European Language Resource Coordination
Connecting Europe Facility

Втори национален семинар ***Координиране на езиковите ресурси в Европа***

Координиране на езиковите ресурси в Европа в България

Светла Коева

Институт за български език
Българска академия на науките

Езикови ресурси

- **Езикови ресурси:** езикови данни в електронна форма, които се използват за **създаване, подобряване и оценка** на системи за **компютърна обработка на езика** (Cucchiarini et al.,2001) като автоматичния превод.

Cucchiarini, C., Daelemans, W .& Strik, H. Strengthening the Dutch Human Language Technology Infrastructure. – In: *The ELRA Newsletter*, 2001, **6**:4, pp. 3-7.

European Language Resource Coordination
Connecting Europe Facility

Езикови ресурси

- В зависимост от тяхното съдържание и структура, езикови ресурси се делят на (Gavrilidou et al., 2012):
  - **корпуси**;
  - **лексикално-семантични ресурси**;
  - описание на езикови правила (**граматики**);
  - **програми за обработка на езика.**

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., Frontini, F., Declerck, T., Francopoulo, G., Arranz, V. & Mapelli, V. The META-SHARE Metadata Schema for the Description of Language Resources. – In: Nicoletta Calzolari, N., Choukri, K., Declerck, T., Doğan, M. M. U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.): *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, ELRA, European Language Resources Association (ELRA), Paris, France, Paris, 5/2012.

Езикови ресурси

- Предназначение на езиковите ресурси за български, събирани в рамките на проекта ***Координиране на езиковите ресурси в Европа***:
  - за **обучение на Платформата за автоматичен превод *eTranslation*** за превод от и на български език;
  - за **подобряване на качеството на автоматичния превод** от и на български език;
  - за **използване на автоматичния превод** от публичната администрация.

European Language Resource Coordination
Connecting Europe Facility

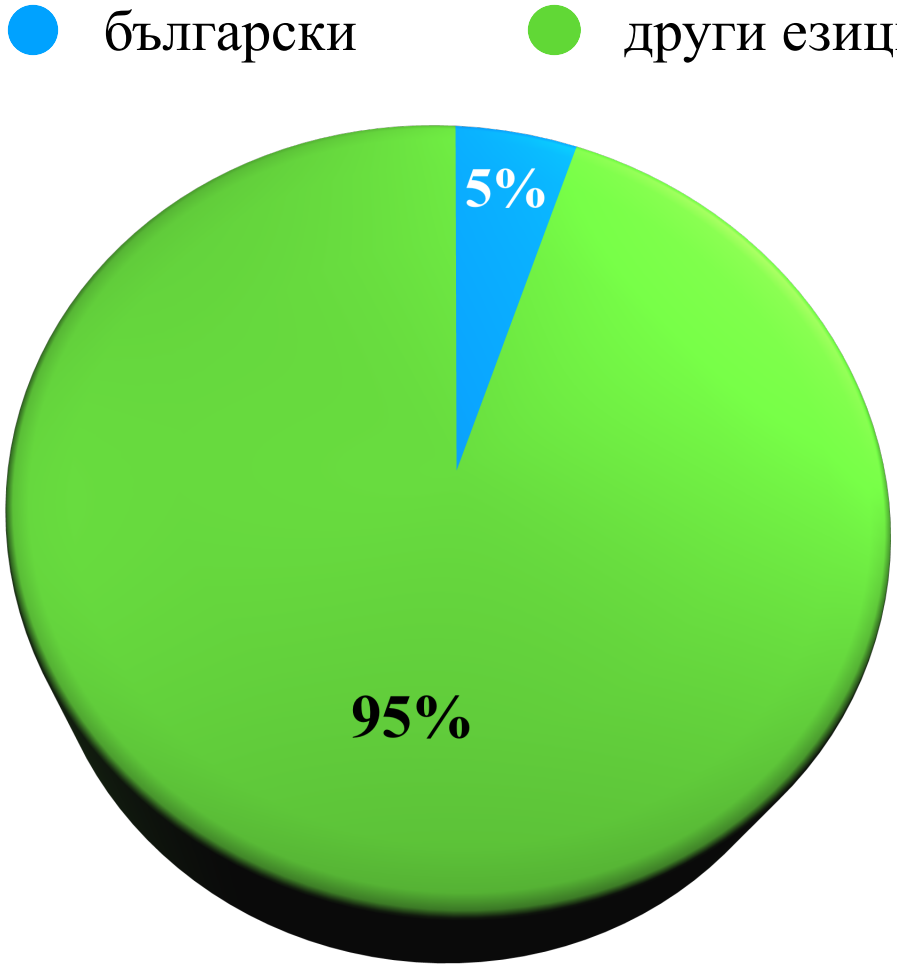
Представители на публичната администрация

