarian | Italian | Latvian | Lithuanian | Modern Greek (1453-) | Polish | Romanian; Moldavian; Moldovan | Spanish; Castilian | Swedish	non-standard/Other_Licence/Terms
262	Parallel Global Voices (Greek - French)	ELRC_262_Parallel Global Voices (Greek - French)	Multilingual Corpus	French | Modern Greek (1453-)	CC-BY_4.0

ID#	Resource Name	Directory Name	Type	Language(s)	Legal Status
263	Parallel Global Voices (Greek - Spanish)	ELRC_263_Parallel Global Voices (Greek - Spanish)	Multilingual Corpus	Modern Greek (1453-) | Spanish; Castilian	CC-BY_4.0
264	The Gaois bilingual corpus of English-Irish legislation	ELRC_264_The Gaois bilingual corpus of English-Irish	Multilingual Corpus	English | Irish	openUnder_PSI
265	PAeSI : Public Administration and Foreign Immigrants	ELRC_265_PAeSI Public Administration and Foreign	Multilingual Corpus	English | French | Italian | Spanish; Castilian	non-standard/Other_Licence/Terms
266	CHARTER OF VALUES OF CITIZENSHIP AND INTEGRATION	ELRC_266_CHARTER OF VALUES OF CITIZENSHIP AND	Multilingual Corpus	English | French | German | Italian | Spanish; Castilian	non-standard/Other_Licence/Terms
267	INFORMATION FOR VICTIMS OF A CRIME	ELRC_267_INFORMATION FOR VICTIMS OF A CRIME	Multilingual Corpus	English | French | German | Italian | Spanish; Castilian	non-standard/Other_Licence/Terms
268	Corpus EPTIC	ELRC_268_Corpus EPTIC	Multilingual Corpus	English | French | Italian	non-standard/Other_Licence/Terms
269	Corpora of legal text	ELRC_269_Corpora of legal text	Multilingual Corpus	English | Italian	non-standard/Other_Licence/Terms
270	Legal Texts	ELRC_270_Legal Texts	Multilingual Corpus	English | French | Italian	non-standard/Other_Licence/Terms
271	Norwegian-English Dictionary for Primary education	ELRC_271_Norwegian-English Dictionary for Primary education	Terminology	English | Norwegian	non-standard/Other_Licence/Terms
272	The Coimisineir Teanga Bilingual Web Corpus	ELRC_272_The Coimisineir Teanga Bilingual Web Corpus	Multilingual Corpus	English | Irish	openUnder_PSI
273	The Coimisineir Teanga Bilingual Corpus of Reference Documents	ELRC_273_The Coimisineir Teanga Bilingual Corpus of	Multilingual Corpus	English | Irish	openUnder_PSI
274	The Coimisineir Teanga Bilingual Corpus of Reports and Press Releases	ELRC_274_The Coimisineir Teanga Bilingual Corpus of	Multilingual Corpus	English | Irish	openUnder_PSI
278	Translations of Lithuanian legislation from Seimas of the Republic of Lithuania	ELRC_278_Translations of Lithuanian legislation from Seimas	Multilingual Corpus	English | Lithuanian	CC-BY_4.0
279	Legal texts from Estonian Ministry of Justice	ELRC_279_Legal texts from Estonian Ministry of	Multilingual Corpus	English | Estonian	CC-BY_4.0

ID#	Resource Name	Directory Name	Type	Language(s)	Legal Status
280	Belgian parallel corpus about taxes, economy, housing and work	ELRC_280_Belgian parallel corpus about taxes, economy,	Multilingual Corpus	Dutch; Flemish | French	openUnder_PSI
281	Belgian parallel corpus about education, health and environment	ELRC_281_Belgian parallel corpus about education, health	Multilingual Corpus	Dutch; Flemish | French	openUnder_PSI
282	Belgian parallel corpus about Belgium and the justice system	ELRC_282_Belgian parallel corpus about Belgium and	Multilingual Corpus	Dutch; Flemish | French	openUnder_PSI
283	Fiscal Dictionary	ELRC_283_Fiscal Dictionary	Terminology	English | Modern Greek (1453-)	CC-BY_4.0
285	Secretariat-General parallel corpus SL-EN and EN-SL (part 1)	ELRC_285_Secretariat-General parallel corpus SL-EN and EN-SL	Multilingual Corpus	English | Slovenian	openUnder_PSI
286	Secretariat-General parallel corpus SL-EN and EN-SL (part 2)	ELRC_286_Secretariat-General parallel corpus SL-EN and EN-SL	Multilingual Corpus	English | Slovenian	openUnder_PSI
287	Unofficial Consolidated legislative texts (Slovene)	ELRC_287_Unofficial Consolidated legislative texts (Slovene)	Monolingual Corpus	Slovenian	CC-BY_4.0
288	Czech Association of Medical Physicists - Physics Glossary	ELRC_288_Czech Association of Medical Physicists -	Terminology	Czech | English	openUnder_PSI
289	ISAP Legal Terminology	ELRC_289_ISAP Legal Terminology	Terminology	Czech | English | French | German	openUnder_PSI
290	Czech Banking Association Terminology	ELRC_290_Czech Banking Association Terminology	Terminology	Czech | English	openUnder_PSI
292	PaWaC - Public Administration Web as Corpus	ELRC_292_PaWaC - Public Administration Web as	Monolingual Corpus	Italian	non-standard/Other_Licence/Terms
294	Polish Food 4 & Food Policy Dataset	ELRC_294_Polish Food 4 & Food Policy	Multilingual Corpus	English | Polish	CC-BY_4.0
295	Polish Food Dataset 2	ELRC_295_Polish Food Dataset 2	Multilingual Corpus	English | Polish	CC-BY_4.0
296	Polish Food DataSet 3	ELRC_296_Polish Food DataSet 3	Multilingual Corpus	English | Polish	CC-BY_4.0
297	Polish-English Internal Aviation Glossaries	ELRC_297_Polish-English Internal Aviation Glossaries	Terminology	English | Polish	openUnder_PSI
298	Multilingual Public Procurement Terminology	ELRC_298_Multilingual Public Procurement Terminology	Terminology	Danish | English | Finnish | French | German | Italian | Modern Greek (1453-) | Polish | Portuguese | Spanish; Castilian | Swedish	openUnder_PSI

