

Семинар на проекта *Координиране на езиковите ресурси в Европа*

Как можем да бъдем полезни?

Христина Добрева

- ✓ проф. Светла Коева, Институт за български език;
- ✓ Светлана Юркова, Генерална дирекция „Писмени преводи“, Европейска комисия;
- ✓ Габриела Митова, Министерство на транспорта, информационните технологии и съобщенията;
- ✓ Любомира Тошкова, Изпълнителна агенция „Морска администрация“;
- ✓ Иглика Караниколова, „Лозанова“ 48 ООД

Единният цифров пазар – част от стратегията на Европейската комисия за развитието на Европа до 2020 г. е многоезиков. Но езиковите бариери издигат сериозни, макар и невидими, граници за него. Езиковите технологии като автоматизирания превод могат да подпомогнат преодоляването на тези бариери между хората и народите, като дават възможност за развитието на многоезична Европа.

Според Европейската комисия:

- 90 % от всички европейски потребители предпочитат да използват информация в интернет на собствения си език;
- 82 % от 4000+ онлайн магазини са едноезични;
- 42 % от интернет потребителите никога не са купували на различен език от собствения си.

Следователно, осигуряването на достъп до информация на различни езици може да донесе значителни ползи, както за икономиката, така и за обществото. Изпълнението на Стратегията за единен цифров пазар и пълноценното му функциониране може да донесе увеличение на икономическия ръст с близо 340 милиарда евро, стотици хиляди нови работни места и жизнено общество, основано на знанието.

Защо са необходими езиковите ресурси ?



Езикови ресурси са необходими за подобряване качеството на машинния превод, както в общи, така и в специфични области. За подобряване услугите на платформата за автоматизиран превод CEF.AT, основните автоматизирани системи за превод трябва да бъдат „захранени“ със съответните материали на всички официални езици на 30-те държави, участващи в Механизма за свързване на Европа (МСЕ).

В зависимост от обхвата на публичните услуги, които ще бъдат подкрепяни от платформата CEF.AT се определят и тематичните области, за които се очаква събиране на езикови ресурси.

Примерни тематични области: права на потребителите, култура, правна база, социална сигурност, здравеопазване, обществени поръчки и т.н.

Как биха могли да бъдат използвани езиковите ресурси?



Ежедневно в държавите членки, асоциираните към Механизма за свързване на Европа страни, неправителствени и частни организации се създава огромно количество ценни езикови ресурси. Голяма част от тези данни би могла да се използва за нуждите на новата платформа CEF.AT.

Дейностите по проект „Координиране на езикови ресурси в Европа“, финансиран по МСЕ предвиждат събиране, както на отворени данни, които биха могли да бъдат предоставени за повторно използване, така и за търговско достъпни набори от данни.

Някои набори от данни, произлизащи от публичните администрации могат да бъдат използвани директно от платформата за автоматичен превод CEF.AT като терминологични и лексикални ресурси, речници и др. Много други ресурси създадени под формата на информационни документи (доклади, наръчници, листовки, данни за административни решения и т.н.), ще се нуждаят от допълнителна обработка, за да се превърнат в езикови ресурси, в случаите, когато са подадени във формат за многократна употреба.

Какви видове данни са полезни за машинния превод?



С цел да отговарят на изискванията за автоматизиран превод, съответните езикови ресурси трябва да бъдат от различни видове:

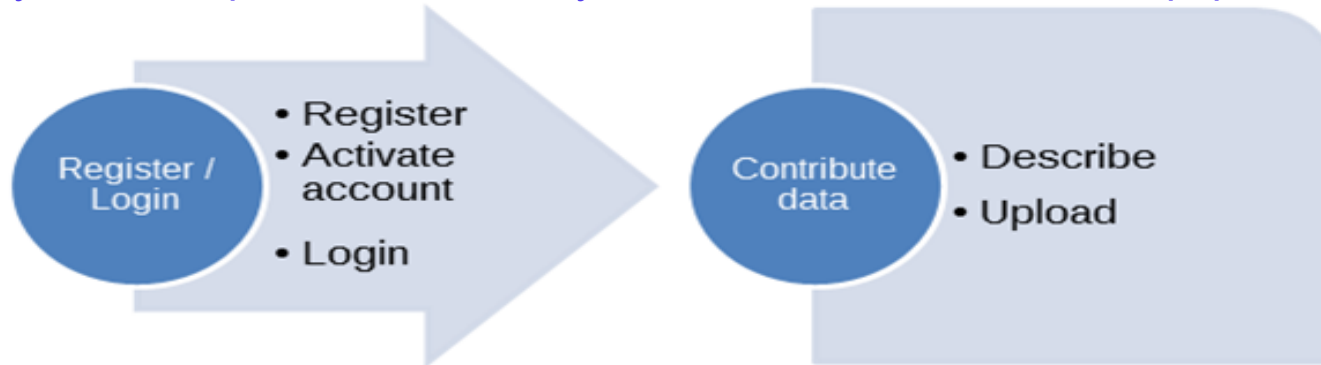
- ✓ **Преводачески памет:** лингвистичните бази данни, които обхващат преводи, направени от човек. Те могат да се използват за улесняване на бъдещи преводни задачи, но и за „обучение“ на автоматизирани системи за превод;
- ✓ **Превод/езикови модели:** статистическа информация, която определя вероятността за използване на дадено понятие или словосъчетание в процеса на превеждане;
- ✓ **Корпуси:** едноезични и многоезични масиви, съпоставими и съгласувани, паралелни документи;
- ✓ **Лексика:** едноезични и многоезични списъци на думи, многозначни думи, изречения и т.н. в общи или специфични области;
- ✓ **Терминологични ресурси:** структуриран набор от понятия, свързани с езикова информация в определена тематична област;
- ✓ **Граматика:** набори от правила за формализиране на езика.

Данните могат да бъдат предоставяни основно по два канала – чрез проекта за Координиране на езикови ресурси в Европа и чрез Европейската комисия.

В рамките на проекта е изградено „хранилище“ за документиране и съхранение на езикови ресурси, които ще бъдат полезни за платформата за автоматичен превод CEF.AT.

Консорциумът е предвидил две опции, от които следва да изберете:

- ✓ Да изпратите данните чрез електронна поща на адрес data@lr-coordination.eu;
- ✓ Да качите данните си директно на сайта на проекта чрез опростена уеб форма (<http://www.lr-coordination.eu/resources>). След подаване на данните представител на консорциума се свързва с Вас за получаване на допълнителна информация.





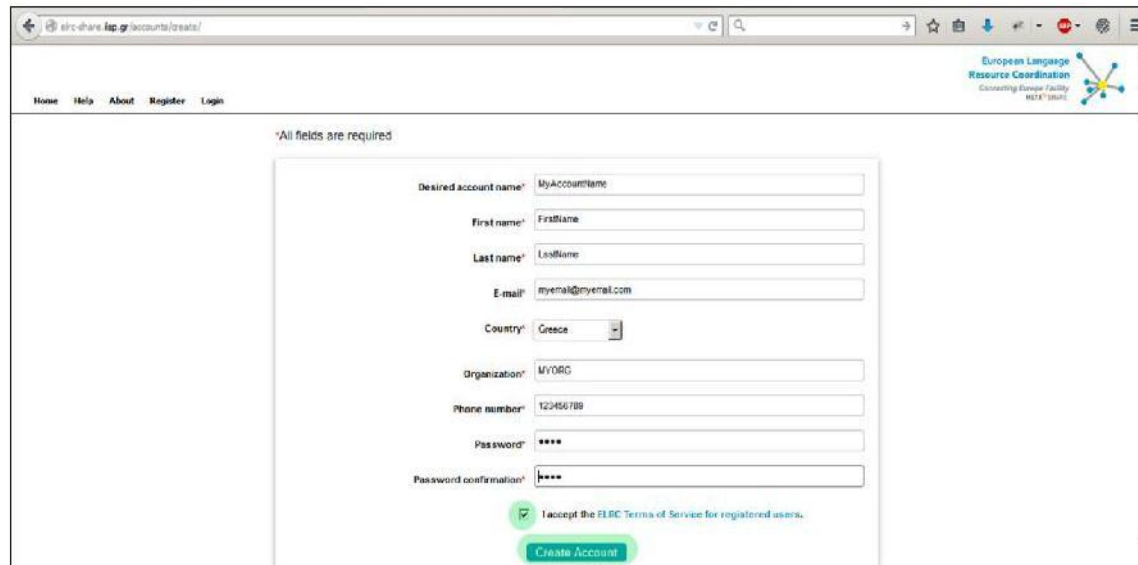
Можете да предоставяте данни само като регистриран потребител. Регистрацията е от основно значение за идентификацията на доставчика на езикови ресурси. Чрез нея се следи за злоупотреби, поддържа се сигурна и надеждна среда, създава се регистър на доставчиците и се управлява употребата на самите ресурси.

