Тестове с невромашинен превод в Европейската комисия

Борислав Георгиев
ГД „Писмени преводи“, ЕК

За какво ще говорим?

1. Системата за машинен превод eTranslation
2. Невромашинен превод
3. Наблюдения и примери

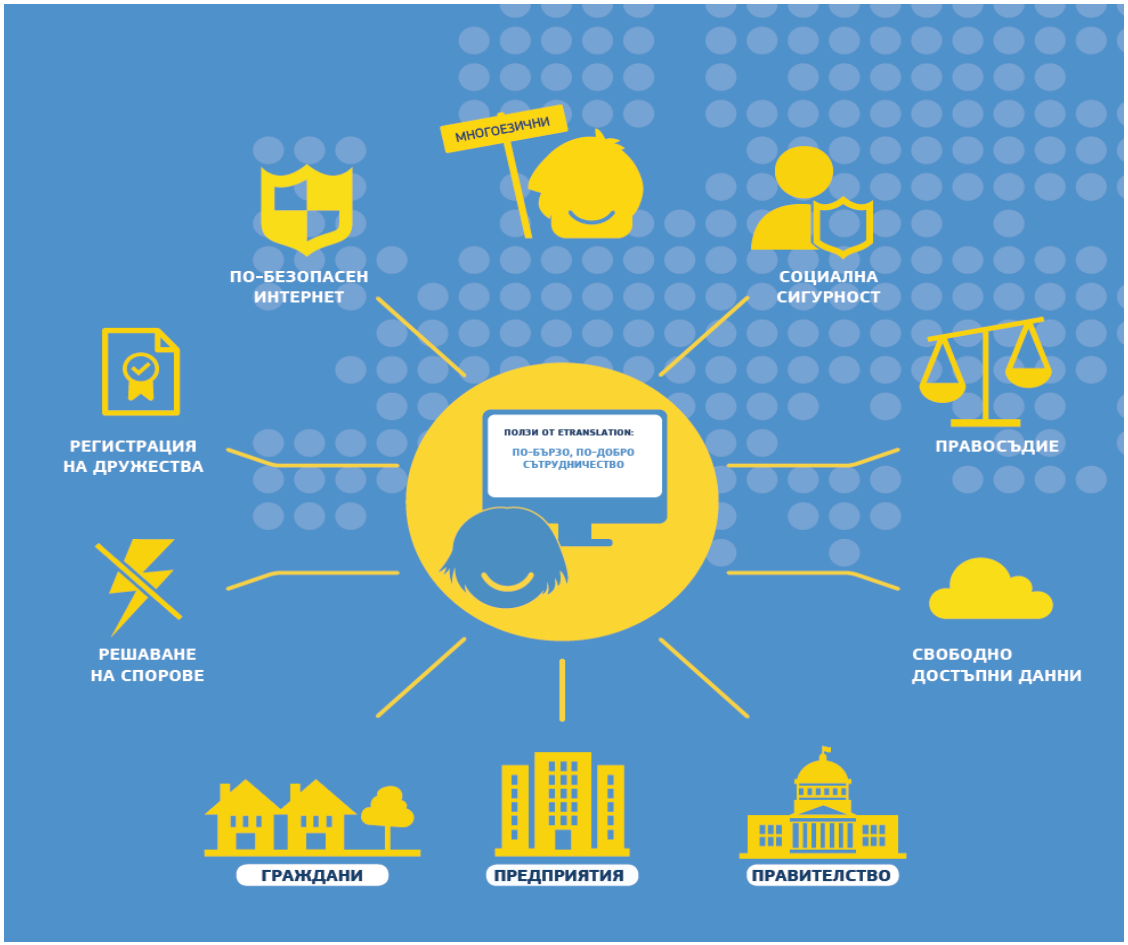
Системата за машинен превод eTranslation