ID#	Resource Name	Directory Name	Type	Language(s)	Legal Status
299	Polish Ministry of Foreign Affairs Regional Dataset	ELRC_299_Polish Ministry of Foreign Affairs Regional	Multilingual Corpus	English | Polish	openUnder_PSI
300	EJTN Handbook	ELRC_300_EJTN Handbook	Multilingual Corpus	Bulgarian | English	openUnder_PSI
301	Monolingual Greek corpus in the culture domain	ELRC_301_Monolingual Greek corpus in the culture	Monolingual Corpus	Modern Greek (1453-)	underReview
302	English-Finnish corpus from Finnish Information Bank	ELRC_302_English-Finnish corpus from Finnish Information Bank	Multilingual Corpus	English | Finnish	CC-BY_4.0
303	English-Swedish corpus from Finnish Information Bank	ELRC_303_English-Swedish corpus from Finnish Information Bank	Multilingual Corpus	English | Swedish	CC-BY_4.0
304	English-Estonian corpus from Finnish Information Bank	ELRC_304_English-Estonian corpus from Finnish Information Bank	Multilingual Corpus	English | Estonian	CC-BY_4.0
305	Translation memories from The Ministry of Foreign Affairs of Norway	ELRC_305_Translation memories from The Ministry of	Multilingual Corpus	English | Norwegian	CC-BY_4.0
308	Newsletter TRESOR ECONOMICS 2016 Part 1	ELRC_308_Newsletter TRESOR ECONOMICS 2016 Part 1	Multilingual Corpus	English | French	non-standard/Other_Licence/Terms
309	Newsletter TRESOR ECONOMICS 2016 Part 2	ELRC_309_Newsletter TRESOR ECONOMICS 2016 Part 2	Multilingual Corpus	English | French	non-standard/Other_Licence/Terms
310	English-Bulgarian Legal Terms	ELRC_310_English-Bulgarian Legal Terms	Terminology	Bulgarian | English	CC-BY-NC_4.0
311	Newsletter TRESOR-ECONOMICS 2017	ELRC_311_Newsletter TRESOR-ECONOMICS 2017	Multilingual Corpus	English | French	non-standard/Other_Licence/Terms
312	English-Bulgarian Computer Terms	ELRC_312_English-Bulgarian Computer Terms	Terminology	Bulgarian | English	CC-BY-NC_4.0
313	Newsletter TRESOR ECONOMIC - 1 (2012-2013-2014-2015)	ELRC_313_Newsletter TRESOR ECONOMIC - 1 (2012-2013-2014-2015)	Multilingual Corpus	English | French	non-standard/Other_Licence/Terms
314	Newsletter TRESOR ECONOMICS 2016 Part 3	ELRC_314_Newsletter TRESOR ECONOMICS 2016 Part 3	Multilingual Corpus	English | French	non-standard/Other_Licence/Terms
315	English-Bulgarian Economy Terms	ELRC_315_English-Bulgarian Economy Terms	Terminology	Bulgarian | English	CC-BY-NC_4.0
316	Polish Ministry of Foreign Affairs Historical Dataset	ELRC_316_Polish Ministry of Foreign Affairs Historical	Multilingual Corpus	English | Polish	openUnder_PSI

ID#	Resource Name	Directory Name	Type	Language(s)	Legal Status
317	Polish Court Rulings Corpus	ELRC_317_Polish Court Rulings Corpus	Monolingual Corpus	Polish	openUnder_PSI
318	Polish Ministry of Foreign Affairs Youth 2011 Report	ELRC_318_Polish Ministry of Foreign Affairs Youth	Multilingual Corpus	English | Polish	openUnder_PSI
319	Public Procurement Dataset 2	ELRC_319_Public Procurement Dataset 2	Multilingual Corpus	English | Polish	openUnder_PSI
320	Civil Aviation Regulations	ELRC_320_Civil Aviation Regulations	Multilingual Corpus	English | Polish	openUnder_PSI
321	Public Procurement Dataset 1	ELRC_321_Public Procurement Dataset 1	Multilingual Corpus	English | Polish	openUnder_PSI
322	Health Multilingual Terminologies	ELRC_322_Health Multilingual Terminologies	Terminology	English | French | German | Italian | Spanish; Castilian	CC-BY-ND_4.0
323	The Foclóir New English-Irish Dictionary	ELRC_323_The Focl¢ir New English-Irish Dictionary	Terminology	English | Irish	non-standard/Other_Licence/Terms
324	The UCD Bórd na Gaeilge Corpus of bilingual PDFs and Word documents	ELRC_324_The UCD B¢rd na Gaeilge Corpus	Multilingual Corpus	English | Irish	openUnder_PSI
325	University of Vienna Termbanks	ELRC_325_University of Vienna Termbanks	Terminology	Croatian | English | French | German | Slovenian	openUnder_PSI
326	Glossaries created for "AAA Offresi" project	ELRC_326_Glossaries created for _AAA Offresi_ project	Terminology	Italian | Spanish; Castilian	non-standard/Other_Licence/Terms
327	Glossary: Terminology of cadastral services	ELRC_327_Glossary Terminology of cadastral services	Terminology	German | Italian	non-standard/Other_Licence/Terms
328	Slovak corpus of texts from the Ministry of Culture of the Slovak Republic	ELRC_328_Slovak corpus of texts from the	Monolingual Corpus	Slovak	openUnder_PSI
329	English-Slovak parallel corpus of texts from The Ministry of Culture of the Slovak Republic	ELRC_329_English-Slovak parallel corpus of texts from	Multilingual Corpus	English | Slovak	openUnder_PSI
330	Slovak corpus of texts from the Ministry of Justice of the Slovak Republic	ELRC_330_Slovak corpus of texts from the	Monolingual Corpus	Slovak	openUnder_PSI
331	English-Slovak parallel corpus of texts from The Ministry of Justice of the Slovak Republic	ELRC_331_English-Slovak parallel corpus of texts from	Multilingual Corpus	English | Slovak	openUnder_PSI
332	Corpus RIZIV	ELRC_332_Corpus RIZIV	Multilingual Corpus	Dutch; Flemish | French	openUnder_PSI
333	CATEX (German-Italian parallel corpus of legal and administrative texts)	ELRC_333_CATEX (German-Italian parallel corpus of legal	Multilingual Corpus	German | Italian	non-standard/Other_Licence/Terms