Можете да се регистрирате през сайта на проекта **ELRC-SHARE repository** или директно на elrc-share.ilsp.gr, чрез натискане на бутон „Регистрация”.



The screenshot shows the homepage of the ELRC-SHARE repository. At the top, there is a navigation menu with links for Home, Help, About, Register, and Login. The main header features the text "ELRC-SHARE Repository" over a background image of various national flags. Below the header, a welcome message reads: "Welcome to the ELRC-SHARE repository! The ELRC-SHARE repository is used for documenting, storing, browsing and accessing Language Resources that are collected through the European Language Resource Coordination and considered useful for feeding the CEF AT platform. If you want to contribute resources, all you have to do is register now and go on to describe and upload your data with a simple form." A central diagram illustrates the process flow: "Register / Login" (with sub-points: Register, Activate account, Login) leads to "Contribute" (with sub-points: Describe, Upload). At the bottom, a note states: "All data resources gathered by this initiative will be provided exclusively to the European Commission for use in the CEF Automated Translation platform."

За създаване на профил е необходимо да попълните всички оказани полета.
Следва получаване на потвърждение на Вашата електронна поща.



The screenshot shows a web browser window at the URL `ellrc-share.lap.gr/accounts/create/`. The page features a navigation menu with links for Home, Help, About, Register, and Login. A logo for European Language Resource Coordination is in the top right corner. The main content area contains a registration form with the following fields: Desired account name (MyAccountName), First name (FirstName), Last name (LastName), E-mail (myemail@myemail.com), Country (Greece), Organization (MYORG), Phone number (123456789), Password (****), and Password confirmation (****). A checkbox for accepting the ELLRC Terms of Service is checked. A green 'Create Account' button is at the bottom of the form.

В електронното съобщение има връзка за активация.

След натискане Вашият профил ще бъде активиран и ще имате права за поставяне на данни.



European Language
Resource Coordination
Connecting Europe Facility
META™ to LBR™

Home Contribute Resources Help About Your Profile, m Logout

Your account has been activated.
You can now proceed to contributing resources.

