European Language Resource Coordination
Connecting Europe Facility

Семинар на проекта *Координиране на езиковите ресурси в Европа*

Как можем да бъдем полезни?

Христина Добрева

European Language Resource Coordination
Connecting Europe Facility

Участници

- ✓проф. Светла Коева, Институт за български език;
- ✓Светлана Юрукова, 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

European Language Resource Coordination
Connecting Europe Facility

Единен цифров пазар и многоезичието

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.

European Language Resource Coordination
Connecting Europe Facility

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 памети:** 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.

Как можете да участвате? 1/8

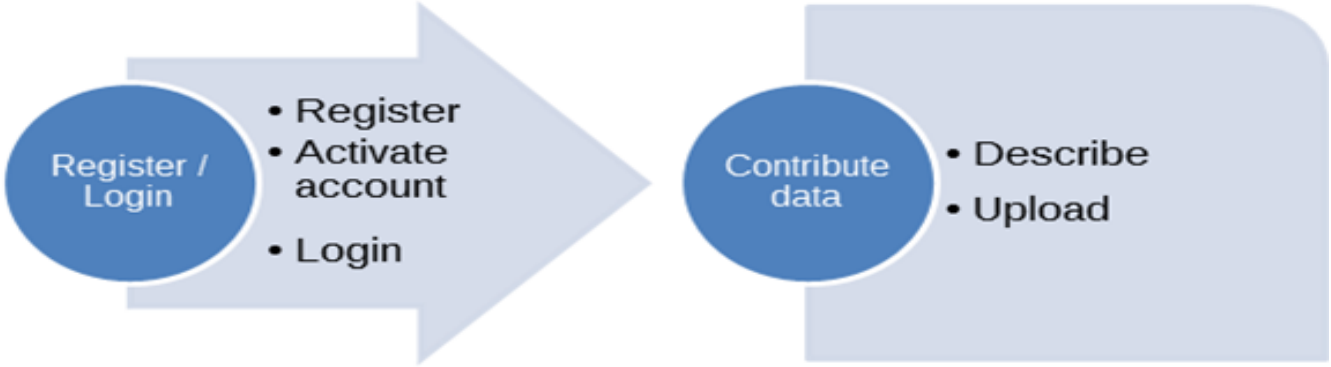
Данните могат да бъдат предоставяни основно по два канала – чрез проекта за Координиране на езикови ресурси в Европа и чрез Европейската комисия.

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.



Как можете да участвате? 2/8

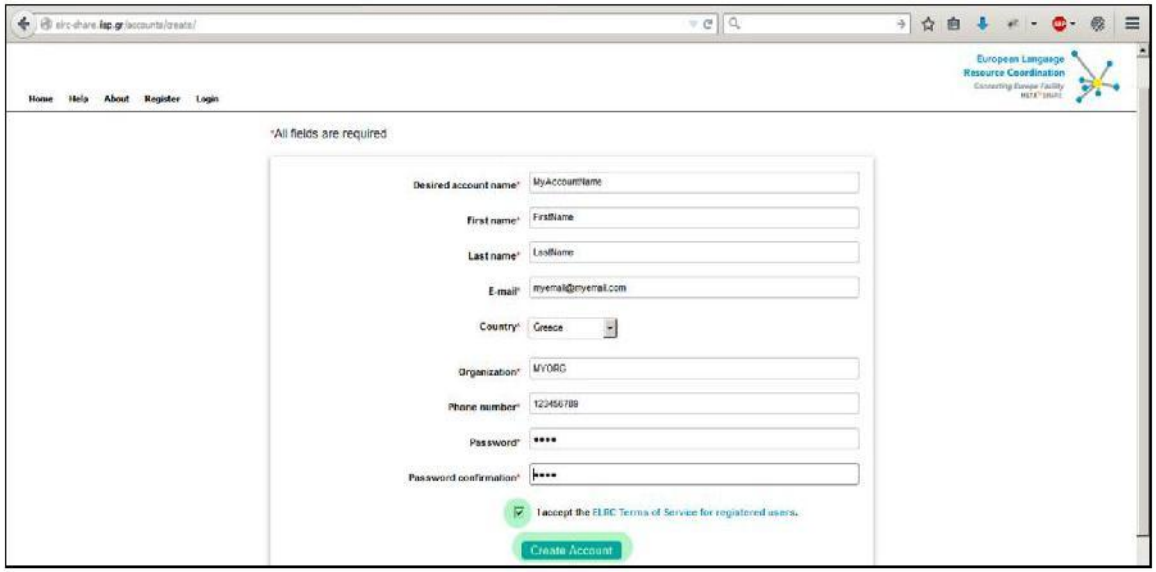
Можете да предоставяте данни само като регистриран потребител. Регистрацията е от основно значение за идентификацията на доставчика на езикови ресурси. Чрез нея се следи за злоупотреби, поддържа се сигурна и надеждна среда, създава се регистър на доставчиците и се управлява употребата на самите ресурси.

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

Как можете да участвате? 3/8

За създаване на профил е необходимо да попълните всички оказани полета.

Следва получаване на потвърждение на Вашата електронна поща.