ID#	Resource Name	Directory Name	Type	Language(s)	Legal Status
334	Legal terminology	ELRC_334_Legal terminology	Terminology	German | Italian	non-standard/Other_Licence/Terms
335	Spanish-Portuguese website parallel corpus	ELRC_335_Spanish-Portuguese website parallel corpus	Multilingual Corpus	Portuguese | Spanish; Castilian	openUnder_PSI
336	Spanish-Italian website parallel corpus	ELRC_336_Spanish-Italian website parallel corpus	Multilingual Corpus	Italian | Spanish; Castilian	openUnder_PSI
337	Maltese-English website parallel corpus	ELRC_337_Maltese-English website parallel corpus	Multilingual Corpus	English | Maltese	openUnder_PSI
338	Spanish-French website parallel corpus	ELRC_338_Spanish-French website parallel corpus	Multilingual Corpus	French | Spanish; Castilian	openUnder_PSI
339	Spanish-English website parallel corpus	ELRC_339_Spanish-English website parallel corpus	Multilingual Corpus	English | Spanish; Castilian	openUnder_PSI
340	Glossary - Legal terminology on children protection	ELRC_340_Glossary - Legal terminology on children	Terminology	English | French | Italian	non-standard/Other_Licence/Terms
341	TGlossary	ELRC_341_TGlossary	Terminology	German | Italian	non-standard/Other_Licence/Terms
342	Spanish-German website parallel corpus	ELRC_342_Spanish-German website parallel corpus	Multilingual Corpus	German | Spanish; Castilian	openUnder_PSI
345	The Udáras na Gaeltachta Corpus of bilingual PDFs and Word documents	ELRC_345_The Ud ras na Gaeltachta Corpus of	Multilingual Corpus	English | Irish	openUnder_PSI
347	Translations of Hungarian from public websites	ELRC_347_Translations of Hungarian from public websites	Multilingual Corpus	Czech | Dutch; Flemish | English | Finnish | French | German | Hungarian | Polish | Slovenian | Swedish	openUnder_PSI
348	Electronic Exchange of Social Security Information documents in Czech-English	ELRC_348_Electronic Exchange of Social Security Information	Multilingual Corpus	Czech | English	openUnder_PSI
349	Gabra lexicon	ELRC_349_Gabra lexicon	Terminology	English | Maltese	CC-BY_4.0
350	Malta Government Gazette	ELRC_350_Malta Government Gazette	Multilingual Corpus	English | Maltese	openUnder_PSI
351	Laws of Malta	ELRC_351_Laws of Malta	Multilingual Corpus	English | Maltese	openUnder_PSI
352	Bilingual extracts from Malta International Airport Newsletter	ELRC_352_Bilingual extracts from Malta International Airport	Multilingual Corpus	English | Maltese	non-standard/Other_Licence/Terms

ID#	Resource Name	Directory Name	Type	Language(s)	Legal Status
353	Monolingual documents from the Government of Lithuania	ELRC_353_Monolingual documents from the Government of	Monolingual Corpus	Lithuanian	CC-BY_4.0
354	TERMIS: Slovene-English terminology in the field of public relations	ELRC_354_TERMIS Slovene-English terminology in the field	Terminology	English | Slovenian	CC-BY-SA_4.0
355	Documents for Translation Tendering Batch 2	ELRC_355_Documents for Translation Tendering Batch 2	Monolingual Corpus	Dutch; Flemish	openUnder_PSI
356	The Vocabulary of Safety and Health at Work (TSK 35)	ELRC_356_The Vocabulary of Safety and Health	Terminology	English | Finnish | French | German | Swedish	CC-BY-NC-ND_4.0
357	The Terminological Vocabulary of Kela – Benefit-related Concepts, 4th edition (TSK 49)	ELRC_357_The Terminological Vocabulary of Kela -	Terminology	Finnish | Swedish	CC-BY-NC-ND_4.0
358	Croatian monolingual corpus of the Official journal of the Republic of Croatia	ELRC_358_Croatian monolingual corpus of the Official	Monolingual Corpus	Croatian	openUnder_PSI
359	DA-EN Danish Ministry of Finance	ELRC_359_DA-EN Danish Ministry of Finance	Multilingual Corpus	Danish | English	non-standard/Other_Licence/Terms
360	DA-EN Danish Ministry of Economic Affairs and the Interior	ELRC_360_DA-EN Danish Ministry of Economic Affairs	Multilingual Corpus	Danish | English	non-standard/Other_Licence/Terms
361	DA-EN Danish Ministry of Finance 2	ELRC_361_DA-EN Danish Ministry of Finance 2	Multilingual Corpus	Danish | English	non-standard/Other_Licence/Terms
362	DA-EN Danish Health Authority	ELRC_362_DA-EN Danish Health Authority	Multilingual Corpus	Danish | English	non-standard/Other_Licence/Terms
363	DA-EN Danish Ministry of Higher Education and Science	ELRC_363_DA-EN Danish Ministry of Higher Education	Multilingual Corpus	Danish | English	CC-BY-NC_4.0
364	DA-EN Danish Ministry of Higher Education and Science 2	ELRC_364_DA-EN Danish Ministry of Higher Education	Multilingual Corpus	Danish | English	CC-BY-NC_4.0
365	DA-EN Danish Ministry of Higher Education and Science 3	ELRC_365_DA-EN Danish Ministry of Higher Education	Multilingual Corpus	Danish | English	CC-BY-NC_4.0
366	DA-EN Danish Ministry of Higher Education and Science 4	ELRC_366_DA-EN Danish Ministry of Higher Education	Multilingual Corpus	Danish | English	CC-BY-NC_4.0
367	English-Icelandic parallel corpus from Statistics Iceland	ELRC_367_English-Icelandic parallel corpus from Statistics Iceland	Multilingual Corpus	English | Icelandic	CC-BY_4.0

ID#	Resource Name	Directory Name	Type	Language(s)	Legal Status
370	Croatian-English terminology collection (medical sciences)	ELRC_370_Croatian-English terminology collection (medical sciences)	Terminology	Croatian | English	CC-BY_4.0
371	Croatian-English terminology collection (social sciences)	ELRC_371_Croatian-English terminology collection (social sciences)	Terminology	Croatian | English	CC-BY_4.0
372	Croatian-English terminology collection (natural sciences)	ELRC_372_Croatian-English terminology collection (natural sciences)	Terminology	Croatian | English	CC-BY_4.0
373	Croatian-English terminology collection (technical sciences)	ELRC_373_Croatian-English terminology collection (technical sciences)	Terminology	Croatian | English	CC-BY_4.0
374	English-Danish EASTIN-CL Multilingual Ontology of Assistive Technology	ELRC_374_English-Danish EASTIN-CL Multilingual Ontology of Assistive	Terminology	Danish | English	CC-BY-SA_4.0
375	English-Estonian EASTIN-CL Multilingual Ontology of Assistive Technology	ELRC_375_English-Estonian EASTIN-CL Multilingual Ontology of Assistive	Terminology	English | Estonian	CC-BY-SA_4.0
376	English-Latvian EASTIN-CL Multilingual Ontology of Assistive Technology	ELRC_376_English-Latvian EASTIN-CL Multilingual Ontology of Assistive	Terminology	English | Latvian	CC-BY-SA_4.0
377	English-Lithuanian EASTIN-CL Multilingual Ontology of Assistive Technology	ELRC_377_English-Lithuanian EASTIN-CL Multilingual Ontology of Assistive	Terminology	English | Lithuanian	CC-BY-SA_4.0
378	The Icelandic Terminology bank	ELRC_378_The Icelandic Terminology bank	Terminology	English | Icelandic	CC-BY-SA_3.0
379	Parallel corpus (Bulgarian - English) in the public administration domain	ELRC_379_Parallel corpus (Bulgarian - English) in	Multilingual Corpus	Bulgarian | English	openUnder_PSI
380	Slovak-English collection of demographic terms	ELRC_380_Slovak-English collection of demographic terms	Terminology	English | Slovak	underReview
381	Dutch Parallel Corpus	ELRC_381_Dutch Parallel Corpus	Multilingual Corpus	Dutch; Flemish | English | French	non-standard/Other_Licence/Terms