Първи национален семинар, 18 март 2016: 130 участници

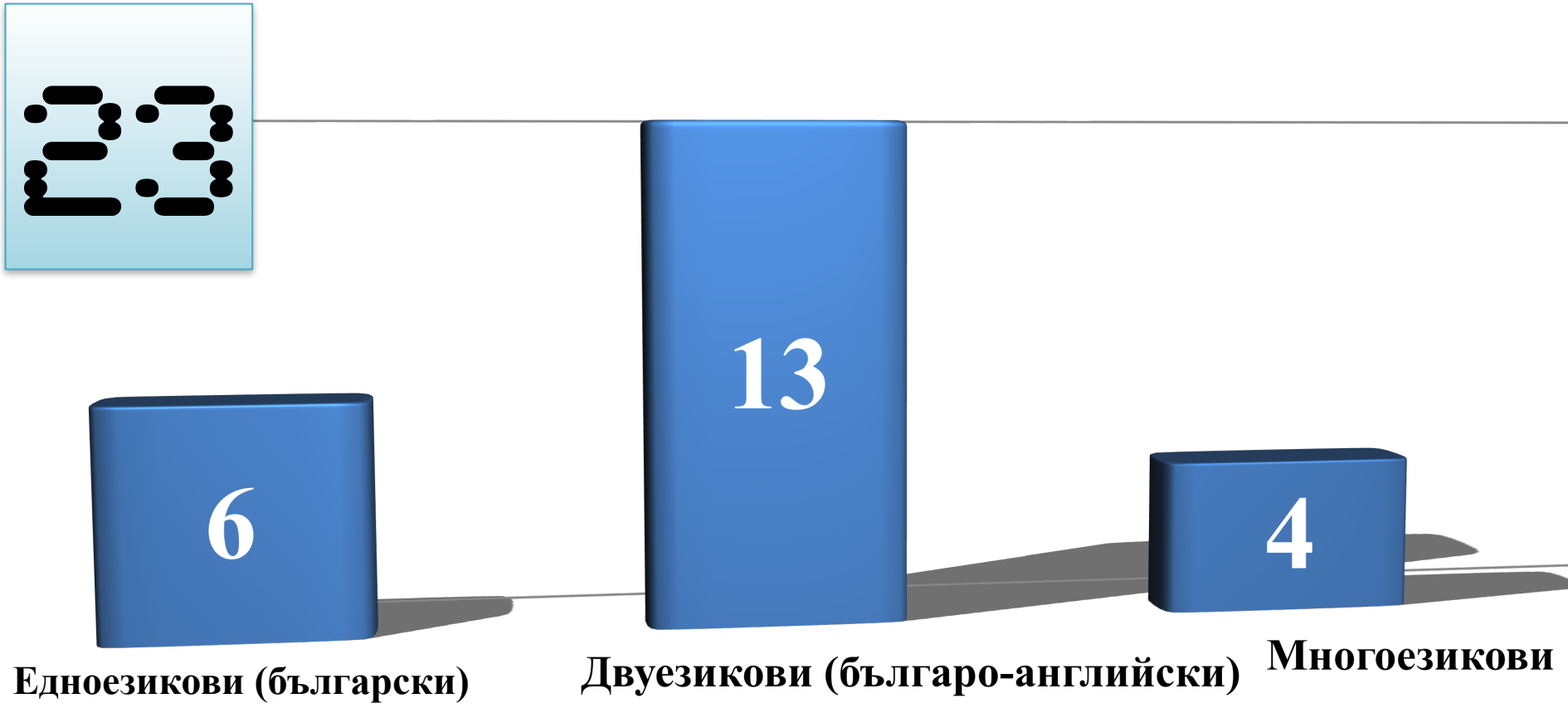
- Министерски съвет на Република България; Президентство на Република България; Министерство на транспорта, информационните технологии и съобщенията; Министерство на правосъдието; Министерство на финансите; Министерство на околната среда и водите; Министерство на енергетиката; Министерство на земеделието и храните; Министерство на външните работи; Министерство на отбраната; Министерство на вътрешните работи; Министерство на регионалното развитие и благоустройството; Министерство на туризма; Министерство на младежта и спорта;
- Национален институт на правосъдието; Национална агенция за приходите; Българска агенция за безопасност на храните; Висш съдебен съвет; Национален статистически институт; Агенция „Митници“; Изпълнителна агенция по морска администрация; Изпълнителна агенция „Автомобилна администрация“; Български институт за стандартизация; Главната дирекция „Гражданска въздухоплавателна администрация“, Национална агенция за професионално образование и обучение; Българска агенция за инвестиции; Агенция „Пътна инфраструктура“; Агенция за държавна финансова инспекция; Изпълнителна агенция „Електронни съобщителни мрежи и информационни системи“; Изпълнителна агенция „Железопътна администрация“; Изпълнителна агенция по горите; Държавна агенция за закрила на детето; Институт по публична администрация; Изпълнителна агенция по сортоизпитване, апробация и семеконтрол.

