Какви данни са необходими?

Проф. д-р Светла Коева
Институт за български език
„Проф. Любомир Андрейчин“
Българска академия на науките

Какви данни са необходими?

- За да се подобри качеството на услугите, предоставяни от *Платформата за автоматичен превод CEF.AT*, системите за машинен превод трябва да бъдат обучени върху **качествени езикови данни на всички официални езици** на страните, участващи в Механизма за свързване на Европа.

Какви данни са необходими?

- **Корпуси**: едноезикови и многоезикови корпуси със сравними, паралелни и съотнесени документи.
- **Речници**: едноезикови и многоезикови речници, в които лексикалните единици са съотнесени с различен тип информация.
- **Корпусите и речниците** се използват при **обучението на системи за машинен превод**.

Какви данни са необходими?

- **Преводна памет**: езикови бази от данни, които съдържат преводи, направени от специалисти. Те могат да се използват за улеснение при нови преводи и при обучението на системи за машинен превод.
- **Езикови и преводни модели**: статистическа информация за последователностите от думи, които се срещат в даден език, и преводните съответствия на думи и изрази в два езика. Използват се за статистически машинен превод.

Корпуси

- Голяма колекция от езикови примери в електронна форма, избрани по определени критерии, така че да представляват адекватно даден език, негово състояние, тематична област или група езици.

Многоезикови корпуси

- Съдържат текстове на повече от един език.

English

български

Паралелни корпуси

- Представляват текстове, които са преводни еквиваленти и може да са съотнесени по изречения или по думи.

Съотносими корпуси

- Съвкупност от сходни по съдържание текстове на повече от един език.

Работният екип на България за номинирането на Ирина Бокова за генерален секретар на ООН включва 16 български дипломати. Това заяви ръководителят на екипа Райко Райков, ръководител на генерална дирекция „Глобални въпроси”.

Ирина Бокова бе номинирана за генерален **секретар** на **ООН** с писмо от министъра на външните работи на 9 февруари 2016 г., във връзка с Решение № 404 от 19 юни 2014 година на Министерския съвет.

български

Группа по координации работы кандидатуры Болгарии на пост Генерального секретаря Организации Объединенных Наций (ООН) за период 2017-2021 года Ирины Боковой будет представлена сегодня в МИД.

Ирина Бокова была выдвинута на пост Генерального секретаря **ООН** в письме министра иностранных дел 9 февраля 2016 года, в связи с Решением № 404 от 19 июня 2014 года Совета министров, передает БГНЕС.

русский

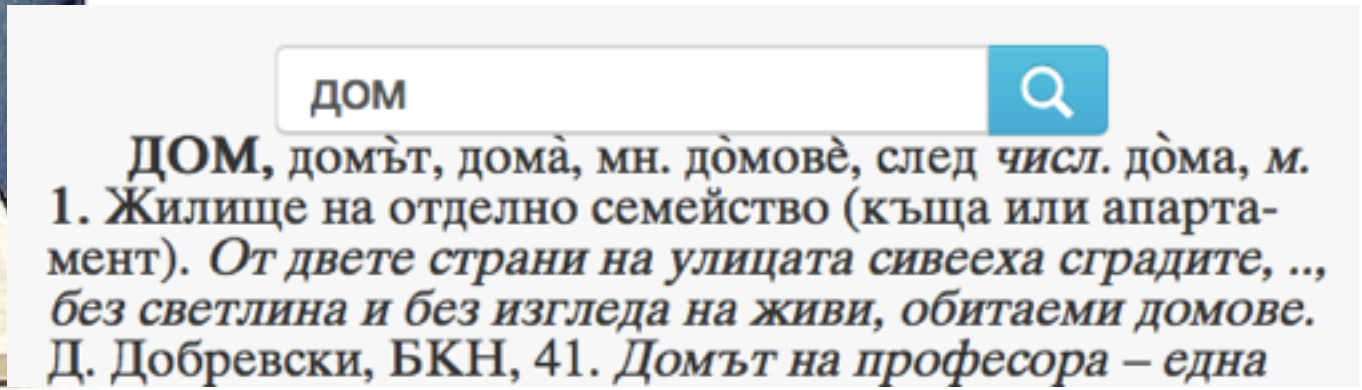
http://novinite.bg

Какви данни са необходими?