ID#	Resource Name	Directory Name	Type	Language(s)	Legal Status
382	Multilingual subtitle data 2BDutch	ELRC_382_Multilingual subtitle data 2BDutch	Multilingual Corpus	Dutch; Flemish | English | German	non-standard/Other_Licence/Terms
383	SoNaR Corpus	ELRC_383_SoNaR Corpus	Monolingual Corpus	Dutch; Flemish	non-standard/Other_Licence/Terms
384	Corpus of Icelandic texts from the Central Bank of Iceland	ELRC_384_Corpus of Icelandic texts from the	Monolingual Corpus	Icelandic	CC-BY_4.0
385	DAESO Corpus	ELRC_385_DAESO Corpus	Monolingual Corpus	Dutch; Flemish	non-standard/Other_Licence/Terms

ANNEX 4: LANGUAGE RESOURCES BY TYPE AND BY DOMAIN

The table illustrates the number of language resources by language and type:

Language	Bi/Multilingual Corpus	Monolingual Corpus	Lexical conceptual resource
Bulgarian	7	5	3
Croatian	2	1	5
Czech	3	0	3
Danish	2	0	2
Dutch; Flemish	7	2	1
English	110	1	44
Estonian	7	0	1
Finnish	11	0	4
French	34	1	11
German	22	0	15
Hungarian	2	0	1
Icelandic	1	0	1
Irish	6	0	1
Italian	13	1	8
Latvian	7	2	1
Lithuanian	3	1	1
Maltese	4	0	1
Modern Greek (1453-)	20	8	8
Norwegian	2	0	1
Polish	22	4	2
Portuguese	5	1	1
Romanian; Moldavian; Moldovan	11	4	1
Slovak	2	2	1
Slovenian	3	1	2
Spanish; Castilian	15	0	4
Swedish	9	0	9

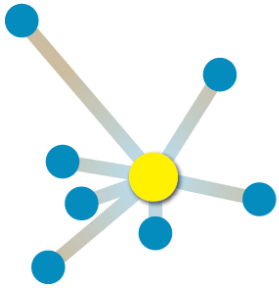
The table below provides an overview of the ELRC language resources by country and domain (EUROVOC).

EUROVOC domain	All countries	Austria	Belgium	Bulgaria	Croatia	Cyprus	Czech Republic	Denmark	Estonia	Finland	France
AGRICULTURE, FORESTRY AND FISHERIES	6	0	0	0	0	0	0	0	0	0	0
BUSINESS AND COMPETITION	2	0	0	1	0	0	0	0	0	0	0
ECONOMICS	11	0	0	1	0	0	0	0	0	2	0
EDUCATION AND COMMUNICA-TIONS	10	0	0	1	0	0	0	0	0	0	0
EMPLOYMENT AND WORKING CONDITIONS	8	0	0	0	0	0	0	0	0	2	0
ENERGY	1	0	0	0	0	0	0	0	0	0	0
EUROPEAN UNION	3	0	0	2	0	0	0	0	0	0	0
FINANCE	9	0	0	0	0	0	1	0	0	2	0
INDUSTRY	5	2	0	0	1	0	0	0	0	0	0
INTERNATIONAL ORGANISATIONS	1	0	0	0	0	0	0	0	0	0	0
INTERNATIONAL RELATIONS	12	0	0	2	0	1	0	0	0	0	0
LAW	38	2	0	5	1	2	1	0	1	0	0
POLITICS	30	1	0	0	2	1	0	0	0	2	0
PRODUCTION, TECHNOLOGY AND RESEARCH	5	0	0	1	0	0	0	0	0	0	1
SCIENCE	6	0	0	1	2	0	1	0	0	0	0
SOCIAL QUESTIONS	30	0	0	0	1	1	0	0	0	3	0
TRADE	3	0	0	0	0	0	0	0	0	0	0
TRANSPORT	6	0	0	1	1	0	0	0	0	0	0
N/A	55	1	6	0	0	0	1	0	0	0	5

EUROVOC domain	Germany	Greece	Hungary	Iceland	Ireland	Italy	Latvia	Lithuania	Luxembourg	Malta	Netherlands
AGRICULTURE, FORESTRY AND FISHERIES	0	1	0	0	0	0	0	0	0	0	0
BUSINESS AND COMPETITION	0	0	0	0	0	0	0	0	0	0	0
ECONOMICS	0	4	0	0	0	0	1	1	0	0	0
EDUCATION AND COMMUNICA-TIONS	0	2	0	0	0	0	0	0	0	0	0
EMPLOYMENT AND WORKING CONDITIONS	0	1	0	0	0	0	0	0	0	0	0
ENERGY	0	0	0	0	0	0	0	0	0	0	0
EUROPEAN UNION	0	0	0	0	0	0	0	0	0	0	0
FINANCE	0	2	0	0	0	0	3	0	0	0	0
INDUSTRY	0	1	0	0	0	0	0	1	0	0	0
INTERNATIONAL ORGANISATIONS	0	0	0	0	0	0	0	0	0	0	0
INTERNATIONAL RELATIONS	0	2	0	0	0	0	3	0	0	0	0
LAW	0	10	0	0	0	0	0	0	0	0	0
POLITICS	2	8	0	0	0	0	4	1	4	0	0
PRODUCTION, TECHNOLOGY AND RESEARCH	0	0	0	0	0	0	0	0	0	0	0
SCIENCE	0	2	0	0	0	0	0	0	0	0	0
SOCIAL QUESTIONS	0	10	0	1	0	0	1	1	0	0	0
TRADE	0	0	0	0	0	0	0	0	0	0	0
TRANSPORT	2	0	0	0	0	0	0	0	0	0	0
N/A	1	1	1	0	7	13	0	0	0	4	6