European Language Resource Coordination
Connecting Europe Facility

Езикови ресурси към 24.10.2018



Езикови ресурси към 24.10.2018



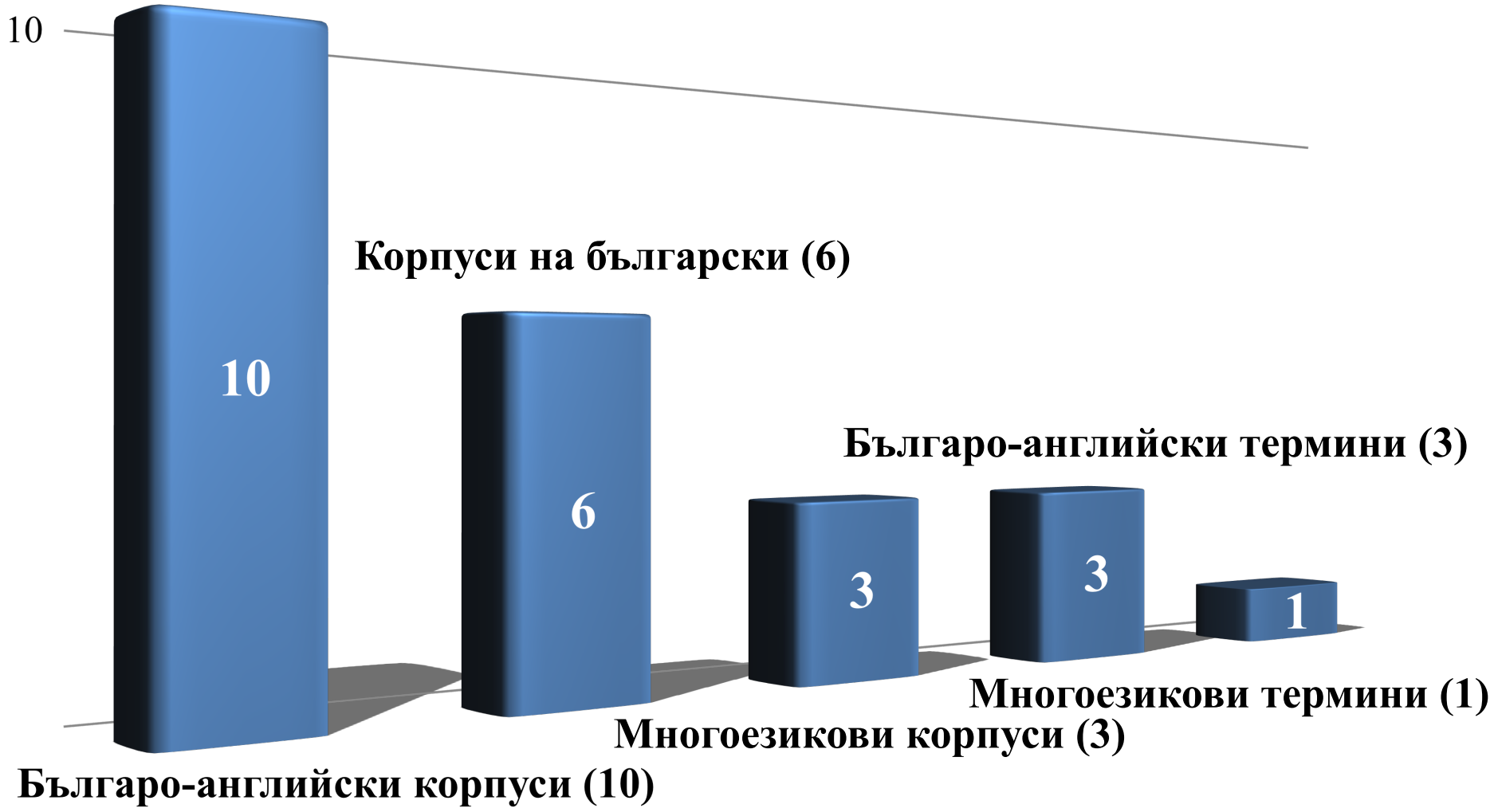
European Language Resource Coordination
Connecting Europe Facility