eTranslation

Градивен елемент на Механизма за свързване на Европа

ЦЕЛ

Многоезични инфра-структури за цифрови услуги



Интегриране в цифрови услуги

ВИДОВЕ

Самостоятелен уеб интерфейс

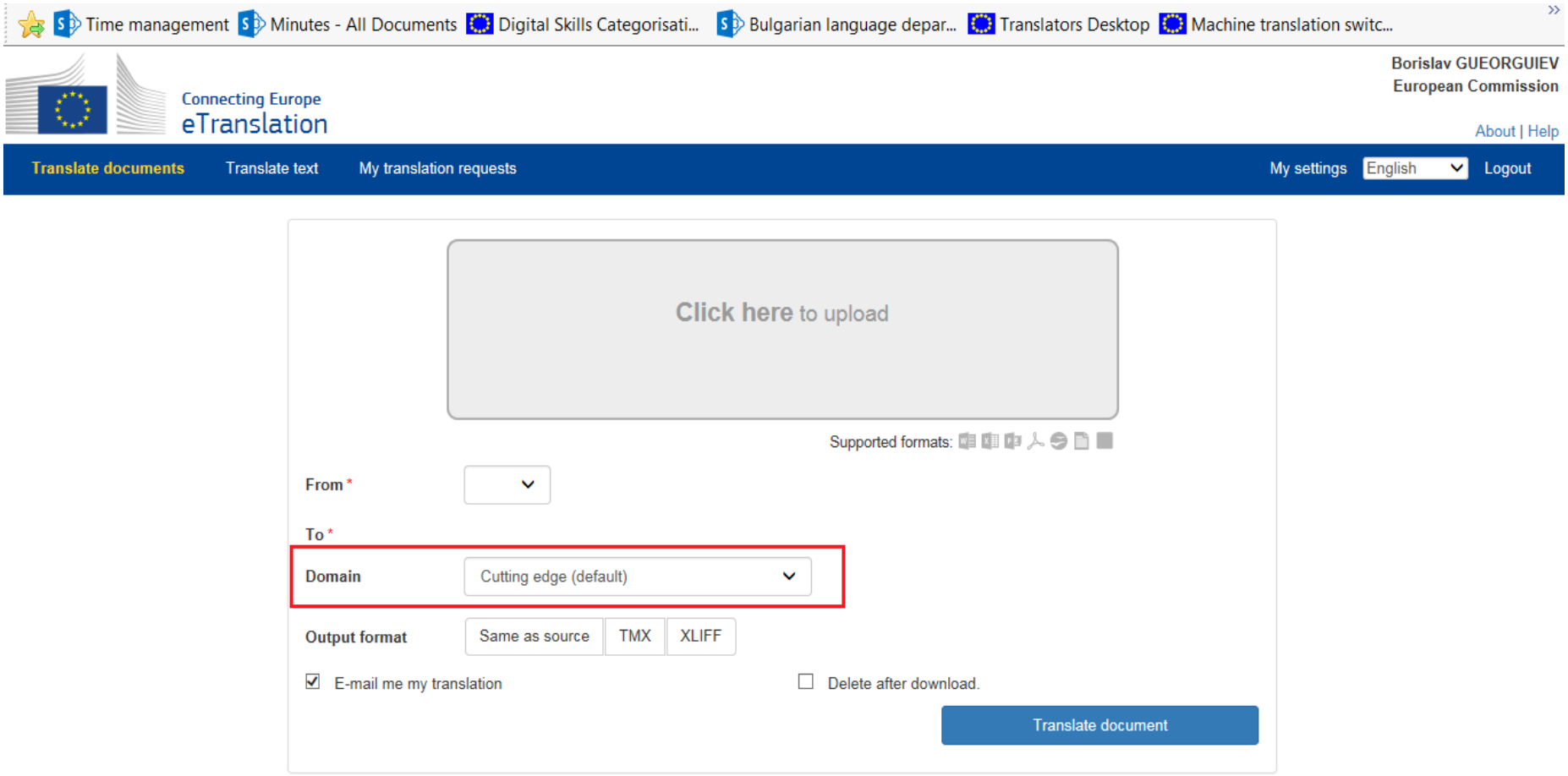
Какво е eTranslation?

- ✓ Система за машинен превод, разработена в Комисията – DGT, CNECT, DIGIT
- ✓ Обхваща всички официални езици на ЕС (552 езикови комбинации)
- ✓ Технология на базата на невронни мрежи → по-добро качество
- ✓ Обучена с данни на ЕС, съхранявани в базата данни Euramis → най-подходяща за свързани с ЕС документи

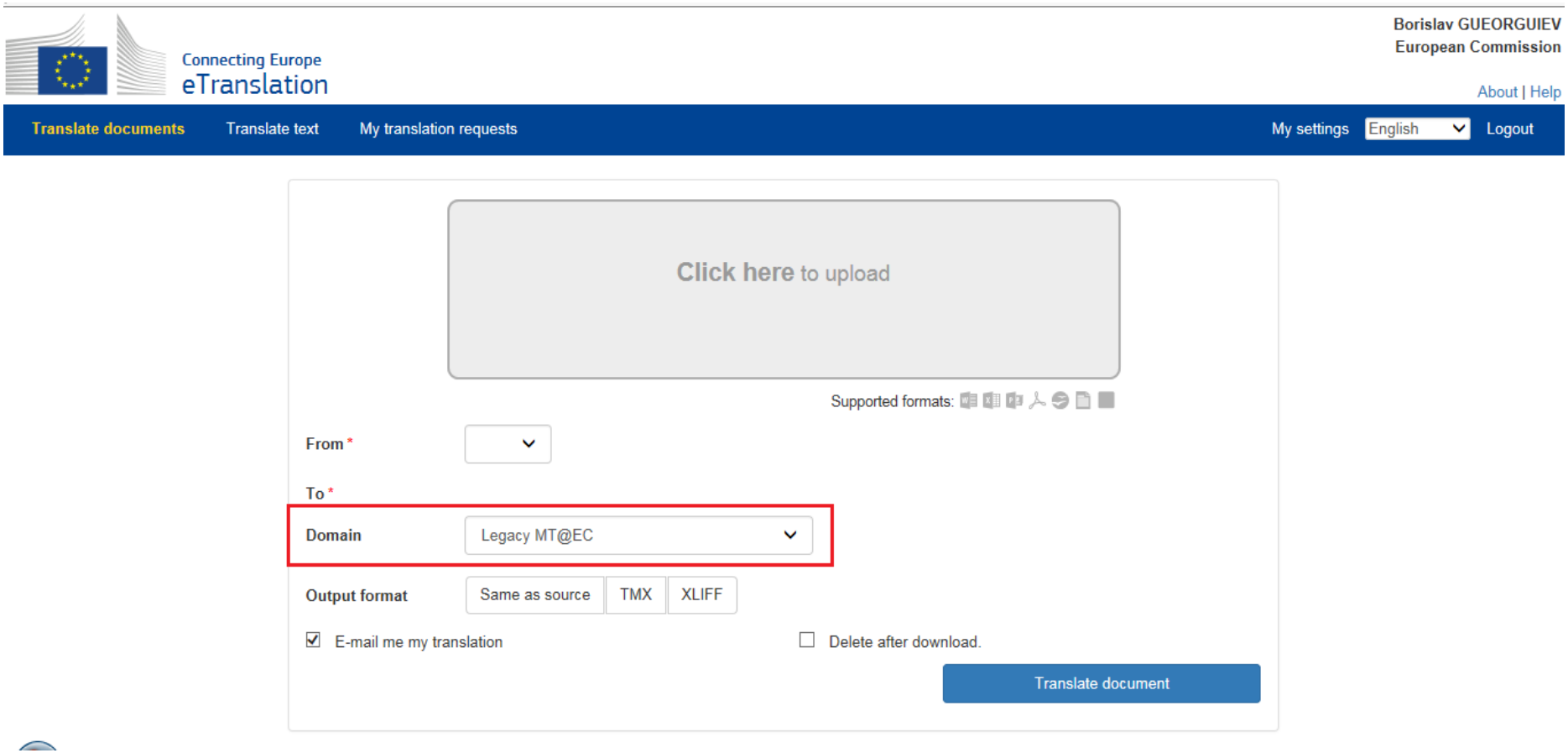
Кой ползва eTranslation?

