MI Szakmai délután

OTP Bank

2023.09.14

Agenda

Érkezés 13:45-14:00

Előadás 14:00-15:30

Szünet 15:30-15:45

Kerekasztal beszélgetés 15:45-17:00

Bemutatkozás: miért foglalkozik egy Bank a nyelvi modellekkel?
Előadó: **Kaszás Zoltán** -OTP Bank Ügyvezető Igazgató

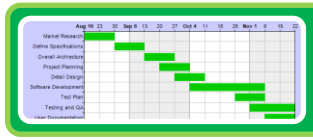
OTP-s NLP „Szuperszámítógép” projekt céljai, tagjai
Előadó: **Nagy Zsombor**; Szenior projektvezető

Nyelvi modellek: miért szükséges a Magyar NLP modell?
Előadó: **OTP AI CoE**

OTP szuperszámítógép választásának szempontjai
Előadó: **Nagy Zsombor; OTP AI CoE**

OTP szuperszámítógép építés lépései
Előadó: **Nagy Zsombor**

Alkalmazott biztonsági megoldások
Előadó**: Oláh István György** IT kockázatkezelési vezető Tanácsadó; Tanár



Kutatási terv és jelen állapot
Előadó: **Nagy Zsombor**

Mi várható eredményként
Előadó: **OTP AI CoE**

Miért foglalkozik egy Bank a nyelvi modellekkel

Kiszolgáltatottság mérséklése, képesség építése, párhuzamos kutatások összehangolása a célok között szerepel

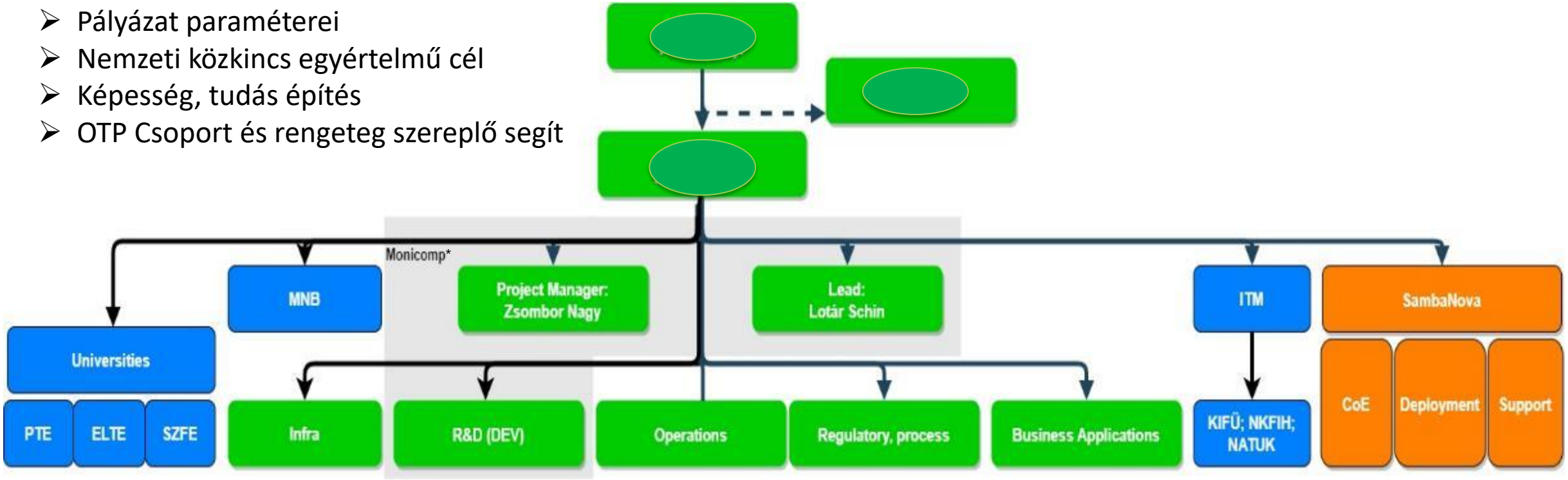
A **kutatási projekt átfogó célja** a digitális képességek fejlesztése a mesterséges intelligencia segítségével, a természetes és a **gépi nyelv közötti átjárás megteremtése és hogy létrejöjjön a mesterséges intelligencia-algoritmusok számára használható magyar nyelvi modell.** Ezzel a OTP csoport óriási lépést tesz a technológiában betöltött vezető szerep felé, egyúttal felkarolja a hazai mesterséges intelligencia-kutatás és -fejlesztés fellendítésének ügyét és elősegíti a magyar nyelvű mesterséges intelligencia alapú megoldások létrehozását is.