Количествени характеристики

104 504 съотнесени единици между български и английски

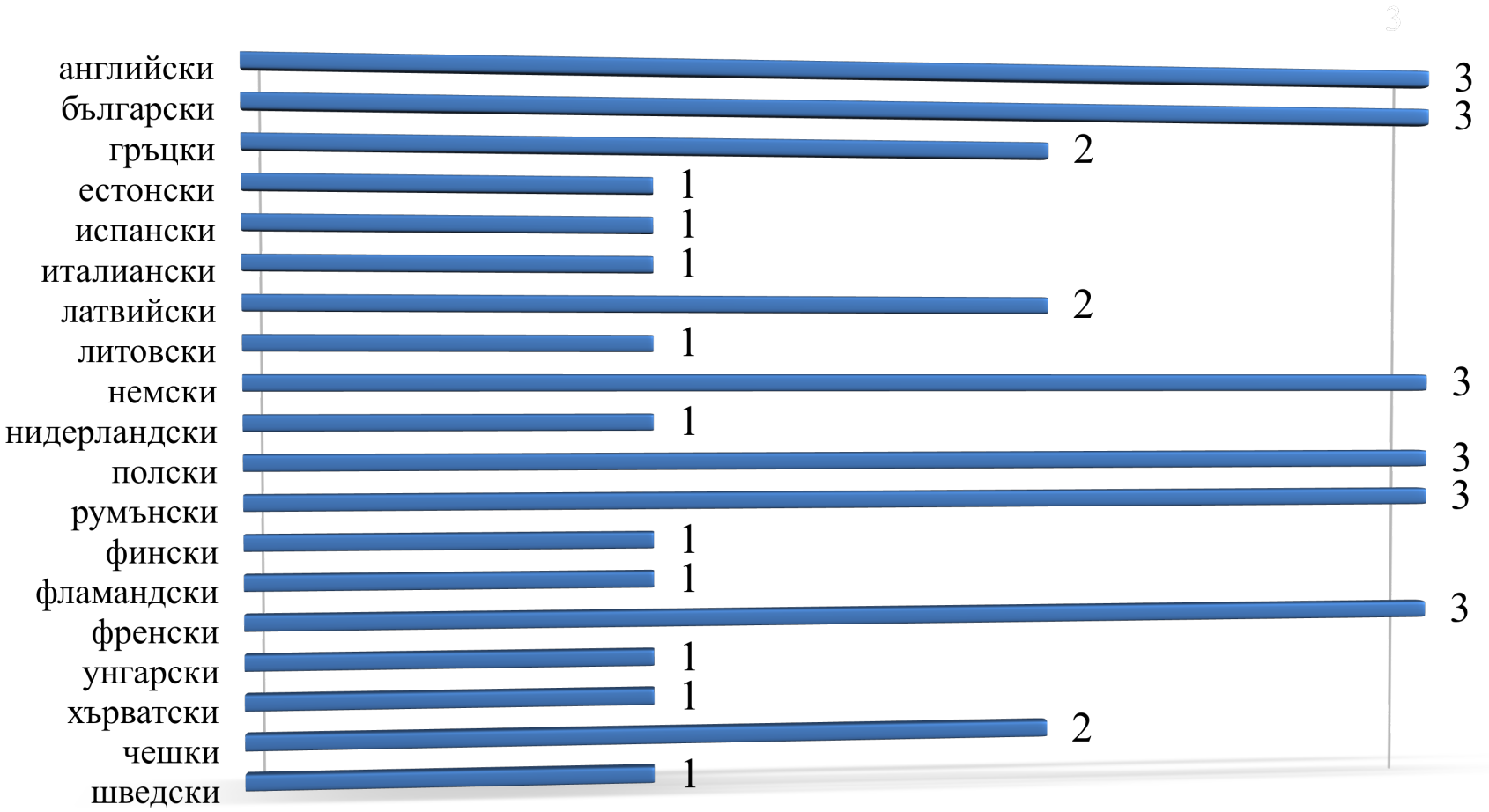
21 659 096 думи на български

Видове езикови ресурси

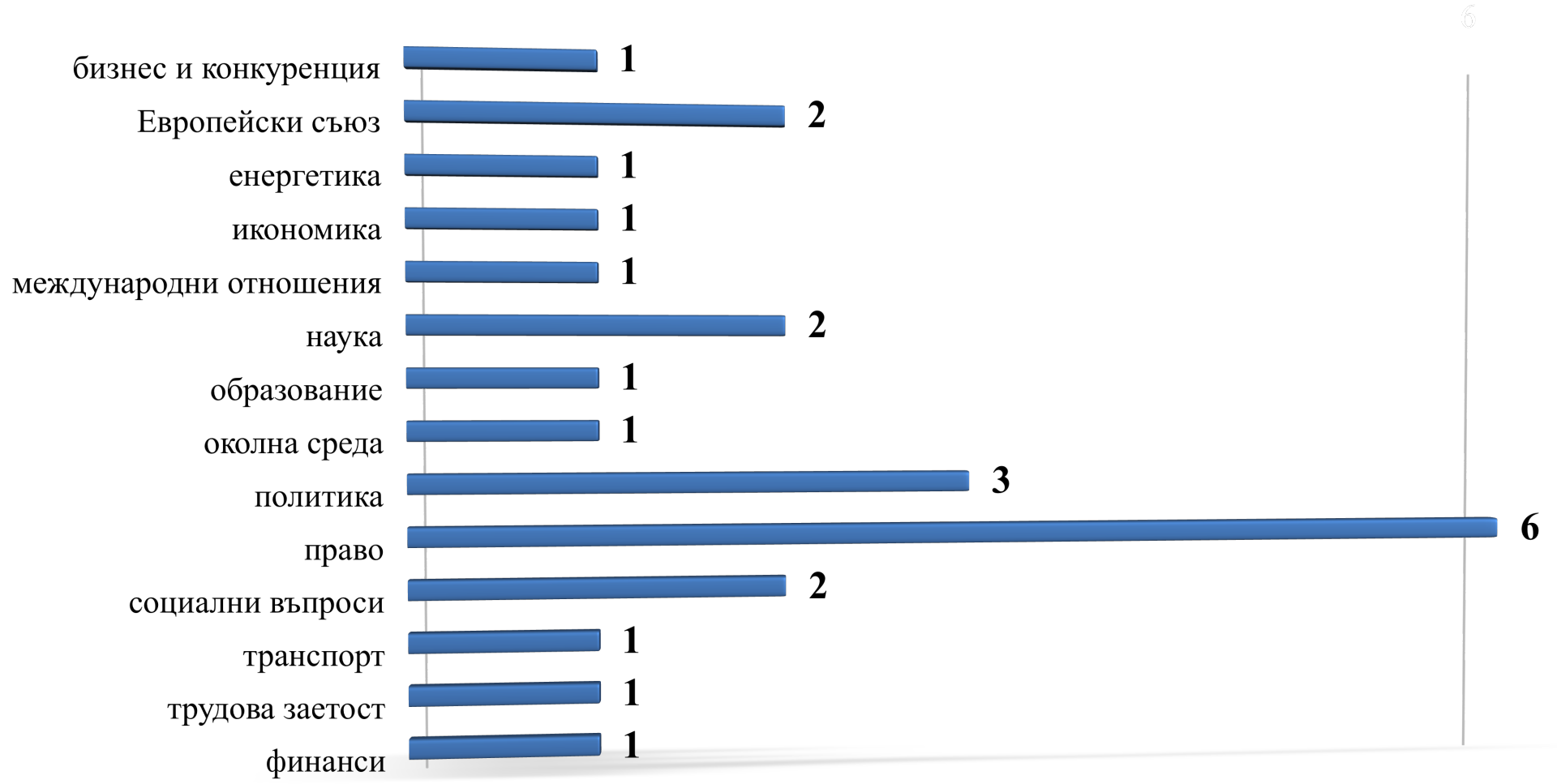


European Language Resource Coordination
Connecting Europe Facility

Многоезикови ресурси