- ✓ Институциите и органите на ЕС
- ✓ Държавната администрация в ДЧ + Исландия и Норвегия
- ✓ Програми в мрежата ЕМТ
- ✓ Онлайн услуги, финансирани или подкрепяни от ЕС (e-Justice, Solvit, TED, EURES, ODR и др.)

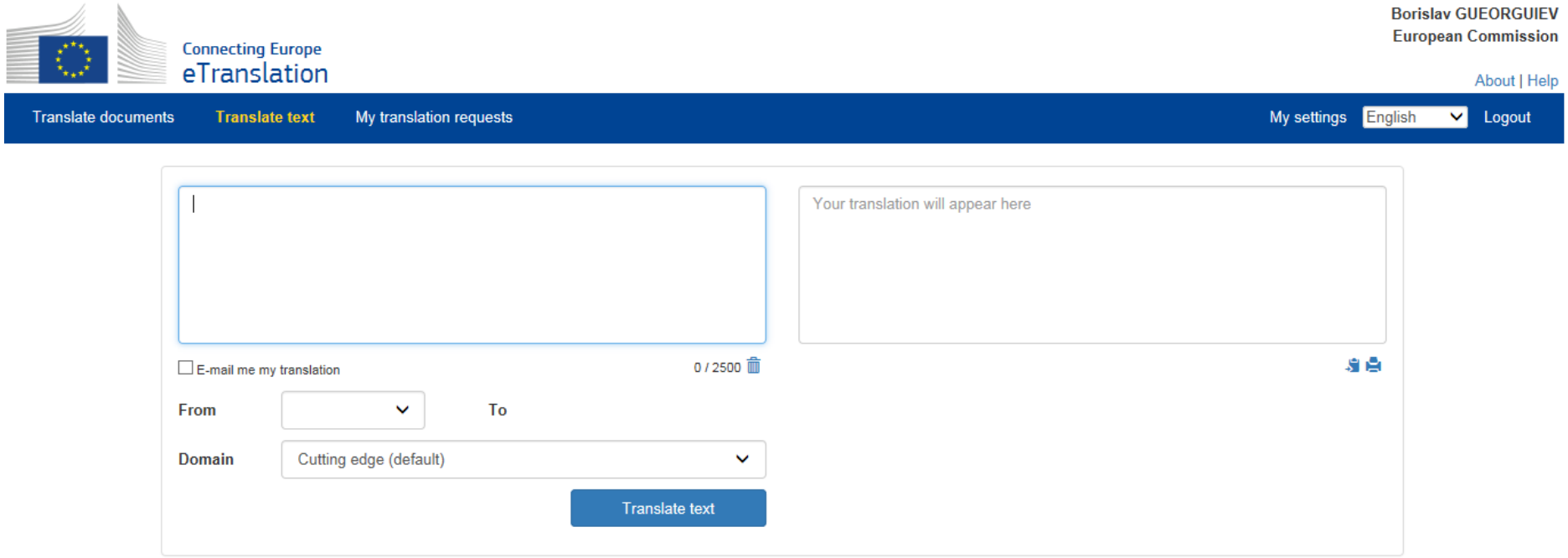
Превод на документ (невромашинен превод)



Превод на документ (статистически превод)



Превод на текст



Невромашинен превод

Невронната мрежа в две изречения

Невронната мрежа е компютърна програма, правеща предвиждания въз основа на моделите, които открива в данните, без да е била специално програмирана за тази цел.