- Копрусите може да съдържат:
  - новини на един или няколко езика;
  - уеб съдържание, поддържано на няколко езика;
  - официални документи на един или няколко езика;
  - закони, правилници, постановления, наредби;
  - архиви;
  - анализи, отчети, бюлетини;
  - често задавани въпроси и отговори;
  - други документи, свързани с дейността на публичната администрация.

Речници

- Електронните речници съдържат различна фонетична, лексикална, граматична или семантична информация, съотнесена с лексикалните единици.



http://ibl.bas.bg/rbe/

Терминологични речници

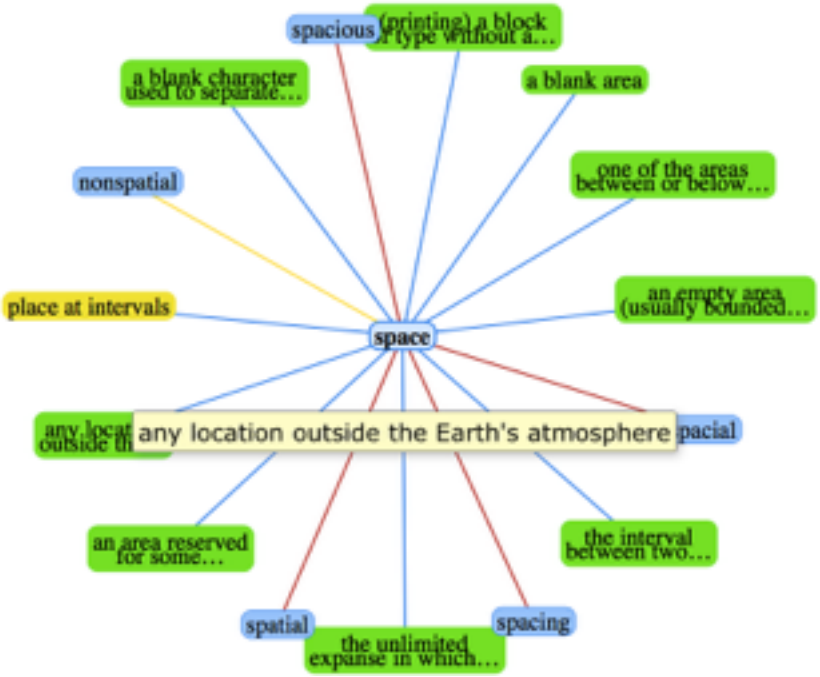
- Съдържат множество от термини от дадена специализирана област, снабдени с подходяща лингвистична информация.

BG	автоматични врати автоматични гранични врати
CS	automatizovaná brána
DE	automatisches Kontrollgate
EN	automated gate automated border gate
ES	barrera automática

http://iate.europa.eu

Онтологии

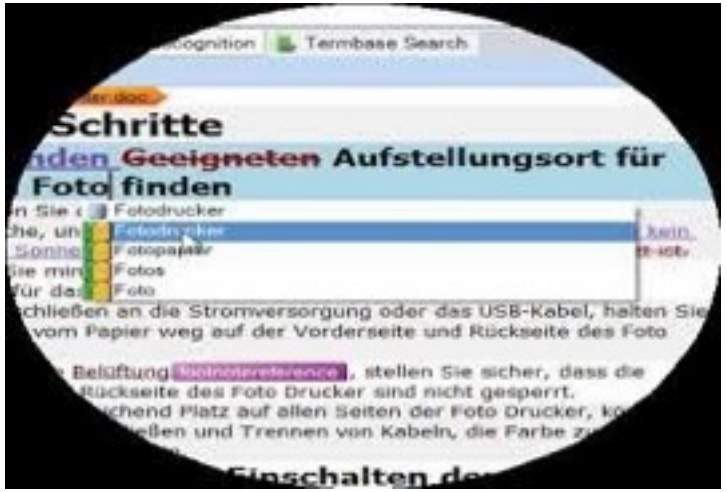
- Представят класове от сферата на човешкото познание, например пространство, време, обект, събитие, действие, количество и т.н., техни подкласове и релациите между тях.



http://wordventure.eti.pg.gda.pl/wne/wne.html

Преводна памет

- Съхранява еквивалентни фрази, изречения или по-големи сегменти. При превод може да се направи справка с предишно решение и да се използва.