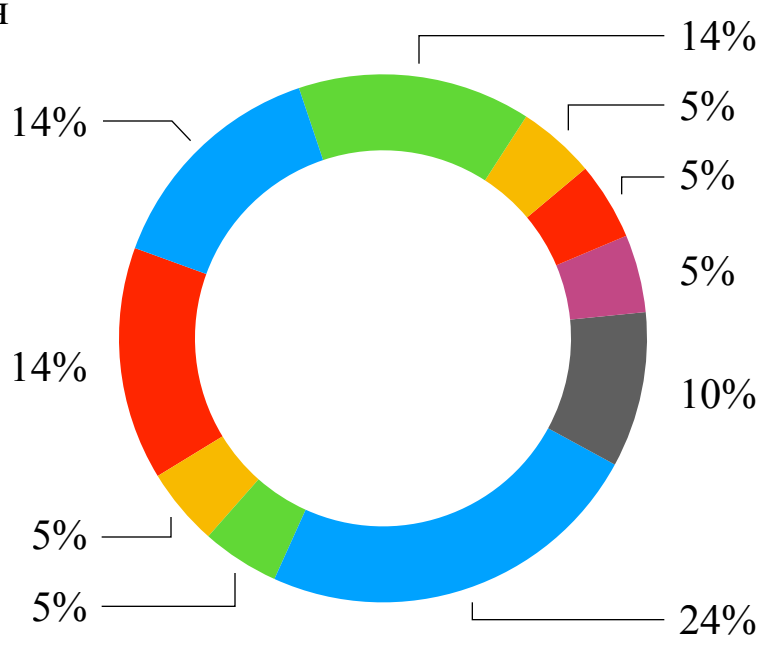


Тематични области



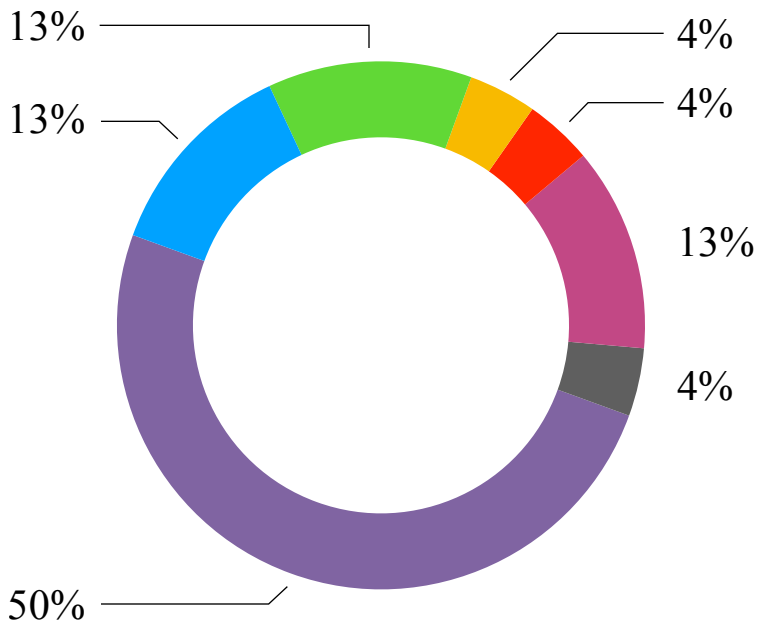
От кои институции са предоставени?

- Министерство на транспорта, информационните технологии и съобщенията
- Министерство на правосъдието
- Национален институт на правосъдието
- Национална агенция за приходите