EUROVOC domain	Norway	Poland	Portugal	Romania	Slovakia	Slovenia	Spain	Sweden	U.K.
AGRICULTURE, FORESTRY AND FISHERIES	0	4	0	0	0	0	0	1	0
BUSINESS AND COMPETITION	0	0	0	0	0	0	0	1	0
ECONOMICS	0	0	0	0	0	0	0	0	0
EDUCATION AND COMMUNICA-TIONS	1	0	0	2	0	1	0	0	0
EMPLOYMENT AND WORKING CONDITIONS	0	0	0	0	0	0	0	5	0
ENERGY	0	1	0	0	0	0	0	0	0
EUROPEAN UNION	1	0	0	0	0	0	0	0	0
FINANCE	0	0	0	0	0	0	0	0	0
INDUSTRY	0	0	0	0	0	0	0	0	0
INTERNATIONAL ORGANISATIONS	0	1	0	0	0	0	0	0	0
INTERNATIONAL RELATIONS	0	4	0	0	0	0	0	0	0
LAW	0	1	4	3	2	3	0	2	0
POLITICS	0	0	0	0	0	0	0	0	0
PRODUCTION, TECHNOLOGY AND RESEARCH	0	0	0	0	0	0	0	0	0
SCIENCE	0	0	0	0	0	0	0	0	0
SOCIAL QUESTIONS	1	2	0	2	2	0	1	3	0
TRADE	0	3	0	0	0	0	0	0	0
TRANSPORT	0	2	0	0	0	0	0	0	0
N/A	0	0	2	1	0	0	0	0	0

ANNEX 5: ELRC ADVISORY REPORT



European Language Resource Coordination

Connecting Europe Facility

ELRC Consultancy Request 1

Author(s):

Kim Harris (DFKI)
Arle Lommel (DFKI)
Christian Dugast (DFKI)
Aivars Bērziņš (TILDE)
Josef van Genabith (DFKI)

Table of Contents

1. DESCRIPTION OF THE TASK

The background: Focus in CEF activities planned so far has been on collecting, processing and converting **existing** language resources for statistical machine translation (SMT). However, for some languages and in some topical areas, there will be no (or very little) raw material (in-domain parallel or quasi-parallel texts) available.

Question: what would be the **cost** of creating a parallel corpus of "sufficient" size, for a given (arbitrary) topical domain and for each of the EU languages, assuming that monolingual texts (source or target language) are available in sufficient volume?
The approach should be analytic and present the assumptions and scenarios/alternatives. In particular, the following parameters (list is not exhaustive) should be estimated, as they affect the cost substantially:

1. The "acceptable" level of **quality** for fully automatic MT scenario (no post-editing) for a typical e-Government scenario in a limited domain for a) exchanging information between administrations, b) exchanging information between administration/economic operator and **client** (citizen or company). We can exclude use scenarios where "perfect" quality is required (e.g. life-critical applications in eHealth or civil protection). So, we can assume that a reasonable amount of errors is tolerated in both scenarios a) and b) above.
2. The **volume** (number of segments/sentences) required for a parallel corpus to train an SMT system, using best available technology, and assuming that a generic (i.e. not domain-adapted) baseline SMT system and the corresponding parallel corpora exist.
3. The **cost** and effort of a creating parallel corpus of size described in point 2 above that would result in the respective SMT delivering quality levels of 1a and 1b above, using best possible available technology and necessary human effort. You can present different scenarios, e.g. assuming manual translation of corpus (using CAT tools), or automated/post-edited translation of corpus, or any automated process deemed feasible (e.g. aligning segments or snippets in non-parallel bilingual corpora employing any available method).
4. If your approach/assumption requires the use of bilingual in-domain **terminologies** to complement parallel texts, you can assume that such in-domain terminology resources either exist or can be automatically extracted from available documents.

Notes: while it is known that different language pairs (e.g. ES-PT compared to FI-GR) will require very different sizes of training corpora, you don’t need to assess all language pairs separately, but to estimate the size/cost of corpus for the "easiest" language pair and for the "most difficult" and trying to estimate the distribution and typical average corpus sizes for representative language pairs (e.g. assuming EN as pivot language and averaging over the required corpus size/cost of EN-X, where X goes through all 24 official EU languages). It is important if you can, however, identify the "outliers", especially the most difficult/expensive language pairs, because the best solution for those might be to apply rule-based or other alternative MT technologies.

Deliverable: the estimate should be presented in a table format where the different categories of language pairs are presented separately, and the different methodologies (point 3 above) are presented separately. The values in the cells of the table should be the estimated cost in EUR.
A description of methodology, assumptions and sources/references should be presented.

2. SHORT SUMMARY OF THE TASK

Given a general domain MT system MT(GD) trained on general domain data GD, given a domain D which is substantially different from GD (so that MT(DG) output for source side data from D is not of sufficient quality), and a targeted MT output quality level Q on data from domain D, how much in-domain data ID of type D is required to tune MT(GD • ID) to achieve quality level Q on data from domain D.

Here GD • ID is a composition of training data GD and ID, in the simplest case concatenation, in sophisticated cases state-of-the-art domain adaptation methodologies. Sophisticated domain adaptation methodologies include (i) the computation of possibly multiple domain specific MT models over GD and model combination with an MT model computed on ID or (ii) supplementing ID with "suitable" parts of GD to produce a domain adapted/tuned version MT(GD • ID) of MT(GD) adapted to D.

3. ASSESSMENT OF THE TASK

To the best of our knowledge there is currently no general solution to the problem as stated in **2 Short Summary of the Task** above, and reformulated in terms of given and to predict below:

Given:

- a general description of the differences/distance between GD and D (perhaps in terms of an information theoretic measure such as perplexity, entropy, cross-entropy, Kullback-Leibler divergence etc.),
- a description of the size of the data GD and D
- an MT technology (e.g. PB-SMT)
- a description of the specificity of the GD and D domains and the MT models computed on GD and D (are these narrow or wide/diverse domains), e.g. in terms of translation table entropy
- a targeted average quality level Q (e.g. BLUE score) of the tuned MT output on D,
- a domain adaptation strategy "•" (e.g. difference in cross-entropy),

Predict:

- size(D), the size of in domain data ID of type D required to tune MT(GD • ID) to achieve quality level Q on data from domain D.

There is currently no analytic formula that from the given information computes the desired prediction.

A general solution to the task is an open research problem at PhD level.