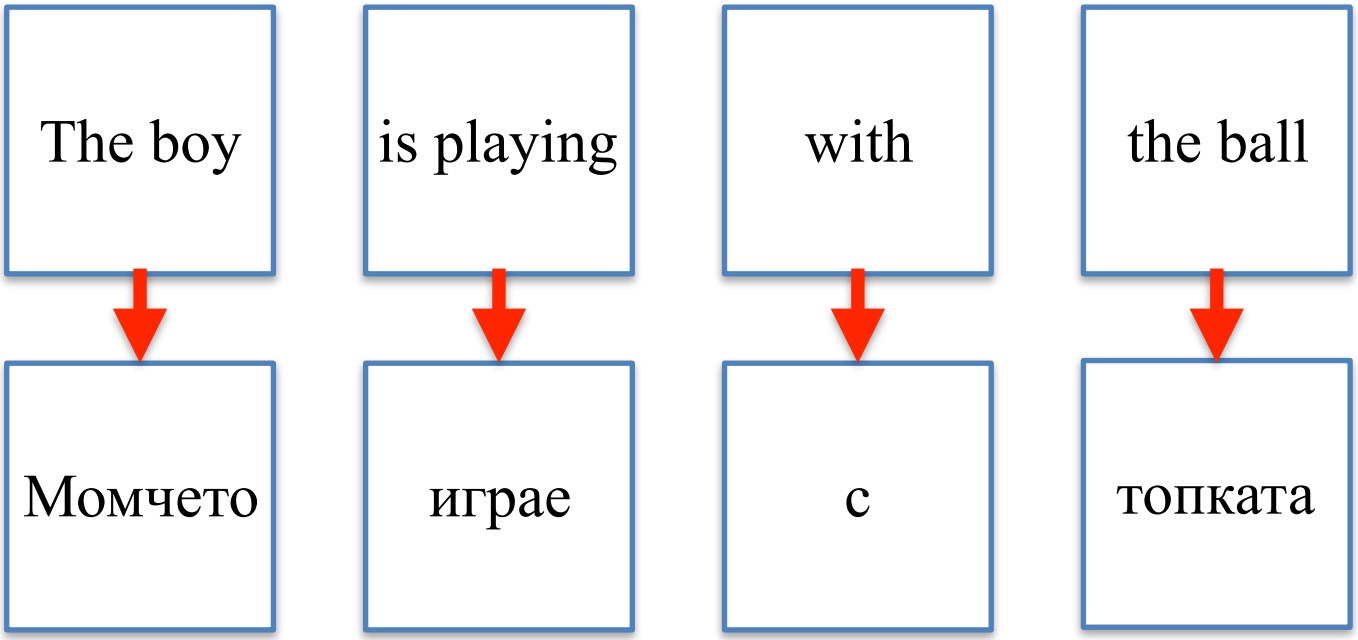
Езикови модели

- Съдържат множество от правилни последователности от думи на езика, на който се превежда.

  - *хубава жена*
  - *хубавата жена*
  - *хубава жена от Пловдив*
  - *хубавата жена от Пловдив*
  - *тази хубава жена*
  - *тази хубава жена от Пловдив*

Преводни модели

- Преводният модел се използва, за да се предскаже вероятността за превод на дума или последователност от думи от един език на друг.

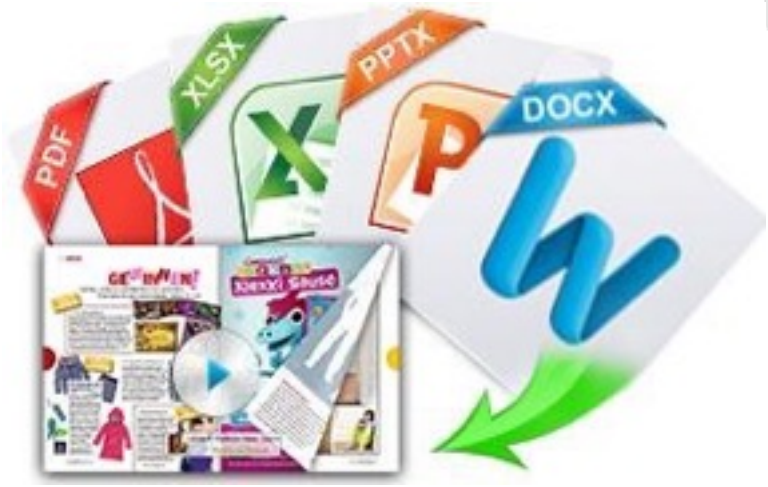


Какви данни са необходими?