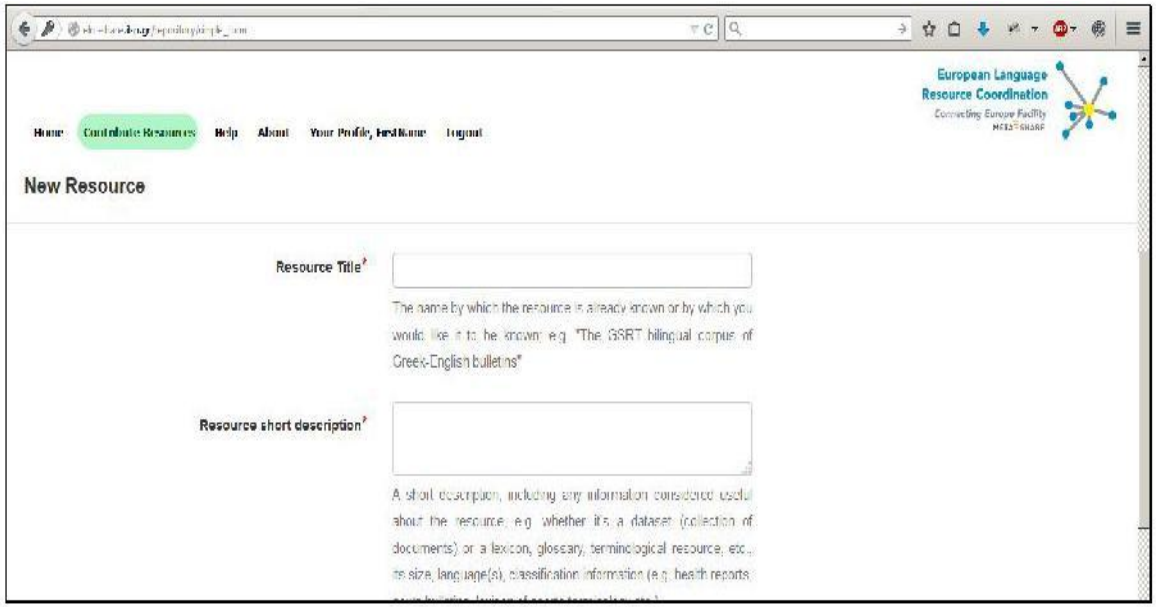
Как можете да участвате? 4/8

В електронното съобщение има връзка за активация.

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

Как можете да участвате? 5/8

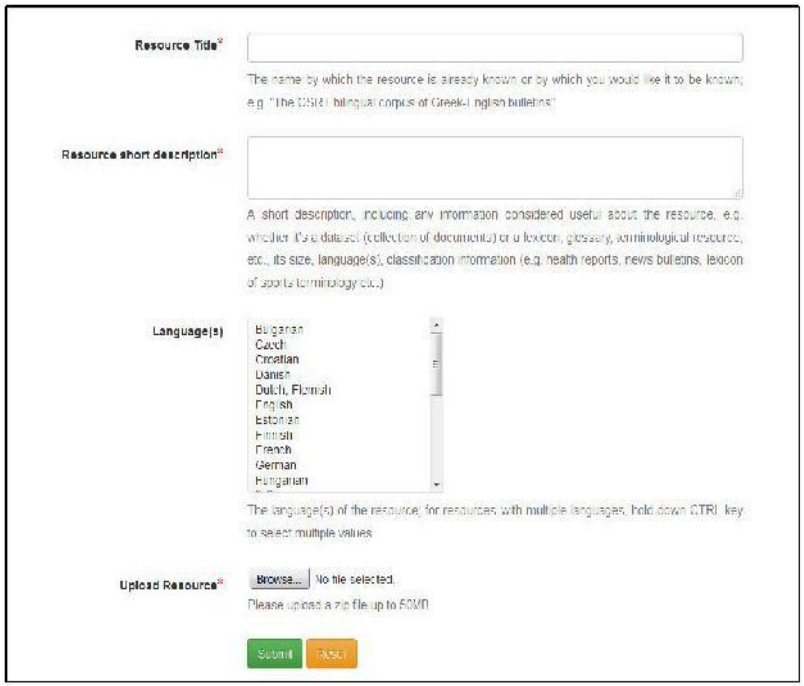
При подаването на данни се изисква попълване на кратка уеб форма. За достъпване на формата трябва да натиснете бутона **„Подаване на ресурси”** (Contribute Resources) от главното меню.



European Language Resource Coordination
Connecting Europe Facility

Как можете да участвате? 6/8

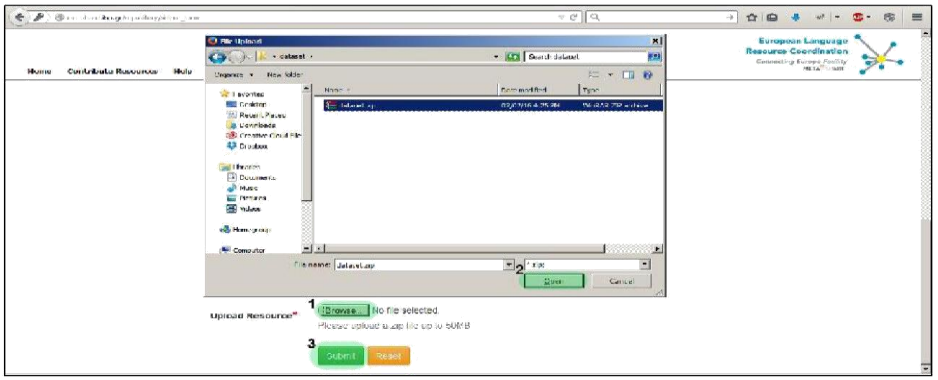
Формулярът е много опростен и съдържа полета за попълване на информация, описваща ресурса (а именно: име на ресурса, езици, кратко описание на ресурсите) и бутон за качването на самия ресурс. Полета със звездичка (*) са задължителни.



Как можете да участвате? 7/8

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.



Как можете да участвате? 8/8

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 съответсвие с условията им за лицензиране.