A magyar beszélt nyelv felhasználói köre relatíve szűk, ezért a globális piac nem érdekelt jó minőségű magyar nyelvfeldolgozás piaci alapú előállításában, emiatt a fejlesztés a globális kiszolgáltatottság mérséklésére irányul, mind a magyar nyelvfeldolgozás, mind az adatkezelés és a platform technológiák használata területén. A fejlesztés arra a hiányterületre épül, hogy az MI kutatásánál, fejlesztésénél a világnyelvekre, különösen az angolra fókuszálnak. Ez viszont az olyan kevés ember által beszélt nyelveknél, mint amilyen a magyar is, nagy hátrányt jelent bármilyen szövegértésre, hangfeldolgozásra épülő alkalmazás lokalizálásánál. Annak érdekében, hogy a nemzetközi vívmányok kisebb nyelvi piacokon, így Magyarországokon is alkalmazhatóak legyenek, szükséges a mesterséges intelligencia alapú nyelvi megoldások alkalmazása.

Jelen kutatás egy kellően nagy HPC kapacitással (56 petaflops~ 1 milló laptop) rendelkező szolgáltatás segítségével lendületet szeretne ennek adni a hazai fejlesztésnek, és a hosszútávon a lehető legnagyobb hasznos méretű általános nyelvi modell előállítása a cél.

OTP-s NLP „Szuperszámítógép" projekt céljai
Projekt tagjai-Működési modell

- Pályázat paraméterei
- Nemzeti közkincs egyértelmű cél
- Képesség, tudás építés
- OTP Csoport és rengeteg szereplő segít



A projekt munkacsoport jellegű működéssel indul el, előre definiált, ám rugalmasan alakítható feladatokkal. A tevékenységek előre haladtával javasoljuk a működési modell rendszeres felülvizsgálatát, az igényekhez igazítását és a szükséges kompetencia mix finomhangolását.
ITM és SambaNova-n kívüli fontos külső kapcsolódások jelennek meg: az MNB-vel, a banki MI szabályozási környezet kialakítására és a magyar nyelvtudományi közösség, egyetemeken keresztül, a nyelvészeti kutatások felhasználására és a magyar nyelvi modell kialakítására, K+F tevékenység sikeres elvégzésére.

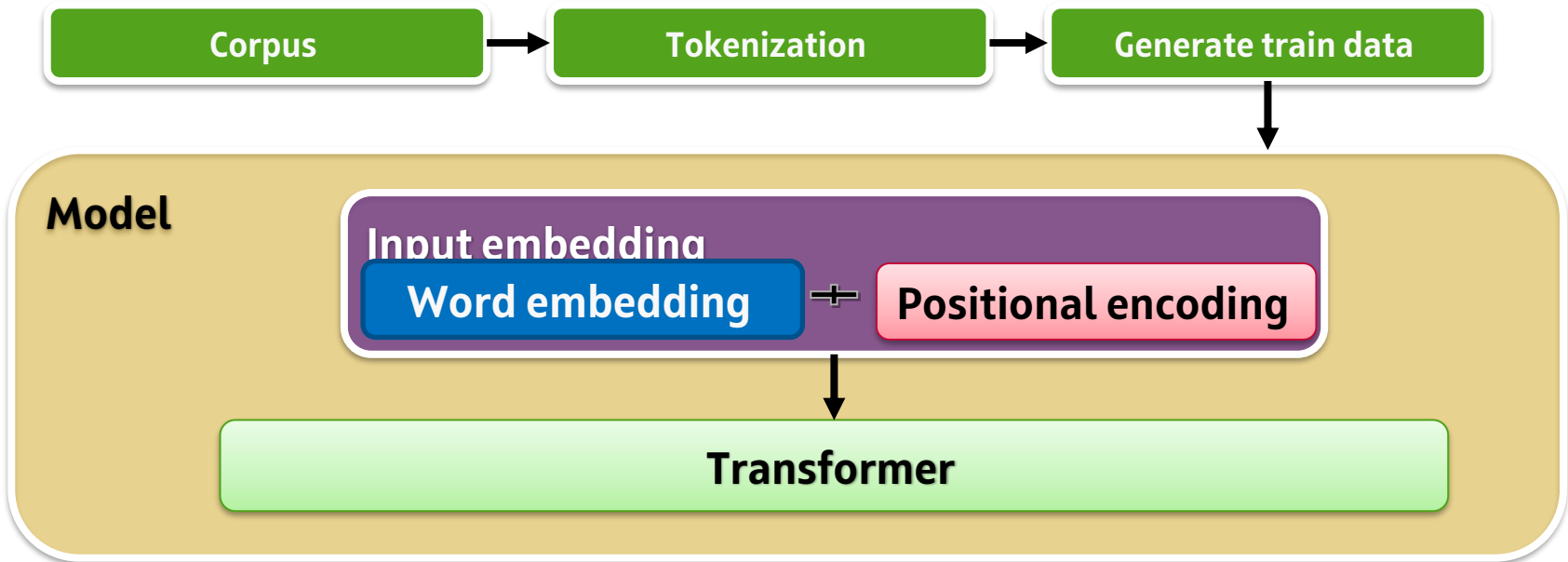
Nyelvi modellek: miért szükséges a Magyar NLP modell?

Miért jó?