New Resource

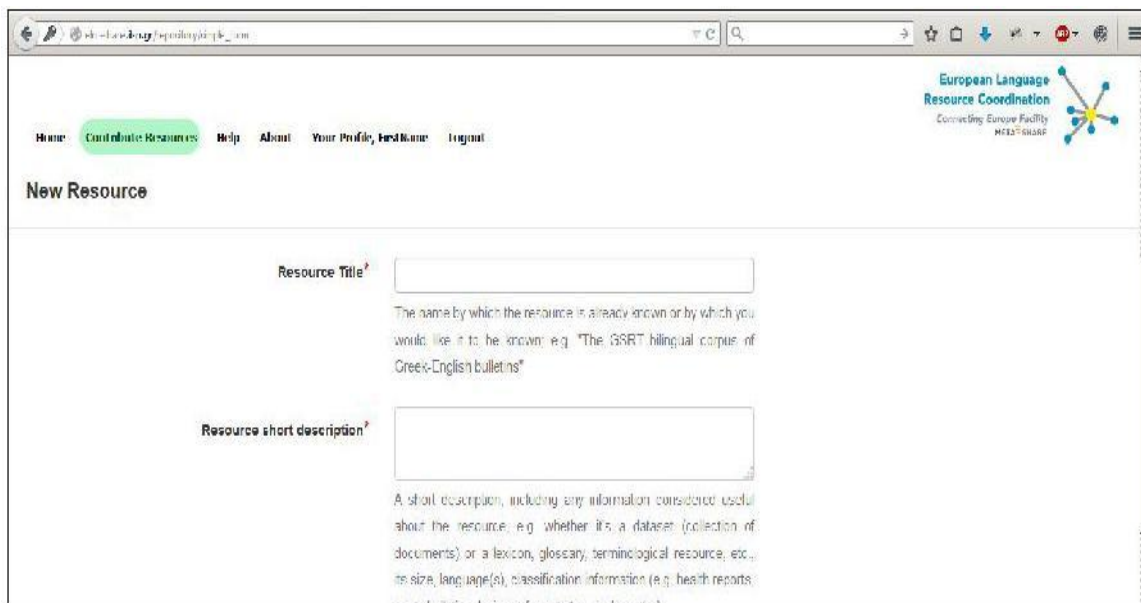
Resource Title*

The name by which the resource is already known or by which you would like it to be known, e.g. "The GSKL bilingual corpus of Greek-English bilinguals"

Resource short description*

A short description, including any information considered useful about the resource, e.g. whether it's a dataset (collection of documents) or a lexicon, glossary, terminological resource, etc., its size, language(s), classification information (e.g. health reports, cover bulletin, issues of assets seminar, etc.)

При подаването на данни се изисква попълване на кратка уеб форма. За достъпване на формата трябва да натиснете бутона „**Подаване на ресурси**” (Contribute Resources) от главното меню.

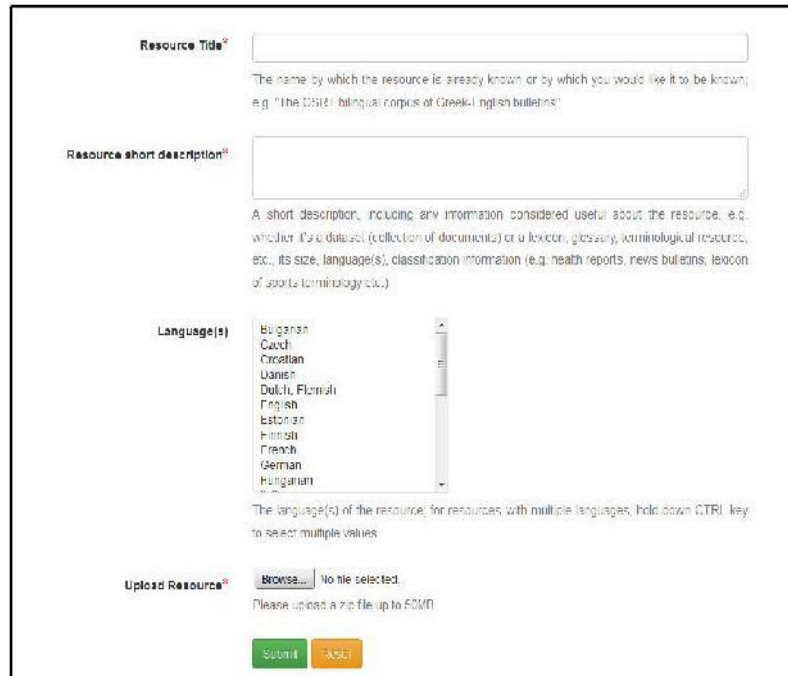


The screenshot shows a web browser window displaying the 'New Resource' form. The browser's address bar shows the URL 'http://www.elpa.europa.eu/portal/index.jsp'. The page header includes a navigation menu with 'Home', 'Contribute Resources', 'Help', 'About', 'Your Profile, first Name', and 'Logout'. The 'Contribute Resources' button is highlighted in green. The main content area is titled 'New Resource' and contains two input fields:

- Resource Title?**: A text input field with a red asterisk. Below it, a small text box explains: "The name by which the resource is already known or by which you would like it to be known: e.g. 'The GSRT bilingual corpus of Greek-English bulletins'".
- Resource short description?**: A larger text input field with a red asterisk. Below it, a small text box explains: "A short description, including any information considered useful about the resource, e.g. whether it's a dataset (collection of documents) or a lexicon, glossary, terminological resource, etc., its size, language(s), classification information (e.g. health reports, ...)".