- Проект „Автоматичният превод на CEF за Председателството на СЕ“
- Пресцентър и информационно бюро на Република Кипър
- Институт за обработка на език и реч, Гърция
- Шведски езиков съвет
- Университетът във Виена
- Институт за български език



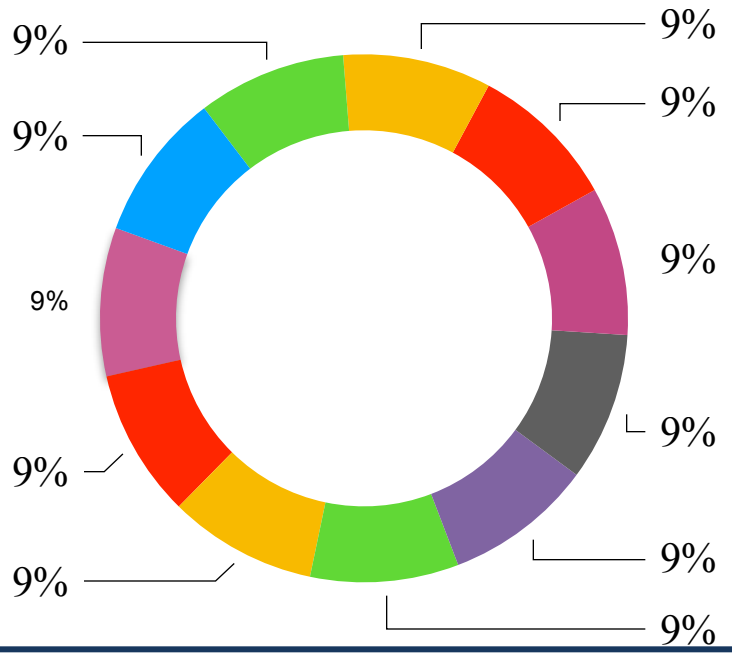
От кои институции са предоставени?

- Министерство на транспорта, информационните технологии и съобщенията
- Министерство на правосъдието
- Национален институт на правосъдието
- Национална агенция за приходите
- Институт за български език
- Проект „Автоматичният превод на CEF за Председателството на СЕ“
- Други страни



От кои институции са предоставени?