Miért kell saját fejlesztésű magyar modell?

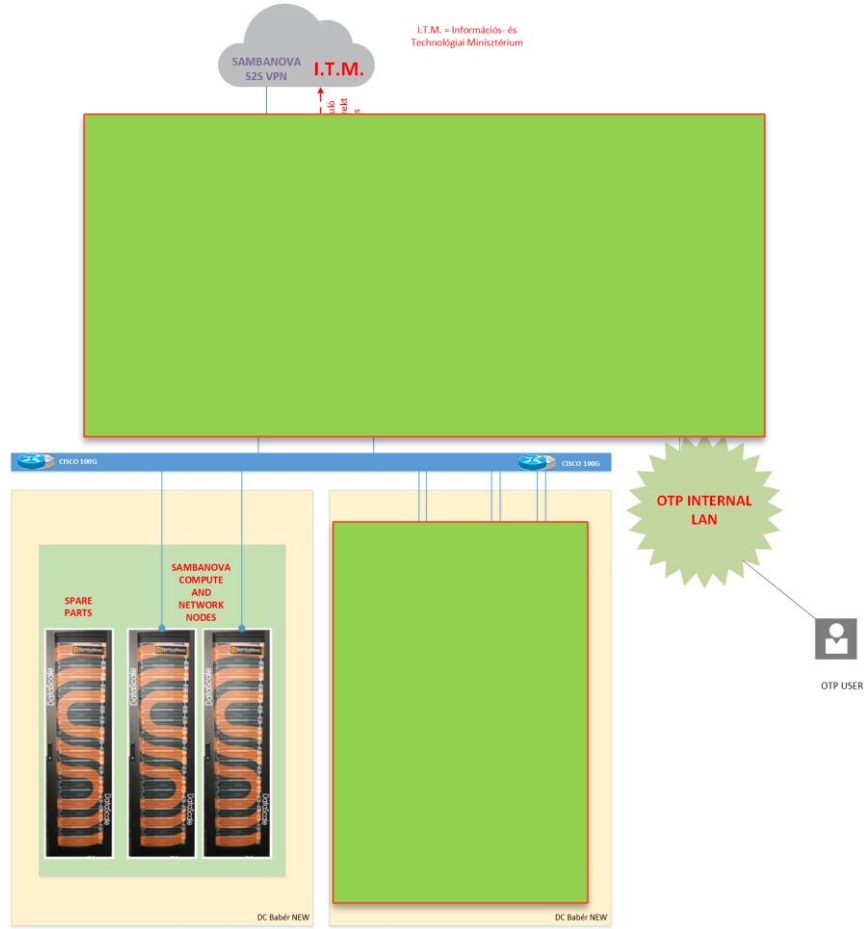
Miért lesz minőségibb?

Ez mégis hogyan néz ki?



OTP szuperszámítógép választásának szempontjai
OTP szuperszámítógép építés lépései

- Nem általnos célú HW-szolgáltatás
- Gyors implementáció volt a szempont
- Covid időben jött és volt HW
- Lehetett egyből dolgozni vele
- CoE Stanford támogatással



- HW szállítás
- 3 rack, 750 kg 1 DB
- Kevlár szőnyeg
- 15-21 KW/h felvétel- mégis hatékony

One SN10-8 system has 8 RDUs. Each RDU has > 300 BF16 TFLOPs. That means that one system has > 2,400 BF16 TFLOPs.
A full Dataflow-as-a-Service rack has 4 x SN10-8 systems. That means that a full rack has > 19,200 BF16 TFLOPs (19.2 PFLOPs)
3 racks has > 57,600 BF16 TFLOPs (57.6 PFLOPs)