С един слой "неврони" могат да се откриват само прости модели, а с повече от един — модели на моделите.

Основна архитектура на невромашинния превод

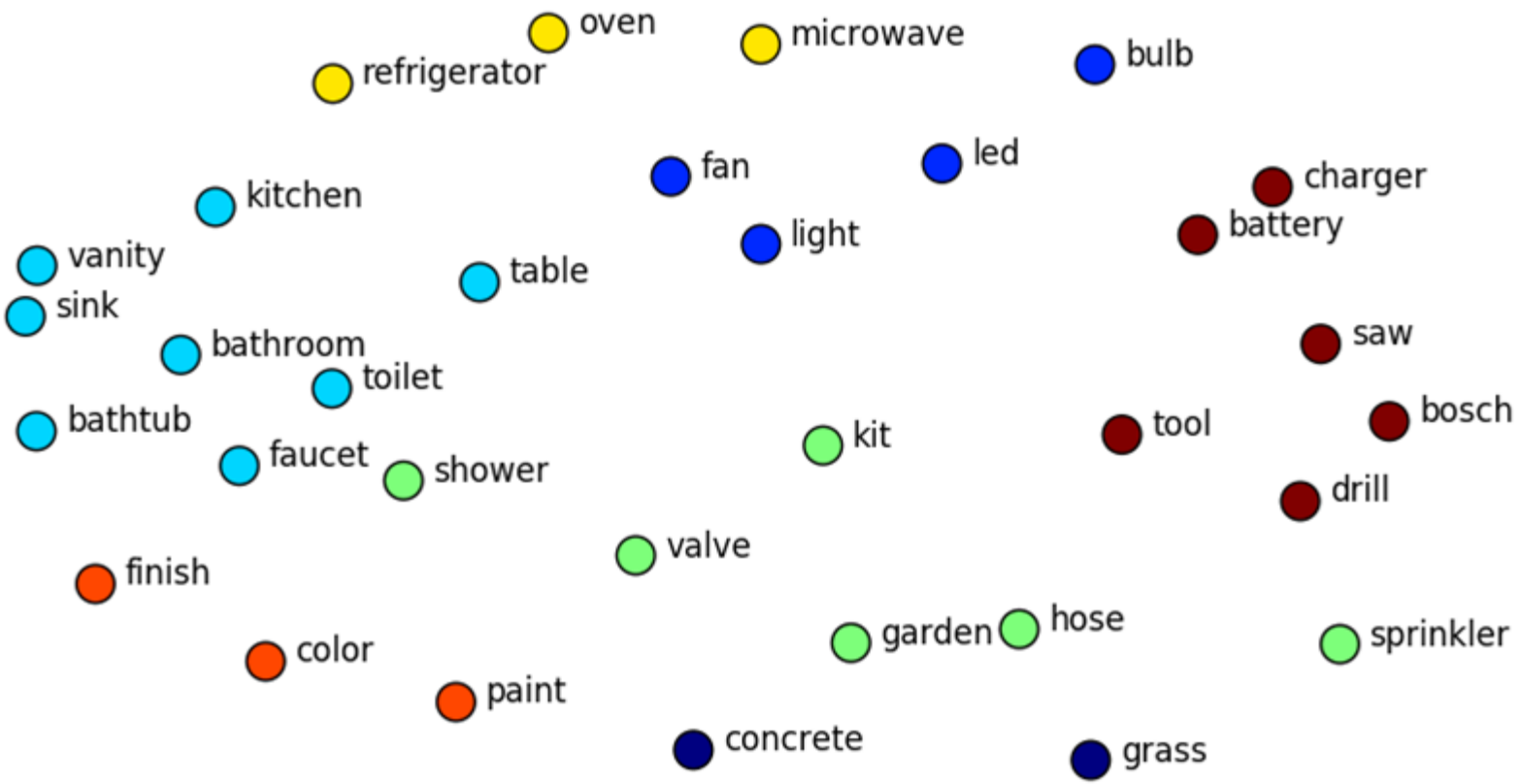
- Основава се на кодиране и декодиране
- При кодирането входящото изречение се представя под формата на вектор
- При декодирането се получава изходящото изречение, което се изгражда дума по дума въз основа на:
  - входящият вектор
  - предишната дума в изходящото изречение
  - цялото изходящо изречение до момента

Кодиране на думите (1)

Отделните думи също се кодират като вектори:

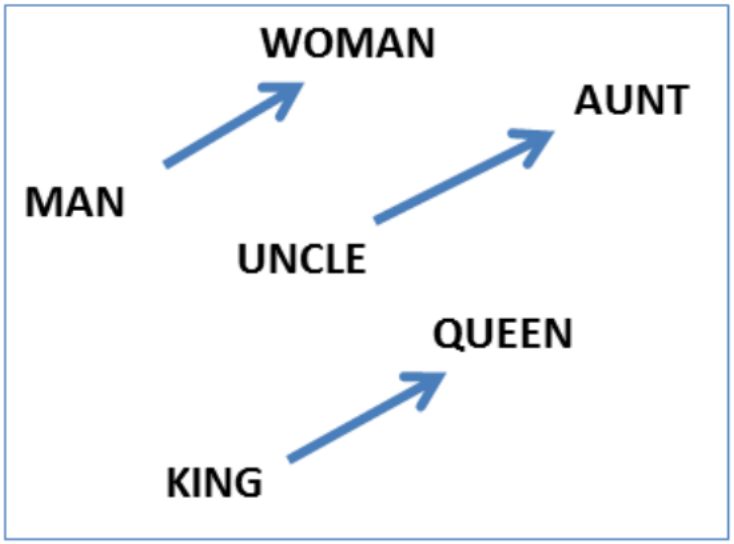
- “mouse” и “mice” се представят по подобен начин, тъй като се използват в подобен контекст
- представянето се основава на значението, а не на изписването.

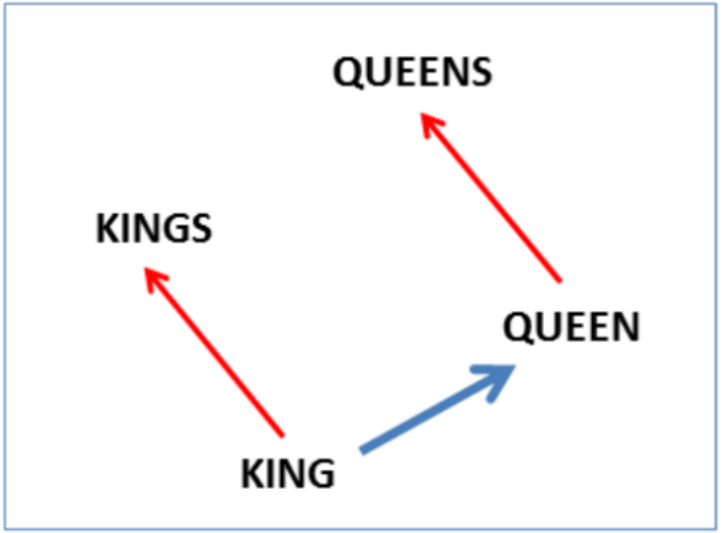
Кодиране на думите (2)



Семантично различни думи се представят по сходен начин.

Кодиране на думите (3)





Понякога векторите улавят връзките между думите (мъжки и женски род или единствено и множествено число).

Кодиране на думите (4)

king – man + woman = queen

Berlin – Germany + Japan = Tokyo

Byte pair encoding (1)

- *Някои думи не се сегментират:*
  - **it, this, the**
- *Други се разделят на най-често срещаните си подчасти:*
  - **con-tract, servi-ces, author-ities**
- *В най-лошия случай думата се разделя на отделни знаци:* ***Szymon –> s-z-y-m-o-n***
- *Така получените думи се компресират допълнително, за да се намали размерът на данните*

Byte pair encoding (2)

- *Помага за намаляване на изчислителните разходи за превода*
- *Улеснява обучението на невронната мрежа*
  - ***когато се намали разнообразието на входящите данни, е по-лесно да се намират модели***
- *По-лесно се вземат предвид лингвистичните феномени*
  - ***граматичните окончания са отделни подчасти и се обработват отделно***

Наблюдения и примери

Положителни страни на НМП (1)

В сравнение със СМП при НМП се наблюдава:

- ***по-добър словоред и синтаксис***
- ***по-добро съгласуване по род и число и време***
- ***по-добро членуване, включително по-правилно използване на пълен член***
- ***малко по-добра пунктуация***

Положителни страни на НМП (2)

Предимствата на НМП спрямо СМП са:

- ***пише се по-малко, тъй като почти не се налагат промени в окончанията на думите***
- ***при добър оригинал дава явно по-добри предложения***
- ***НМП превежда дори сгрешени в оригинала думи:***

Structural Reform Support **Por**gramme — **Програмата** за подкрепа на структурните реформи

destechnologies digitales — цифровите **технологии**

Примери за сполучлив превод

EGNOS is an important enabler for the modernisation of ATM in Europe.

EGNOS играе важна роля за модернизирането на УВД в Европа.

The central authority may also transmit information from its national authorities to another Member State via IMI (Article 3).

Централният орган може също така да предава информация от своите национални органи на друга държава членка чрез IMI (член 3).

Слабости на НМП

- *"Обръщане" на превода*
- *Терминологична непоследователност*
- *Изпускане и добавяне на думи*
- *Заместване на думи с подобни на тях*
- *Създаване на "нови" думи*
- *Оставяне на интервал преди точка или запетая*

Като цяло НМП се отличава с голяма изобретателност!

Примери (1)

- ***"Обръщане" на превода:***
  - *The **obsolete** definition – **неостарялото** определение*
  - ***Fees are not charged** by the Ministry of Finance – Министерството на финансите **не заплаща такси за***
  - *They are identified either **by** the project or institution which provided the data or **by** topic. - Те се идентифицират или **от** проекта, или **от** институцията, която е предоставила данните, или **от** темата.*
- ***Терминологична непоследователност:***
  - ***ENP South** country – държава от **ЕПС от страна на юг**, държава от **южното съседство на ЕПС**, държава **от Южното измерение на ЕПС***

Примери (2)

- ***Изпускане и добавяне на думи:***

  - Cooperation on satellite navigation between **European Neighbourhood Partnership South countries** and EU... **–** Сътрудничеството в областта на спътниковата навигация между **държавите от Европейското партньорство за съседство** и държавите — членки на ЕС...