4. WHERE TO GO FROM HERE?

Given this, we propose to split the task into two parts which are manageable:

- Per word translation cost estimates for the parameters of the task under consideration
- Best practice estimations, which allow a best practice guesstimate of total expected cost for MT Corpus Data creation

5. PER WORD TRANSLATION COST ESTIMATES

The following are the parameters that we consider essential to the per word cost estimates:

- **Market Costs:** to reflect current market costs, per word cost estimates have been collected from two reputable commercial Language Service Providers (LSPs, anonymized in the Tables below) that operate across Europe. We present three quotes (low, average, high) to show some (and in some cases lack of) relevant variation.
- **Cost Savings through Commissioning Translation for each Language Pair through LSP local to (one of the Members of) the Language Pair:** cost estimates reported are based on quotes from LSPs operating Europe wide. Going instead through a local LSP vendor for each language pair is likely to result in cost savings. These are not reported here due to the considerable administrative overhead in dealing with 24+ local LSPs.
- **Crowd Sourcing:** we do not think that crowd-sourcing with lay web-workers are an option in the current scenario due to the difficulties of reliably predicting translation quality, volume and delivery dates. Furthermore, given that the data in focus is specific domain data we think it unlikely to attract suitable domain and language experts. Therefore we do not provide crow-sourcing based estimates, but professional LSP quotes.
- **Specification of the Targeted Translation Outcome:** professional translation is done to specification. Depending on the requirements, translations can be required to be fully idiomatic (that is indistinguishable from text in the domain at stake originally authored in the target language), or adequate and grammatically correct (but not necessarily fully idiomatic). Other specifications are possible. Here we assume that the result of the translation is adequate and grammatically correct (but not necessarily fully idiomatic). This can be advantageous for MT as a translation that is closer to the source can be easier to derive MT models from than translations that involve extensive “cultural transfer” (insertion of additional information such e.g. introducing the phrase “The German Chancellor” before “Angela Merkel” in an English Translation where the German source only has the proper name) or highly idiomatic target language. Incidentally, costing these translation options does not produce changes in the quotes.
- **Terminology:** as specified in the Consultancy Request, terminological resources are not costed. They are assumed to exist and to be made available to the LSPs to provide terminology support for the creation (translation) of the data.

- **Bi-text Data:** we assume that only bi-text data need to be created. Comparable mono-lingual data sets for training domain specific language models are likely to exist.
- **Translation into English and Translation out of English:** to produce domain specific bi-text data for all 24 official EU languages requires 24x23=552 datasets for a single domain, assuming that the datasets can be used to train MT systems in both language directions. Further assuming 10 different domains of interest, the number of data sets multiplies to 5520. These are clearly too many data sets. Because of this, and following the specification of the Task (see above) for all official EU languages we will consider translation into English, and translation out of English. This produces 23x2=64 language pairs to be costed.
- **Difficulty (Domain Specificity) of Text/Domain/Genre:** within a given language pair, translation difficulty differs with respect to domain and genre. A highly technical text is usually more difficult to translate than a text in general language and this is reflected in translation cost: translators that are domain experts are often in less supply than translators that can translate general language texts. In our cost estimates we therefore provide three domain costs: general domain, technical domain and one for highly technical domains. Together with the 23x2 into English and out of English translation language pairs, this raises the per word cost estimation data points to 23x2x3=138.
- **Translation Memories:** per word cost estimates do not assume the existence of suitable translation memory (TMs) resources: if suitable MTs of effective size existed for the domain under consideration, then the required MT training data exists and no manual creation of training data is required. That said, pricing assumes that during the creation of the required data TMs are incrementally constructed (this is common practice in professional LSP workflows) with some (limited) impact on translation recycling and the resulting TMs are delivered as end product. We specifically say “limited impact” as the desired in-domain data created needs to show some variation to cover the domain of interest while keeping size (and cost) of the data within bounds. It is common for LSPs to distinguish three levels of translation tool support: standard technology (translators use common technology that increases throughput and simplifies interoperability), non-standard (proprietary of customized environment) and, finally, no support technology. This is costed in the quotes. Together with the 23x2 into English and out of English translation language pairs, and the 3 domain specificity rates, this raises the per word cost estimation data points to a total of 23x2x3x3=414.
- **Machine Translation and Post-Editing:** for the same reasons as for TMs above, per word cost estimates do not assume the existence of suitable machine translation (MT) resources: if suitable and effective MT systems existed for the domain under consideration, then the required MT training data exists and no manual creation of training data is required. Furthermore, general domain MT is not considered effective to support automatic quality translation to support translation cost-savings due to MT post-editing in the per word cost estimates: MT and post-editing would only start to be effective at some stage during the data creation so that given the limited size of the total required in-domain data the cost overhead of training MT

and integrating MT into a post-editing workflow, say half-way through, are likely to offset savings. Sophisticated incremental learning and updates of the MT from human post-edits is currently not a main-stream offering at professional LSPs.

- **Format:** insertion, removal, change or preservation of layout and format information impacts cost. We distinguish three levels: plain text or simple formatting, PDF or complex formatting and OCR-ed data. Together with the 23x2 into English and out of English translation language pairs, the 3 domain specificity rates and the three technology rates, this raises the per word cost estimation data points to a total of 23x2x3x3x3=1242.
- **Outside Domain Expert:** highly domain specific material may need to be proof-read by a domain expert. This is captured of an additional dimension for domain expert yes or no. Together with the 23x2 into English and out of English translation language pairs, the 3 domain specificity rates, the three technology rates and three format levels, this raises the per word cost estimation data points to a total of 23x2x3x3x3x2=2484.
- **Interactive per-Word Cost Estimation Spreadsheet:** 2484 cost estimate data points are too many to be comfortable presented in a single spreadsheet table. Therefore we present an interactive spreadsheet that allows the user to set language direction, domain specificity, technology support, format levels and domain expert dimensions. The spreadsheet also allows the user to specify data sizes required in terms of number of segments and average number of words per segment.
- **Difficult and Easy Language Pairs:** translation cost is not the same for each language pair. This is due to a number of reasons: some languages are considered more difficult, language pairs which are distant tend to be harder to translate and for some language pairs sufficient numbers of professional translators are not easy to find. On the other hand, some languages are linguistically close (e.g. Spanish-Portuguese) and therefore easier to translate. While this report does not provide per word translation estimates for all EU language pairs (see above), we do consider extreme points of the spectrum of language pairs with respect to translation difficulty and cost. In addition to X->English and English->X the interactive spreadsheet provides a separate section for “extreme” cases such as Finish->Greek or Spanish->Portuguese.
- **Statistical Machine Translation:** throughout this report we assume standard Phrase-Based Statistical Machine Translation (PB-SMT, as e.g. implemented in the Moses System) models as the underlying base technology. Different models (e.g. Hierarchical Phrase Based SMT, Rule-Based MT, Neural MT) would potentially further increase the parameter space (some models may be inherently better suited to “difficult” translation pair languages than others).