Alkalmazott biztonsági megoldások

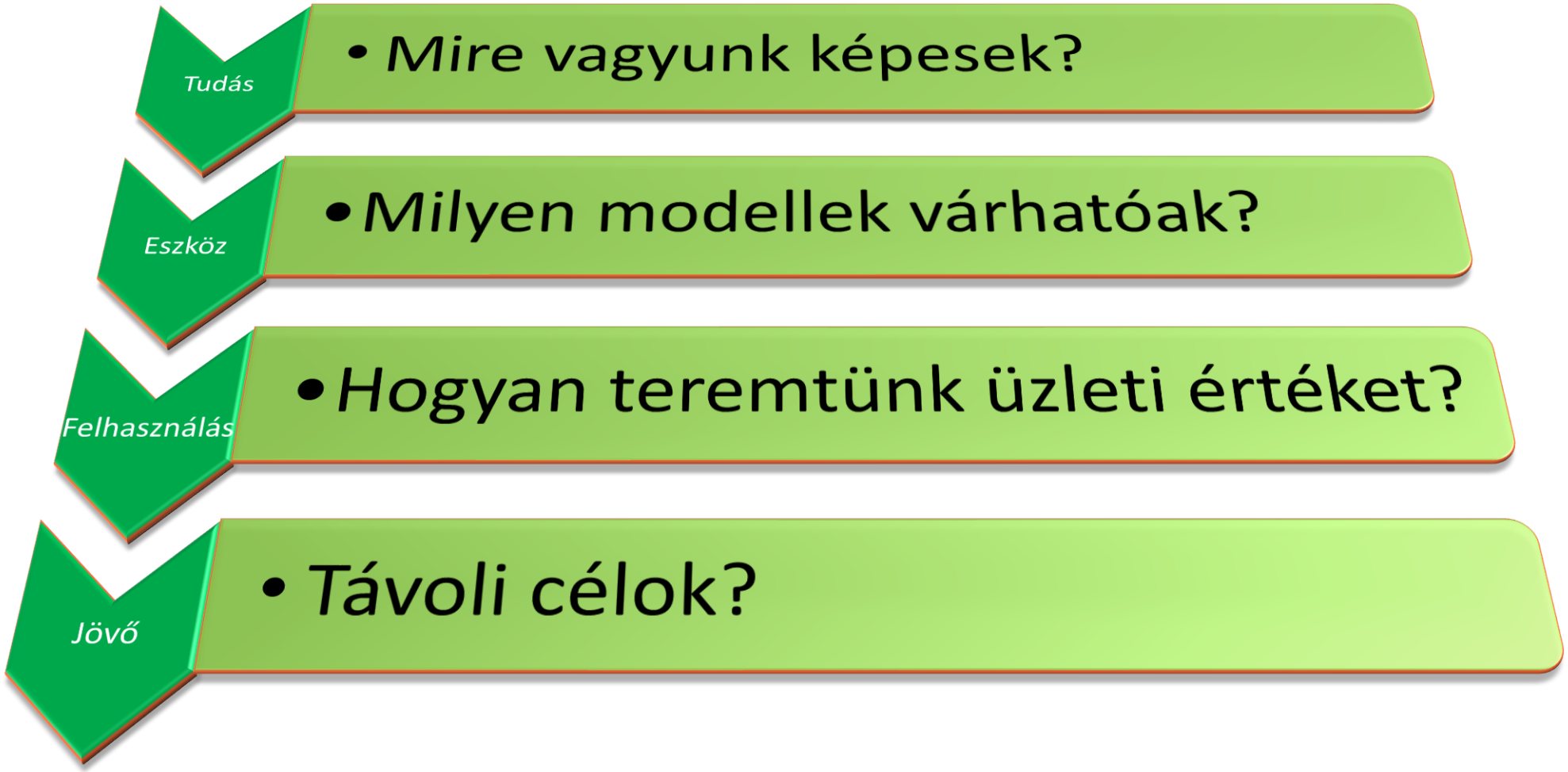
Ha banki rendszer lenne:

- Alkalmazás ID
- Rto, Rpo, SLA, BIA, Ibtv, MNB fogyasztóvédelem, …..,
- HLD, LLD
- Adatleltár, Adatok BSR , BRD osztály + BRD
- SIEM + SOC, IT monitoring, egyéb pl. speciális elemző eszközök
- BCP, DRP, MAVT
- Szabályzatok, auditok, felügyeleti ügyek, csoport ügyek
- ….

Nem banki rendszer:

- Terület, fizikai biztonság
- Energia, klíma, + felügyelet
- Adatkapcsolat
- Önálló védelem és IT monitoring és üzem
- Önálló AAA
- Adathordozó ? ekkor sem…