  - … but use the statistical machine translation technology used earlier in **MT@EC. –** … но използва технологията за статистически машинен превод, която се използва по-рано на **електронен адрес**: MT@EC.

  - discussed with **ENP South countries** representatives and stakeholders in several workshops – представители и заинтересовани страни от държавите от Южна **Африка**, участващи в ЕПС, в няколко семинара

Примери (3)

- ***Изпускане и добавяне на думи (продължение):***
  - **Public** Revenue Collection Code – Кодекс за събиране на постъпления (липсва думата "**публични**")
  - District **Civil** Court judge – съдия от районния съд (липсва думата "**граждански**")
  - **standard** form – формуляр (липсва думата "**стандартизиран**")
  - place where there are **attachable** assets – всяко място, където има имущество (липсва думата "**секвестируемо**")
  - transmission – изпращане **на информация** (в случая верният превод е "**предаване**")

Примери (4)

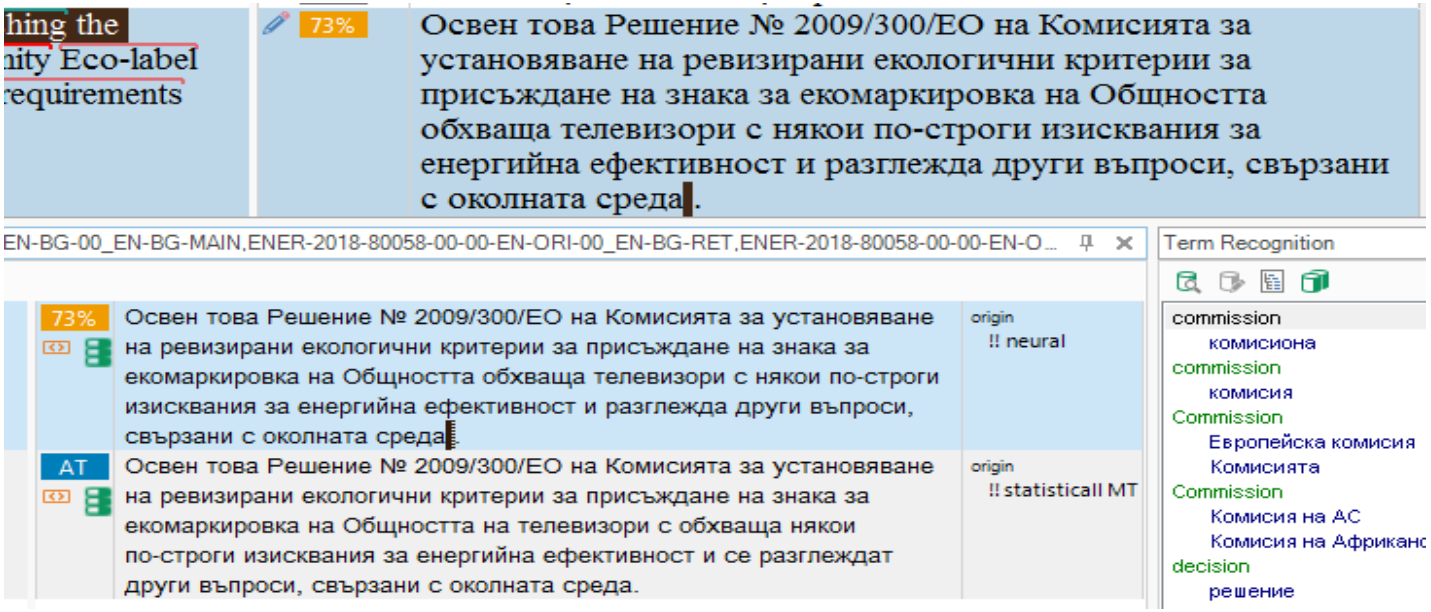
- ***Заместване на думи с подобни на тях:***

  - Cooperation on satellite navigation between European Neighbourhood Partnership South countries and **EU** is framed in the EuroMed transport cooperation. – Сътрудничеството в областта на спътниковата навигация между държавите от Европейското партньорство за съседство и **държавите — членки на ЕС**, е поставено в рамките на Евро-средиземноморското сътрудничество в областта на транспорта.