- Някои типове данни могат да се използват директно от *Платформата за автоматичен превод CEF.AT* (преводна памет, корпуси на един език, паралелни корпуси със съотнесени единици, терминологични ресурси, речници).
- Други типове данни (новини, нормативни документи и др.) могат да станат основа за създаване на нови езикови ресурси (паралелни корпуси, съотнесени на определено равнище; преводни и езикови модели).

Какъв формат е нужен?

- Документите трябва да са достъпни във формат, който да улеснява **повторната им употреба**.



Път на данните

- Идентифициране на подходящи езикови данни;
- Изясняване на авторските права;
- Документиране на метаданните;
- Предварителна обработка на текстовете; (отстраняване на линкове, реклами, темплейти, изображения и др.);
- Разделяне на текстовете на изречения и съотнасяне на откритите преводни еквиваленти;
- Създаване на езикови и преводни модели.

Идентифициране


Official Journal
of the European Union
C 326
Information and Notices
Volume 55
26 October 2012
Contents
page
Consolidated versions of the Treaty on European Union and the Treaty on the Functioning of the European Union 1
Consolidated version of the Treaty on European Union 13
Consolidated version of the Treaty on the Functioning of the European Union 47
Protocols 201
Annexes 331
Declarations annexed to the Final Act of the Intergovernmental Conference which adopted the Treaty of Lisbon, signed on 13 December 2007 337
Tables of equivalences 363
Charter of Fundamental Rights of the European Union 391
EN

Английска версия

Българска версия


Официален вестник
на Европейския съюз
C 326
Информация и известия
Година 55
26 октомври 2012 г.
Съдържание
Страница
Консолидирани текстове на Договора за Европейския съюз и на Договора за функционирането на Европейския съюз 1
Консолидиран текст на Договора за Европейския съюз 13
Консолидиран текст на Договора за функционирането на Европейския съюз 47
Протоколи 201
Приложения 331
Декларации, приложени към заключителния акт на Междуправителствената конференция, която прие Договора от Лисабон, подписан на 13 декември 2007 г. 337
Таблици на съответствието 363
Харта на основните права на Европейския съюз 391
BG

http://eur-lex.europa.eu/

Документиране

- Метаданни:
  - заглавие;
  - автор;
  - тема;
  - тематична област;
  - дата на създаване;
  - формат;
  - издател;
  - авторски права.

http://dublincore.org/documents/dces/

Обработка

[...]

INFORMATION FOR NATIONAL PARLIAMENTS

Article 1

Commission consultation documents (green and white papers and communications) shall be forwarded directly by the Commission to national Parliaments upon publication. The Commission shall also forward the annual legislative programme as well as any other instrument of legislative planning or policy to national Parliaments, at the same time as to the European Parliament and the Council.

Article 2

Draft legislative acts sent to the European Parliament and to the Council shall be forwarded to national Parliaments.

For the purposes of this Protocol, "draft legislative acts" shall mean proposals from the Commission, initiatives from a group of Member States, initiatives from the European Parliament, requests from the Court of Justice, recommendations from the European Central Bank and requests from the European Investment Bank, for the adoption of a legislative act.

Draft legislative acts originating from the Commission shall be forwarded to national Parliaments directly by the Commission, at the same time as to the European Parliament and the Council.

Draft legislative acts originating from the European Parliament shall be forwarded to national Parliaments directly by the European Parliament.

Draft legislative acts originating from a group of Member States, the Court of Justice, the European Central Bank or the European Investment Bank shall be forwarded to national Parliaments by the Council.

[...]

[...]

ИНФОРМАЦИЯ, ПРЕДНАЗНАЧЕНА ЗА НАЦИОНАЛНИТЕ ПАРЛАМЕНТИ

Член 1

Консултативните документи на Комисията (зелени книги, бели книги и съобщения) се изпращат директно от Комисията на националните парламенти на държавите-членки при публикуването им. Комисията изпраща на националните парламенти и годишната законодателна програма, както и всеки друг инструмент за законодателно планиране или за политическа стратегия, като едновременно с това ги изпраща и на Европейския парламент и на Съвета.

Член 2

Проектите на законодателни актове, адресирани до Европейския парламент и до Съвета, се изпращат на националните парламенти.

За целите на настоящия протокол "проект на законодателен акт" означава предложенията на Комисията, инициативите на група държави-членки, инициативите на Европейския парламент, исканията на Съда, препоръките на Европейската централна банка и исканията на Европейската инвестиционна банка, които целят приемането на законодателен акт.

Проектите на законодателни актове, инициирани от Комисията, се изпращат пряко от Комисията на националните парламенти едновременно с изпращането им на Европейския парламент и на Съвета.