Mi várható eredményként



Függelék

Ez mégis hogyan néz ki?
Input embedding

Word embedding + **Positional encoding**

ügyfél →

→ $\begin{bmatrix}0.37\\0.99\\0.01\\0.08\end{bmatrix}$

Péter nem figyelt, és **elesett**. → Pozíció 5

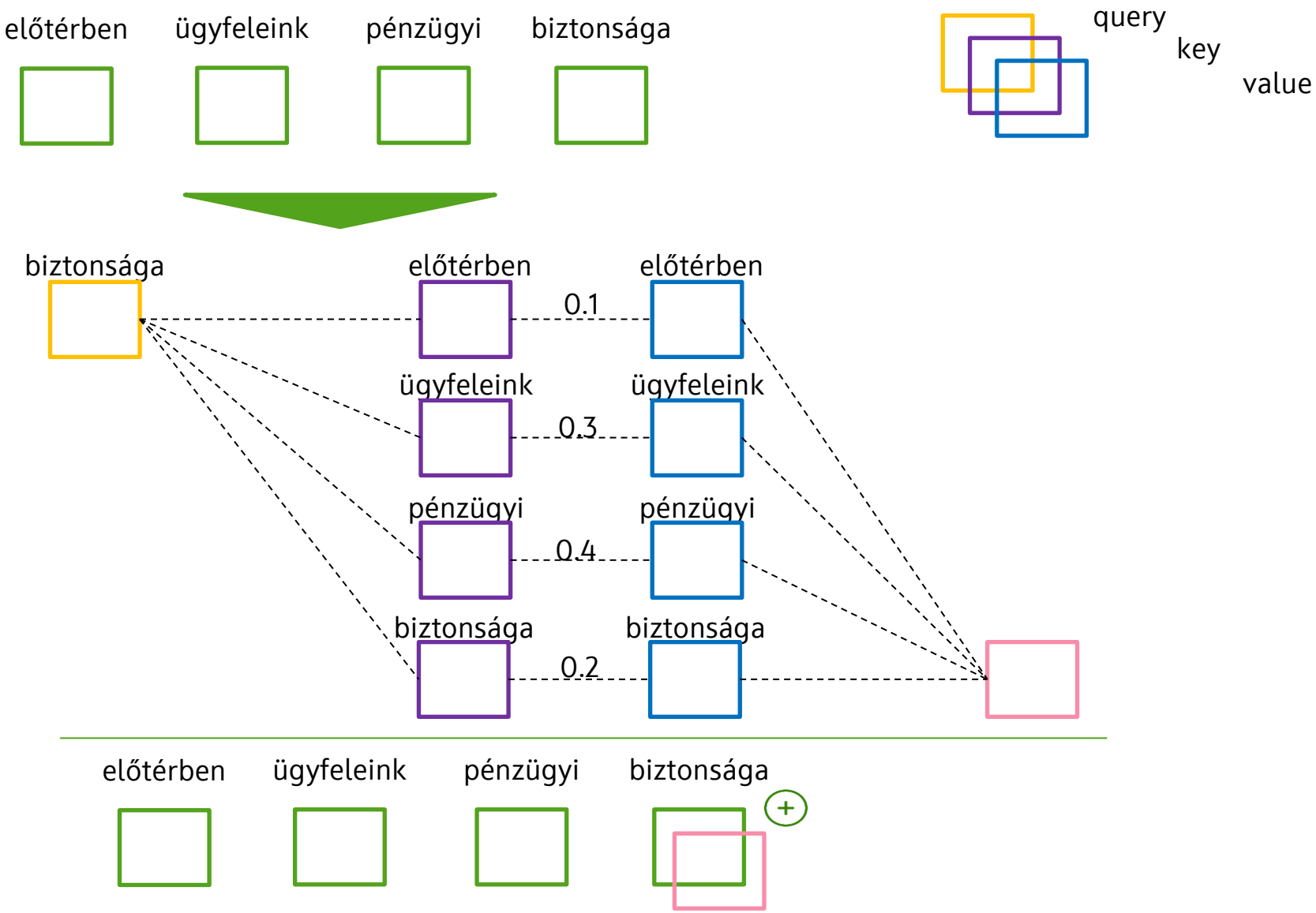
A király **elesett** a csatában. → Pozíció 3

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

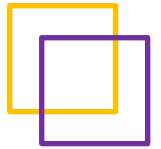
=

$\begin{bmatrix}0.42\\0.84\\0.12\\0.81\end{bmatrix}$ → Az ügyfél embedding-je, kontextussal

Ez mégis hogyan néz ki?
Attention



Ez mégis hogyan néz ki?
LM head



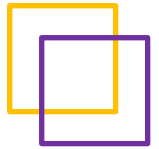
input

output

	előtérben	ügyfeleink	pénzügyi	biztonsága	a
előtérben	0.02	0.53	0.08	0.14	0.23
ügyfeleink	0.07	0.00	0.44	0.35	0.14
pénzügyi	0.09	0.21	0.03	0.51	0.16
biztonsága	0.14	0.25	0.11	0.01	0.49

	ügyfeleink	pénzügyi	biztonsága	a
ügyfeleink	0.00	1.00	0.00	0.00	0.00
pénzügyi	0.00	0.00	1.00	0.00	0.00
biztonsága	0.00	0.00	0.00	1.00	0.00
a	0.00	0.00	0.00	0.00	1.00