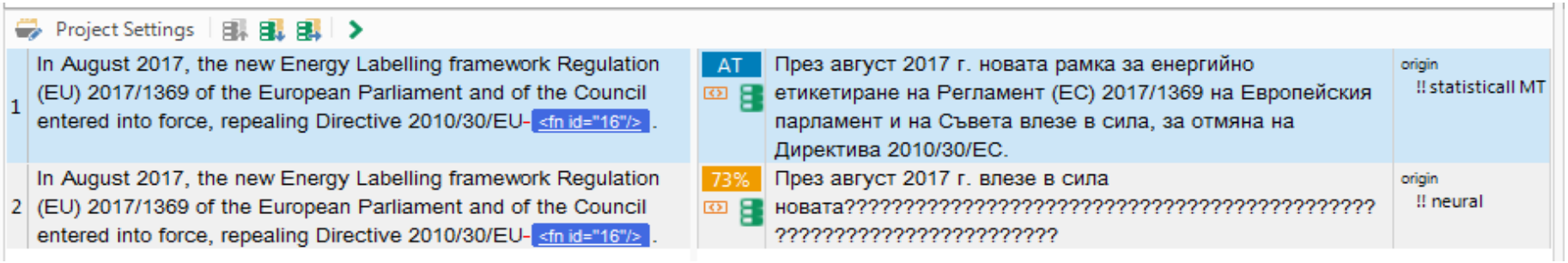
  - The **Ombudsman** strongly believes that, to earn the public's full trust, EU institutions must be citizen-friendly, accountable, transparent, and service-oriented. – **Emily O'Reilly** jest głęboko przekonana, że aby zdobyć pełne zaufanie społeczeństwa, instytucje UE muszą być przyjazne obywatelom i odpowiedzialne oraz działać w sposób przejrzysty i ukierunkowany na usługi.

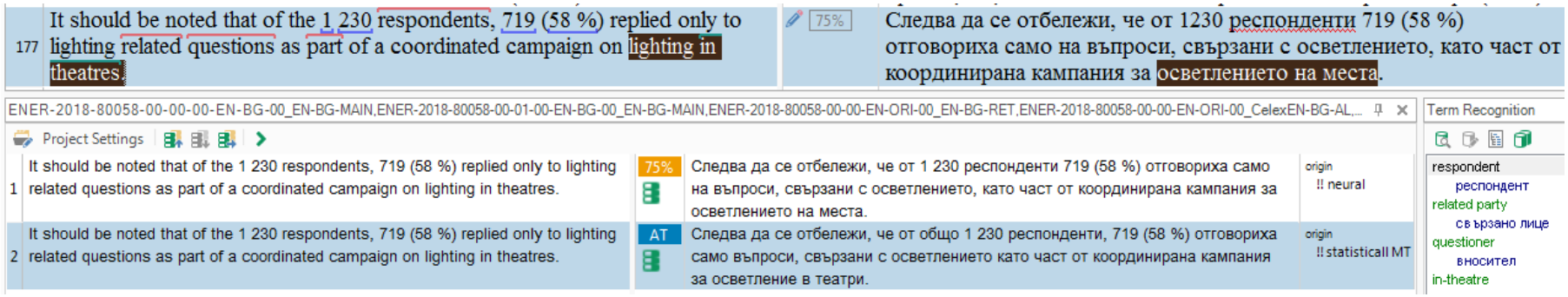
Примери (5)

- ***Създаване на "нови" думи:***
  - … available also in **fablabs** or **citylabs**. – … които са на разположение и в **abestabs** или **сиулаборатории**.
- ***Оставяне на интервал преди точка или запетая:***

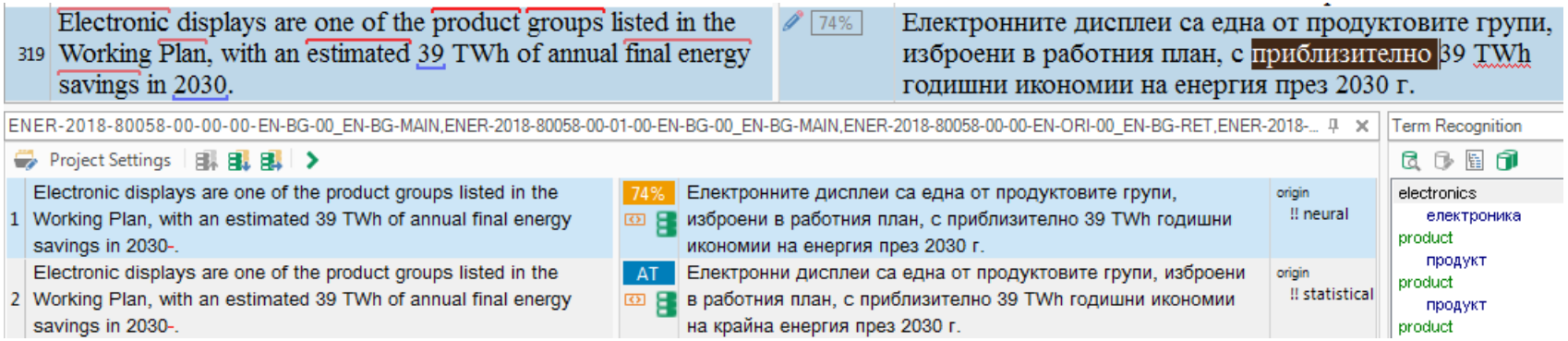


Примери (6)