Формулярът е много опростен и съдържа полета за попълване на информация, описваща ресурса (а именно: име на ресурса, езици, кратко описание на ресурсите) и бутон за качването на самия ресурс. Полета със звездичка (*) са задължителни.



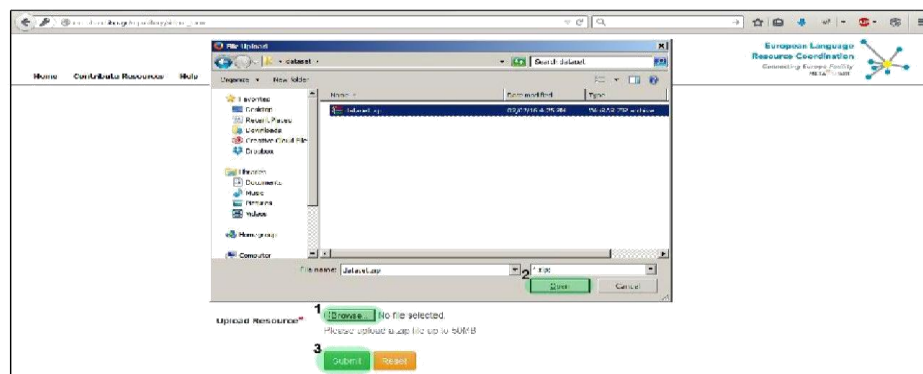
The screenshot shows a web form for submitting a resource. It contains the following fields and elements:

- Resource Title***: A text input field with a placeholder. Below it, a note reads: "The name by which the resource is already known or by which you would like it to be known; e.g. 'The OLSI1 bilingual corpus of Greek-English bulletins'".
- Resource short description***: A larger text area. Below it, a note reads: "A short description, including any information considered useful about the resource, e.g. whether it's a dataset (collection of documents) or a lexicon, glossary, terminological resource, etc., its size, language(s), classification information (e.g. health reports, news bulletins, lexicon of sports terminology etc.)".
- Language(s)**: A dropdown menu showing a list of languages including Bulgarian, Czech, Croatian, Danish, Dutch, Finnish, English, Estonian, Finnish, French, German, and Hungarian. Below the menu, a note reads: "The language(s) of the resource; for resources with multiple languages, hold down CTRL key to select multiple values".
- Upload Resource***: A "Browse..." button next to the text "No file selected." Below this, a note reads: "Please upload a zip file up to 50MB".
- At the bottom, there are two buttons: a green "Submit" button and an orange "Cancel" button.

Попълнете съответната информация, щракнете върху „Разглеждане“, за да качите Вашия ресурс. В прозореца, който се отваря, търсите в папките от Вашия компютър съответния компресиран файл (.zip), съдържащ данните, които желаете да споделите, маркирате и натискате бутона за подаване „Подай“.

Моля, имайте предвид, че разрешени са само компресирани файлове (zipped files) до **50 MB**. Ако се опитате да качите файл надхвърлящ този размер, се появява предупредително съобщение.

Connecting Europe Facility





Предоставените езикови ресурси от различни доставчици първо трябва да преминат през процес на преглед, при който оторизирани редактори могат да обогатят описанието на ресурсите и да потвърдят условията за издаване на лицензи. При неясни случаи могат да се свържат с представител на правната служба. По време на тази процедура данните се считат за „вътрешни“ и не могат да бъдат видяни отвън. Когато документацията е финализирана, упълномощеният редактор прави записа на данни публичен. Така ресурсите стават достъпни за изтегляне от потребителите, в съответствие с условията им за лицензиране.