Ez mégis hogyan néz ki?
Classification head



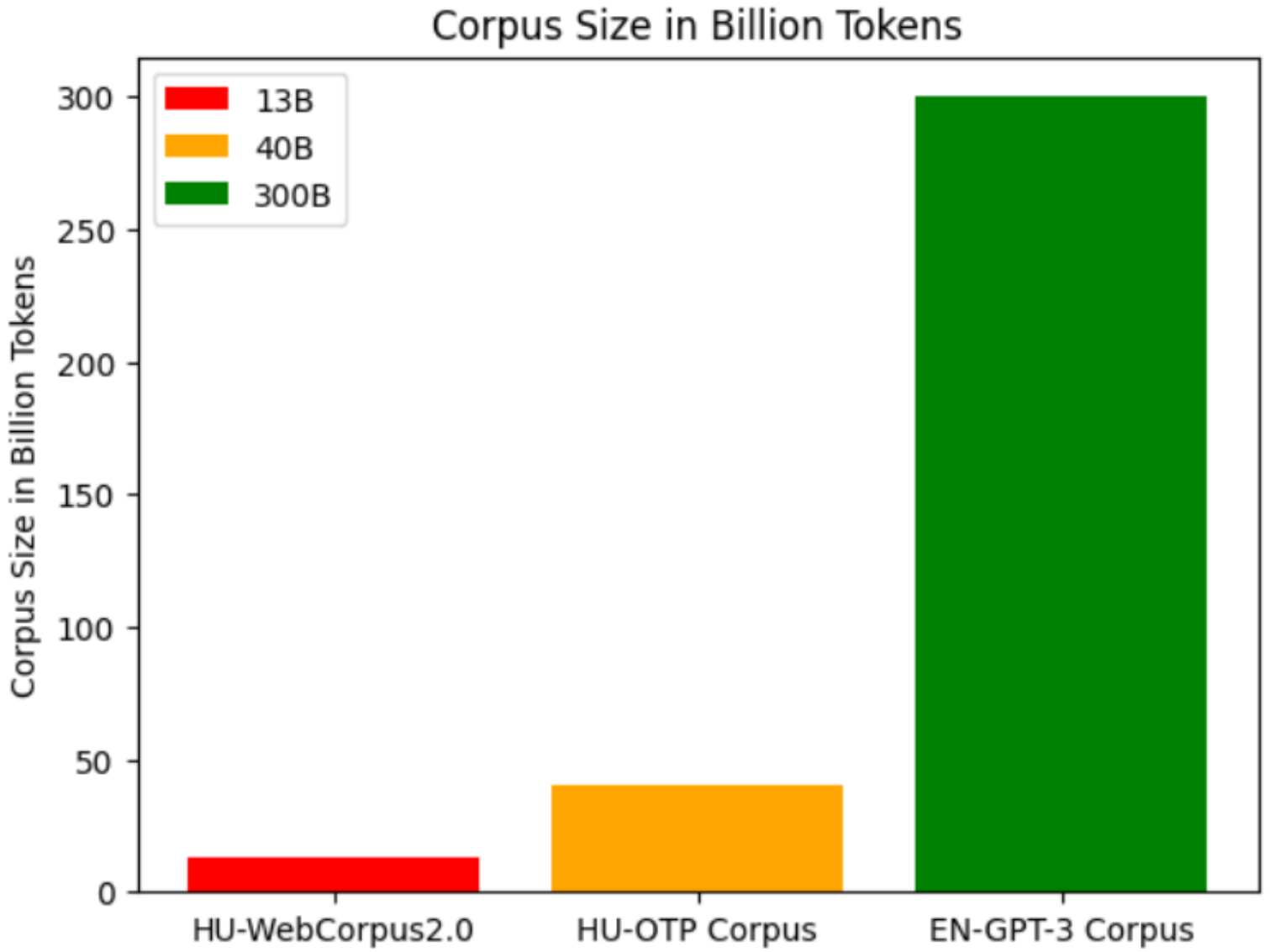
input

output

	pozitív	negatív
előtérben	0.96	0.04
ügyfeleink	0.65	0.35
pénzügyi	0.26	0.74
biztonsága	0.81	0.19
átlag:	**0.67**	**0.33**

1 **0**

Corpus


Corpus Size in Billion Tokens
13B
40B
300B
Corpus Size in Billion Tokens
300
250
200
150
100
50
0
HU-WebCorpus2.0
HU-OTP Corpus
EN-GPT-3 Corpus

Corpus <> Context Window

1.5B model 1024 token
13B model 2048 token

WebCorpus 2.0 has 13B token:

- 1.5B model **12 397 steps**
- 13B model **6 198 steps**

150k steps would be ideal!

How much data do we need?