Project Settings
1
In August 2017, the new Energy Labelling framework Regulation (EU) 2017/1369 of the European Parliament and of the Council entered into force, repealing Directive 2010/30/EU- <fn id="16"/> .
AT
През август 2017 г. новата рамка за енергийно етикетиране на Регламент (ЕС) 2017/1369 на Европейския парламент и на Съвета влезе в сила, за отмяна на Директива 2010/30/ЕС.
origin
!! statisticall MT
2
In August 2017, the new Energy Labelling framework Regulation (EU) 2017/1369 of the European Parliament and of the Council entered into force, repealing Directive 2010/30/EU- <fn id="16"/> .
73%
През август 2017 г. влезе в сила новата??? ????????????????????????
origin
!! neural


177
It should be noted that of the 1 230 respondents, 719 (58 %) replied only to lighting related questions as part of a coordinated campaign on lighting in theatres.
75%
Следва да се отбележи, че от 1230 респонденти 719 (58 %) отговориха само на въпроси, свързани с осветлението, като част от координирана кампания за осветлението на места.
ENER-2018-80058-00-00-00-EN-BG-00_EN-BG-MAIN,ENER-2018-80058-00-01-00-EN-BG-00_EN-BG-MAIN,ENER-2018-80058-00-00-EN-ORI-00_EN-BG-RET,ENER-2018-80058-00-00-EN-ORI-00_CelexEN-BG-AL,...
Term Recognition
Project Settings
1
It should be noted that of the 1 230 respondents, 719 (58 %) replied only to lighting related questions as part of a coordinated campaign on lighting in theatres.
75%
Следва да се отбележи, че от 1 230 респонденти 719 (58 %) отговориха само на въпроси, свързани с осветлението, като част от координирана кампания за осветлението на места.
origin
!! neural
2
It should be noted that of the 1 230 respondents, 719 (58 %) replied only to lighting related questions as part of a coordinated campaign on lighting in theatres.
AT
Следва да се отбележи, че от общо 1 230 респонденти, 719 (58 %) отговориха само въпроси, свързани с осветлението като част от координирана кампания за осветление в театри.
origin
!! statisticall MT
respondent
респондент
related party
свързано лице
questioner
вносител
in-theatre

Примери (7)



Sub-Directorate General – **Под**генерална дирекция (вярно е Генерална **под**дирекция)

Commercial Court – **окръжен** съд (вярно е **търговски** съд)

Статистически и невромашинен превод (1)

Original	SMT	NMT
Ignoring or weakening one of the pillars will undermine the whole construction as they are closely interlinked and interdependent:	Пренебрегване или отслабва един от стълбовете ще подкопае целия строителството, тъй като те са тясно взаимосвързани и взаимозависими:	Пренебрегването или отслабването на един от стълбовете ще подкопае цялостната конструкция, тъй като те са тясно свързани помежду си и взаимно зависими:
The Digital Innovation Hubs will serve as access points to latest digital capacities including high performance computing (HPC), artificial intelligence, cybersecurity, as well as other existing innovative technologies such as Key Enabling Technologies, available also in fablabs or citylabs.	На центровете за цифрови иновации ще служат като звена за връзка с най-новите цифрови възможности, включително в областта на високопроизводителните изчислителни технологии (ВИТ), изкуствения интелект, киберсигурността, както и на други съществуващи иновативни технологии, като например ключовите технологии, също така и в fablabs или citylabs.	Центровете за цифрови иновации ще служат за достъп до последните цифрови технологии, включително високопроизводителните изчислителни технологии (HPC), изкуствения интелект, киберсигурността, както и други съществуващи иновативни технологии, като например главните базови технологии, които са на разположение и в abestabs или сиулаборатории.

Статистически и невромашинен превод (2)

Cybersecurity technologies such as digital identities, cryptography or intrusion detection, and their application in areas such as finance, industry 4.0, energy, transportation, health and care, or e-government are essential to safeguard the security and trust of online activity and transactions by both citizens, public administrations, and companies.	Технологии за киберсигурност, като например цифрови самоличности, криптография или проникване, както и прилагането им в области като финанси, промишленост 4.0, енергия, транспорт, здравеопазване и грижи, или електронното управление, са от първостепенно значение за гарантиране на сигурността и доверието на онлайн дейности и сделки, както от гражданите, публичните администрации и предприятията.	Технологии за киберсигурност, като например цифрови самоличности, криптиране на криптографията или откриване на проникване, и тяхното прилагане в области като финансиране, промишленост 4.0, енергетика, транспорт, здравеопазване и грижи, или електронно управление са от съществено значение за гарантиране на сигурността и доверието на онлайн дейността и трансакциите както на гражданите, така и на публичните администрации и дружествата.
all major car manufacturers are developing self-driving cars, and machine learning techniques are at the heart of all main web platforms and big data applications.	всички основни производители на коли разработват автомобили без шофьор, машинно обучение и методи са в центъра на всички основни уеб платформи и приложения за големи масиви от данни.	всички големи производители на автомобили разработват собствени автомобили, а техниките за машинно обучение са в основата на всички основни уеб платформи и приложения на големи информационни масиви.

Изводи

1. **Невромашинният превод (НМП), както и СМП, рядко може да се използва без редакция.**

2. **Предложенията на НМП са по-"гладки", но е по-вероятно да съдържат "скрити" смислови грешки.**

3. **НМП не се справя добре с оригинали с лошо качество.**

4. **В крайна сметка, всеки преводач преценява сам дали и как да го използва!**

Благодаря за вниманието!