Проектите на законодателни актове, инициирани от Европейския парламент, се изпращат пряко от Европейския парламент на националните парламенти.

Съотнасяне

[...]

S1. INFORMATION FOR NATIONAL PARLIAMENTS

S2. Article 1

S3. Commission consultation documents (green and white papers and communications) shall be forwarded directly by the Commission to national Parliaments upon publication.

S4. The Commission shall also forward the annual legislative programme as well as any other instrument of legislative planning or policy to national Parliaments, at the same time as to the European Parliament and the Council.

S5. Article 2

S6. Draft legislative acts sent to the European Parliament and to the Council shall be forwarded to national Parliaments.

S7. For the purposes of this Protocol, "draft legislative acts" shall mean proposals from the Commission, initiatives from a group of Member States, initiatives from the European Parliament, requests from the Court of Justice, recommendations from the European Central Bank and requests from the European Investment Bank, for the adoption of a legislative act.

S8. Draft legislative acts originating from the Commission shall be forwarded to national Parliaments directly by the Commission, at the same time as to the European Parliament and the Council.

[...]

S1. ИНФОРМАЦИЯ, ПРЕДНАЗНАЧЕНА ЗА НАЦИОНАЛНИТЕ ПАРЛАМЕНТИ

S2. Член 1

S3. Консултативните документи на Комисията (зелени книги, бели книги и съобщения) се изпращат директно от Комисията на националните парламенти на държавите членки при публикуването им.

S4. Комисията изпраща на националните парламенти и годишната законодателна програма, както и всеки друг инструмент за законодателно планиране или за политическа стратегия, като едновременно с това ги изпраща и на Европейския парламент и на Съвета.

S5. Член 2

S6. Проектите на законодателни актове, адресирани до Европейския парламент и до Съвета, се изпращат на националните парламенти.

S7. За целите на настоящия протокол "проект на законодателен акт" означава предложенията на Комисията, инициативите на група държави-членки, инициативите на Европейския парламент, исканията на Съда, препоръките на Европейската централна банка и исканията на Европейската инвестиционна банка, които целят приемането на законодателен акт.

S8. Проектите на законодателни актове, инициирани от Комисията, се изпращат пряко от Комисията на националните парламенти едновременно с изпращането им на Европейския парламент и на Съвета.

Къде са данните?

- Институциите и гражданите, правителствените, неправителствените и частните организации, които работят в страните членки на Механизма на свързване на Европа, създават значителни по обем езикови данни, които биха били полезни за усъвършенстване на *Платформата за автоматичен превод CEF.AT*.

Необходимо количество

- Разпределението на лексиката над определена честота следва **Закона на Зипф**, който формулира закономерностите за разпределението на честотата на думите по следния начин – ако всички думи в езика (или от достатъчно голям текст) се подредят по честотата на срещането си, то честотата на n-тата думата в подредбата е приблизително обратно пропорционална на нейния пореден номер n, тоест около половината думи в даден корпус се срещат само един път, около една четвърт – само два пъти и т.н.

Zipf, G. K. 1935. The psychobiology of language. New York: Houghton Mifflin.

Необходимо количество

- Дж. Синклер илюстрира Закона на Зипф с данни от Браун корпус, в който има 69 002 уникални думи, от които 35 065 се срещат само един път.
- Най-често срещаната в корпуса дума *the* е употребена 69 970 пъти, което е почти два пъти повече от следващата по честота дума *of* – 36 410 пъти.

http://www.ahds.ac.uk/guides/linguistic-corpora/chapter1.htm

Необходимо количество

- Експерименти доказват, че един милиард думи не винаги са достатъчни за надеждни заключения по отношение на рядко срещани, но не необичайни думи като английската дума *tidiness* (подреденост), с което се потвърждава интуицията, че думите са неравномерно разпределени в различните текстове.

Необходимо количество

- Налага се заключението, че по-големият обем на корпусите и езиковите ресурси предполага по-достоверна илюстрация на по-широк кръг езикови явления (с по-висока честотата на срещане и разнообразна дистрибуция в различни тематични области) и по-ефективни решения на класификационни задачи или при машинно обучение.

Заключение

- За да се разработят приложения за машинен превод, които са полезни за европейска публична администрация и за потребителите на публични услуги в интернет, са необходими **подходящи едноезикови и паралелни ресурси, отразяващи специфичните тематични области и отделните европейски езици**.