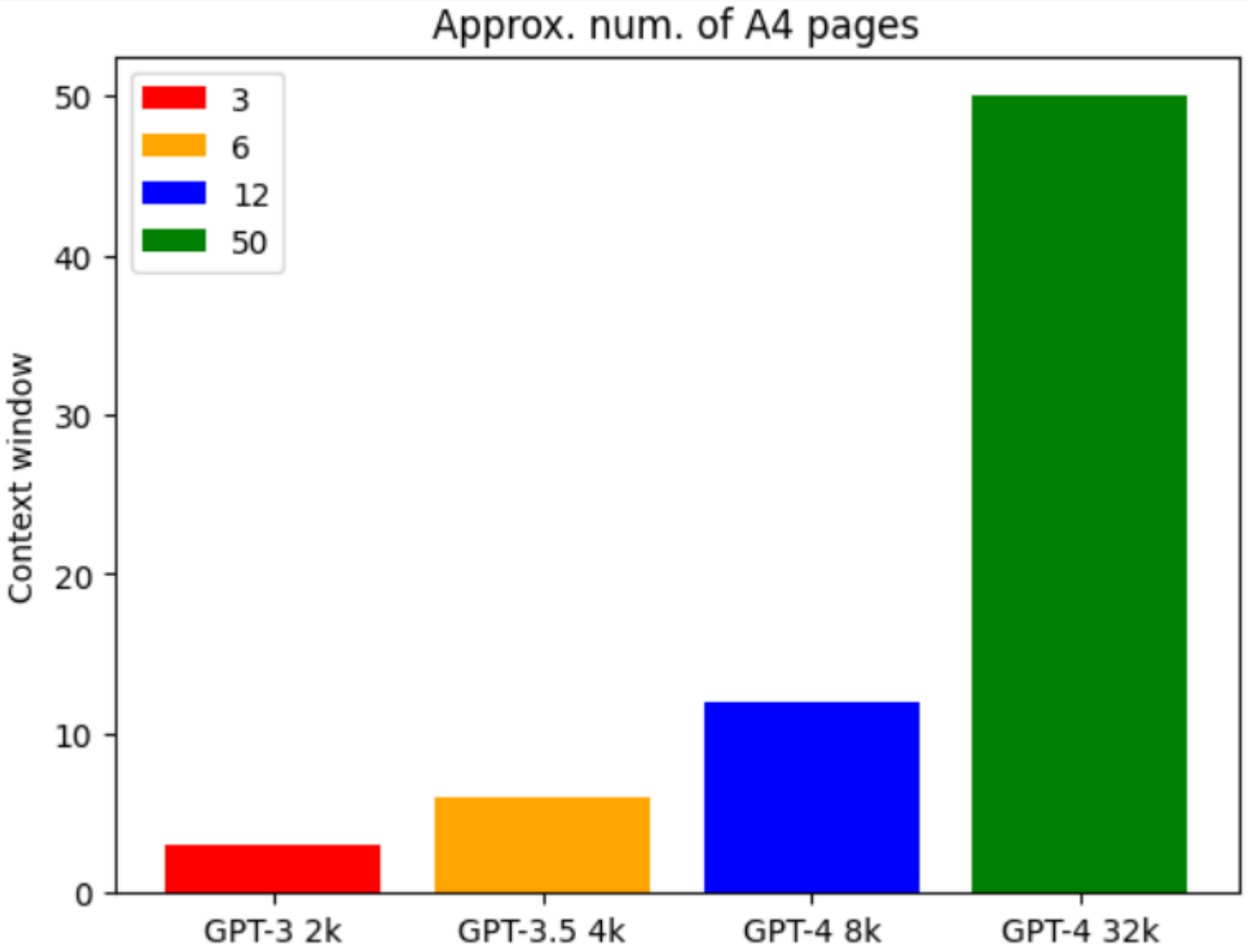
Sambanova’s experience:

- 13B model (300k steps) can overperform 175B model (150k steps)

300k steps -> 1024*300k*2048=**629B tokens!!!**

150k steps -> **314B tokens!!!**

Context window



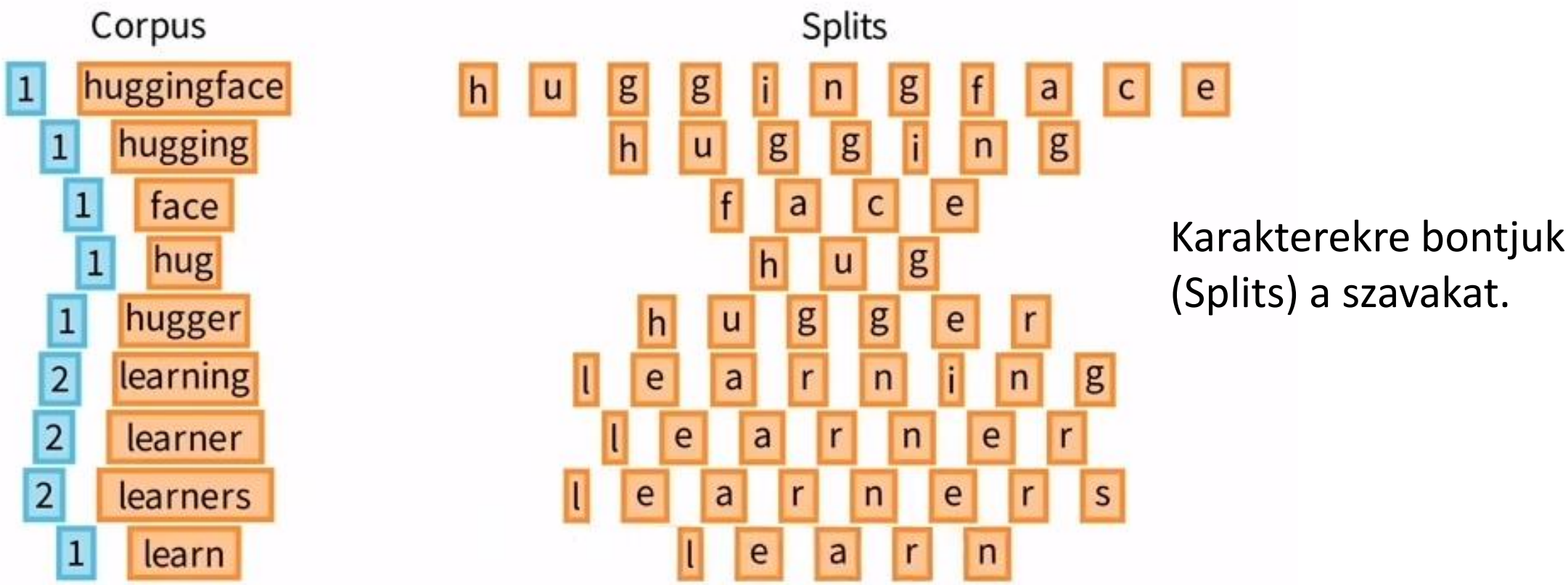
Hogyan készülünk a modell tanítására?