6. Domain Adaptation Best Practice

There is large and varied research literature on domain adaptation, both with respect to language and translation models.

Domain adaptation through the language model only can be successful if the training data for the general domain MT system already to a considerable extent includes data relevant to

the specific domain of interest amongst the other domains in the general domain training data. A domain tuned language model can then act as a filter or a “lens” that biases translations to adhere to what is characteristic to the domain of interest.

In most cases (but not always) both language and translation model are adapted for domain tuning. Therefore in the text below we will not consider them separately.

Overall there four major strands identifiable in the research literature on domain tuning:

6.1. Domain Adaptation through Supplementary Lexical Resources

The idea here is that the characteristics of a particular domain are often reflected in a specialized vocabulary (and its translation – this often referred to as terminology and includes multi-word units MWUs) and that a general domain MT system will often not have seen this vocabulary in its general domain training data. If this is the case the general domain MT system will be confronted with many out-of-vocabulary (OOV) items when translating domain specific data, with adverse effects on the translation output. In order to plug these OOV holes, lexical resources (bi-lingual dictionaries, either hand –crafted or otherwise obtained) can be added to the MT training data. Sometimes the addition of lexica resources is weighted by a factor (often between 1-5, by simply adding the supplementary lexical data that many times to the training data). Addition of domain specific bi-lingual dictionaries can also improve alignment, and as a consequence phrase extraction. For the purposes of this Consultancy Action, availability of domain specific lexical resources is assumed as given. Therefore we will not discuss this further below, except for saying that domain adaptation using supplementary lexical resources, while important, is limited as domain characteristics beyond the terminological level are not addressed.

6.2. Model Selection

General domain training data are often a mixture of data from different domains. It is often possible to automatically extract (e.g. using clustering methods, topic modelling approaches or modelling through latent variables) different more specific domains from the general domain training data. Instead of general domain training data GD we may be able to identify a number of specific domains D_i to D_n. These domains may overlap. We can then train different MT systems MT(D_i) to MT(D_n). Perhaps one of these models is already close to the specific domain of interest for the domain tuning task at hand. We can then combine the different MT models using weights tuned on a development (tuning set) set that reflects the characteristics of the domain of interest. This approach can be combined with supplementary MT training data reflecting the characteristics of the domain at stake as an additional model in the model combination.

6.3. Supplementary Data Selection

This approach assumes the existence of a limited set of domain relevant data D and a (usually) much larger pool of general domain data GD. Simply concatenating D to DG is usually not useful as DG may swamp or dilute D. The idea is to use the domain specific D as

seed data and to incrementally supplement (or grow) G with only those parts of general domain data GD that are “useful” to the domain of interest D, while avoiding the other parts of GD. Useful here usually means similar in some sense. There are many ways of identifying and selecting useful data from GD to extend D, including difference in cross-entropy, perplexity or modelling usefulness in terms of latent variables.

6.4. Incremental Updates

In this approach a general domain MT(DG) system is incrementally tuned or refined to a specific domain D by running the general domain MT(DG) system on input from D and using the result of human post-edits on the output of MT(GD) on D to incrementally and continuously adapt MT(DG) to better suit D. Simple concatenation of each translation input with its corrected (i.e. post-edited) output to the training data GD and retraining the MT system with the original data and each such translated and corrected input-output pair is not effective: (i) a single additional bi-text segment added to the general domain training data will not have any perceivable effect on a large scale general domain MT system and (ii) retraining will simply take much too long. Better approaches are required. These include batch retraining (where dozens or hundreds of post-edits are collected for each retaining step), MT models which allow incremental updates without full etraining or where an additional second translation (and language) model is trained incrementally on just the post-edits and combined (usually weighted) with the original general domain MT model. The additional second translation model can be trained quickly on the fly as the number of post-edited segments is low in comparison with the general domain MT model which may be based on millions of translation segments. The additional second translation model can either take the form of a source-target translation model of the form of a “mono-lingual” MT(GD)-target translation model adapting the output of the MT(GD) system.

6.5. Determining the Size of In-domain Data Required to Tune a General Domain MT System to a Specified Quality Level on a Specific Domain

In abstract terms, predicting the minimal amount of in-domain data ID required for domain adaptation or domain tuning of a general domain MT system MT(GD) (whether using lexical resources, model combination, supplementary data selection or incremental approaches) to a specific domain D with quality level outcome Q in data of type D takes the following form:

Given:

- a description of the size of the data GD and D
- a general description of the differences/distance between GD and D (perhaps in terms of an information theoretic measure such as perplexity, entropy, cross-entropy, Kullback-Leibler divergence etc.)
- an MT technology (e.g. PB-SMT)
- a description of the specificity of the GD and D domains and the MT models computed on GD and D (are these narrow or wide/diverse domains), e.g. in terms of translation table entropy
- a targeted average quality level Q (e.g. BLUE score) of the tuned MT output on D

- a domain adaptation strategy "•" (e.g. difference in cross-entropy),

Predict:

- size(ID), the size of in domain data ID of type D required to tune MT(GD • ID) to achieve quality level Q on data from domain D.

There is currently no analytic formula that from the given information computes the desired prediction.

Currently, the amount of in-domain data ID required to adapt a general domain MT system MT(GD) to guarantee a fixed quality level on data of domain D can only be established experimentally.

A general solution to the task is an open research problem at PhD level.

We are aware of research in quality estimation that may guide the way towards progress in this area: for the more narrow scenario of estimating the quality of an MT system without running the system on data, there are approaches that try to predict MT quality in terms of information theoretic measures of fit between the training and development or test data.

6.6. Rules of Thumb/Guesstimates

As the general problem underlying the consultation task is currently not solved, below we give a few rules of thumb based on reported (often informally) best practice of MT developers.

Tuning on In-Domain Data: modern statistical machine translation systems combine many features with weights (often in log-linear models). If the general domain data available already has reasonable amounts of data suitable to the specific domain of interest, a cheap-and-cheerful (and sometimes successful) first way to adapt a general domain system is to use additional available in-domain data to tune the SMT system (i.e. to set the feature weights of the system to optimally fit the desired domain). This approach is a recommended first step. It will in general not be successful if the general domain data do not contain "hidden data" suitable to the domain of interest. In some cases, i.e. if there is sufficient in-domain data "hidden" in the general domain data, even tuning the target side language model to the domain of interest has the potential to effect substantial gains.

More Data is Better Data: this is true, provided, however, that that data is domain focused. If this is not the case, additional data may introduce translation alternatives that are simply not relevant to the domain under consideration (different translations for "bat" in sports or zoology). The degree of focus of a translation model can be partially captured by translation table entropy.