- Президентство на Република България
- Министерски съвет на Република България
- Министерство на транспорта, информационните технологии и съобщенията
- Министерство на земеделието, храните и горите
- Министерство на образованието и науката
- Министерство на околната среда и водите
- Министерство на правосъдието
- Министерство на отбраната
- Министерство на икономиката
- Национална агенция за приходите
- Национален институт на правосъдието



European Language Resource Coordination
Connecting Europe Facility

Обработка на езиковите ресурси

Стандартизиране на формàта

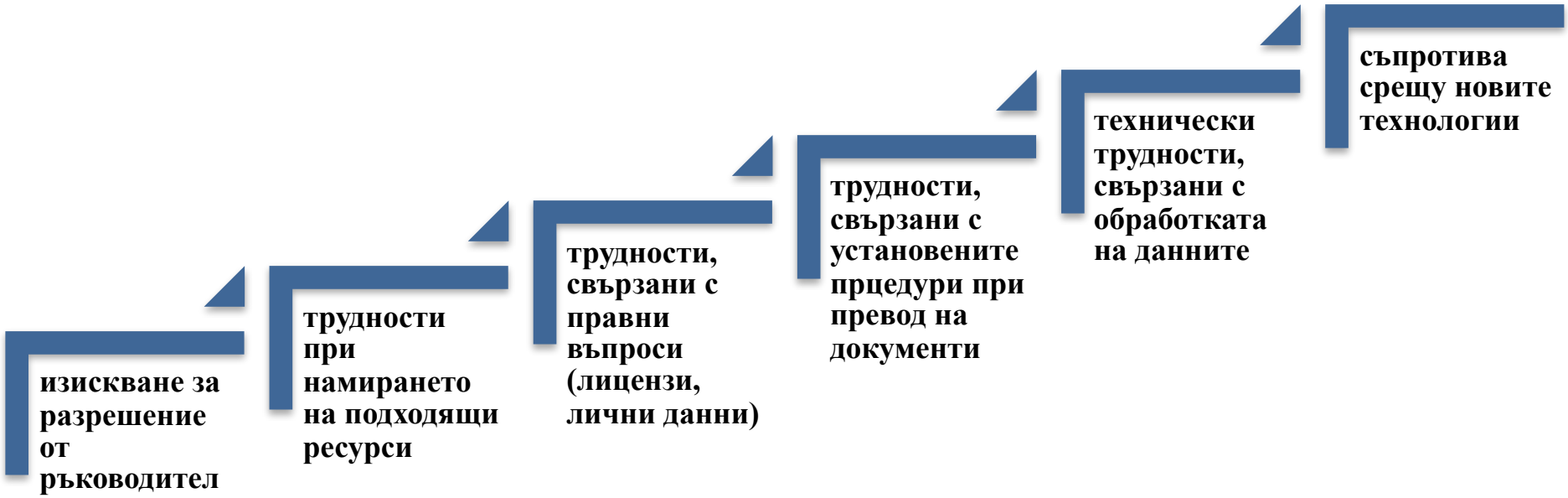
Съотнасяне по изречения

Описание на метаданните

Запазване на ресурсите в хранилището ELRC-споделяне

Препятствия и постижения

Основни препятствия



Препятствия и постижения

Основни постижения

- Единна процедура за предоставяне на ресурси
- Разбиране за важността на ресурсите
- Предоставяне на паралелни езикови ресурси
- Ресурси от разнообразни тематични области
- Паралелни ресурси на български и 18 езика

Обобщение

Добри практики	Подходящи са ресурсите, създадени от професионални преводачи, които използват компютърно подпомогнат превод
	Голяма част от документите, създадени от публичната администрация, са отворени данни
Пречки	Недостатъчно разбиране за важността на споделянето на езикови данни и за обогатяването им с метаданни
	Недостатъчен интерес към автоматичния превод

European Language Resource Coordination
Connecting Europe Facility

Обобщение

Бъдеща работа

- Популяризиране на важността на езиковите данни за подобряване на качеството на автоматичния превод
- Разширяване на кръга от участници в процеса на предоставяне на езикови ресурси за нуждите на автоматичния превод
- Оптимизиране на процедурата за предоставяне на данни, включително за анонимизирането им.
- Преизползване и натрупване на преводна памет при превод на документи от една тематична област

European Language Resource Coordination
Connecting Europe Facility

Благодаря Ви за вниманието!

Имейл: info@lr-coordination.eu
Интернет адрес: www.lr-coordination.eu