Morphologically Rich and Syntactically Varied Languages Require more Training Data: everything else being equal, morphologically rich and syntactically varied languages tend to require more training data than morphologically simpler and syntactically more constrained

languages. This is due to the fact that smaller amounts of training data simply do not expose the MT system to the full variation of possibilities in training.

Narrow Domains can be well Translated with Systems Based on Small Amounts of Training Data: there are reports on using MT for data relevant in the localization industry where 50,000 segments of training data are sufficient to achieve high quality translation output on narrow technical domains with limited vocabulary sizes, short average segment length and repetitive syntax (e.g. for the translation of user interface dialogue boxes in the IT industry). This holds the possibility that for similar domains substantially less than 50,000 segments of in-domain seed data are sufficient to tune general domain systems that do already contain (!) useful "hidden" data using e.g. model combination or supplementary data selection approaches.

The Minimum Size of In-domain Data Required Depends on Domain Target Level Quality and the Amount of "Hidden" Useful Data in the General Domain Training Data: obviously the higher the targeted translation quality level of the domain tuned MT system, the larger the size of the in-domain data required. A second important variable is the amount of "hidden" useful data in the general domain training data: extreme points are (i) when there is no "hidden" useful data in the general domain training data (in this case the additional dedicated in-domain data has to do all the work) and (ii) when there is ample "hidden" useful data in the general domain training data (in which case very little additional in-domain data is required for supplementary data selection, model combination approaches and even feature weight tuning or target side language model adaptation may already show considerable gains).

ANNEX 6: PROJECT PROGRESS INDICATORS

Task	Indi-cator Nr	Indicator Name	Source	Base-line	Target M24	Actual M24	Comments (especially on methodology or targets if month 12/18/24 targets not applicable)
1 Secretariat	1.1	Number of data providers' calls	Secretariat log files	n.a.	no target	4	
	1.2	Percentage of deliverables and reports submitted on time	Deliverables/reports	n.a.	80%	100%	
	1.3	Number of email enquiries answered per month	Ticket System	n.a.	no target	432	On average per month
	1.4	Number of phone calls received per month	Ticket system	n.a.	no target	4	On average per month
	1.5	Timeliness of bi-weekly phone conferences	Agreed dates, 2 per month	n.a.	80%	95%	% of phone conferences per half a year held on time
2 Technical help desk	2.1	Number of unique email enquiries answered per month (multiple emails in a conversation count as one)	email tracking system	n.a.	no target	15	Equals 1,25 on average per month
	2.2	Number of phone calls received per month	Helpdesk log files	n.a.	no target	5	Equals 2,4 on average per month
	2.3	Number of replies	Emails, Phones, Forum	n.a.	100%	100%	
	2.4	Percentage of simple questions solved within set deadlines (1 working day)	Helpdesk log files	n.a.	95%	100%	

Task	Indi-cator Nr	Indicator Name	Source	Base-line	Target M24	Actual M24	Comments (especially on methodology or targets if month 12/18/24 targets not applicable)
	2.5	Percentage of complex questions solved within set deadlines (5 working days)	Helpdesk log files	n.a.	95%	100%	
	2.6	Percentage of very complex questions forwarded to consultancy	Helpdesk log files	n.a.	5%	0%	
	2.7	Average response time for simple query	Helpdesk log files	n.a.	8h	<8h	
	2.8	Average response time for complex query	Helpdesk log files	n.a.	3d	<3d	
3 Language Resource Board	3.1	Number of National Contact Points identified per country		x	2x (100%)	90%	x = the number of NAPs indicated in the initial support letter
	3.2	Percentage of National Contact Points who provided data		n.a.	60%	60%	
	3.3	Percentage of LRB members attending face-to-face meetings		n.a.	70%	>75%	
4 Website	4.1	Bounce rate		n.a	70%	54,03%	On average
	4.2	Number of visits per month		n.a.	tbd	831,67	Per month on average
	4.5	Downtime figures		n.a.	<5%	<4%	At all times

Task	Indi-cator Nr	Indicator Name	Source	Base-line	Target M24	Actual M24	Comments (especially on methodology or targets if month 12/18/24 targets not applicable)
	4.6	Number of users of social media channels		n.a.	tbd	32	
5 Conferences	5.1	Number of participants	attendance lists	120	n.a	124	
	5.2	Number of participants from technology eco-system	attendance lists	<30	n.a	19	
	5.3	Number of participants from public sector administration	attendance lists	>30	n.a	100	
	5.4	Number of participants from LSPs	attendance lists	<30	n.a	5	This indicator was changed after 1st year to minimize no. of LSP participants
6 Workshops	6.1	Number of participants	attendance list	15-50	15-50	31-130	50 for big countries (e.g. France, Germany), 15-20 for small countries (e.g. Malta, Cyprus etc.)
	6.2	Number of participants from technology eco-system	attendance list	5	5	<5	On average
	6.3	Number of participants from public sector	attendance list	25	25	>25	On average
	6.4	Number of participants from LSPs	attendance list	5	5	<5	On average; should be minimized after trial workshops
	6.5	Satisfaction rate - Feedback from evaluation form	feedback form	n.a.	n.a	4 (Satisfied)	On average
	6.6	Number of workshops conducted by month x	actual workshops conducted	n.a.	n.a.	29	U.K. was omitted because of political involvements

Task	Indi-cator Nr	Indicator Name	Source	Base-line	Target M24	Actual M24	Comments (especially on methodology or targets if month 12/18/24 targets not applicable)
	6.7	Number of data sources identified	List of Data Sources	n.a.	See 7. Data manag ement	See 7. Data manag ement	
7 Data management	7.1	Leader board formula with examples	Leader board formula	tbd	n.a.	n.a.	Replaced by a simpler measurement (no. of contributed resources) - see report for results
	7.2	Number of data sourcs iden tified to secure all data needed			600+	1.083	
	7.3	Number of raw resources secured			250	n.a.	
	7.4	Number of resourcessecu red cleaned/packa ged			200	225	
	7.5	Number of resources rejected			n.a.	<15%	This should be less than 15%
8 Consultancy tasks	8.1	Number of questions received from helpdesk	Emails/pho ne calls	tbd	tbd	0	
	8.2	Number of questions received from EC	Emails	tbd	4	1	To be defined and estimated by EC
	8.3	Time needed to answer/respon se time		n.a.	80%	100%	Percentage of timely response
	8.4	Number of advisory reports submitted	Advisory Reports	n.a.	1	1	

European Commission

ELRC European Language Resource Coordination – Final Report
Luxembourg, Publications Office of the European Union

2017 - 92 pages