„Szótár” létrehozása, előre meghatározott mérettel

Képzeljünk el egy szöveges állományt (corpora), amiben a bal oldalon látható szavak szerepelnek. Kék színnel az előfordulásuk darabszámát jelöljük.

Hogyan készülünk a modell tanítására?

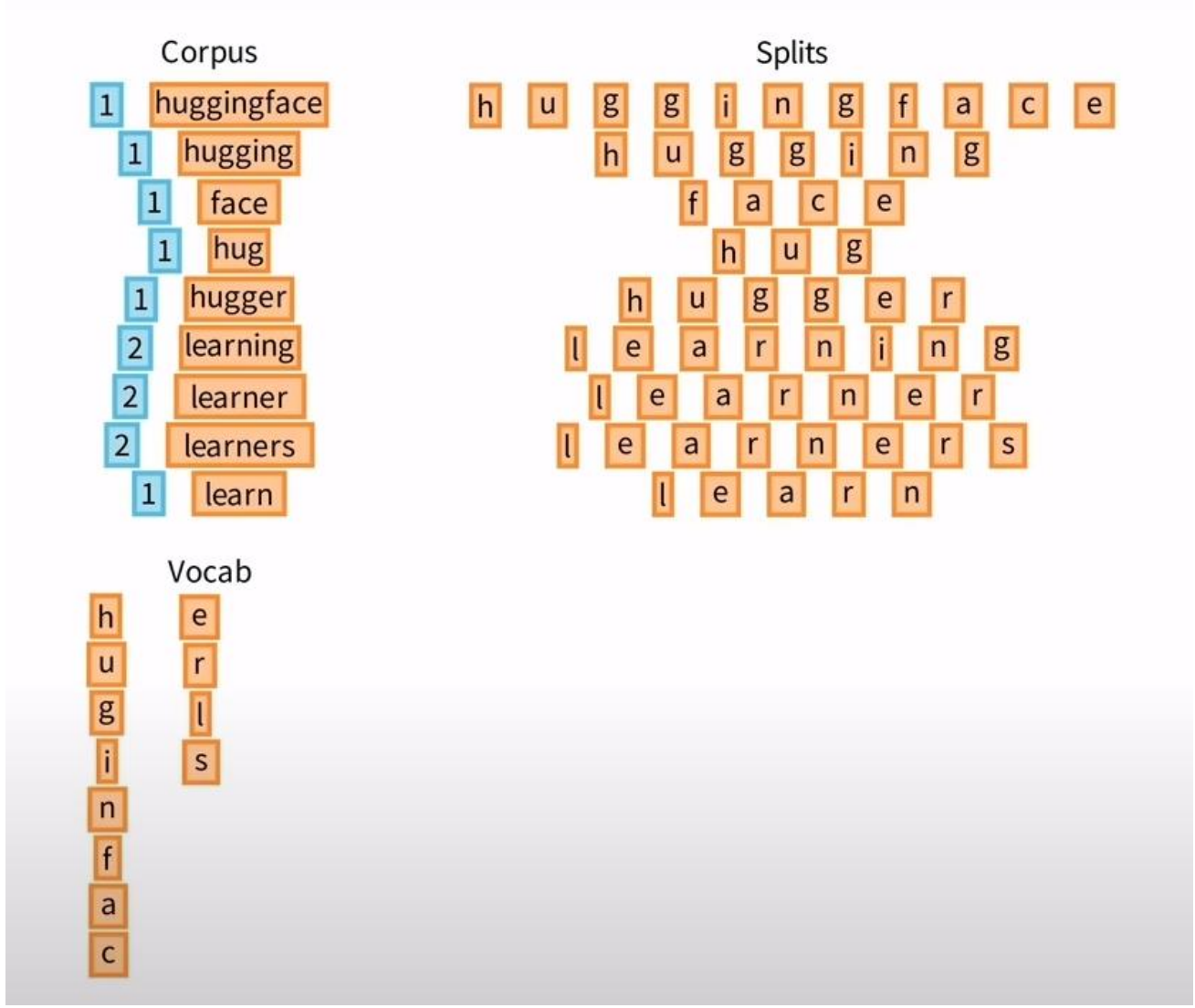
„Szótár” létrehozása, előre meghatározott mérettel



Karakterekre bontjuk (Splits) a szavakat.

Hogyan készülünk a modell tanítására?

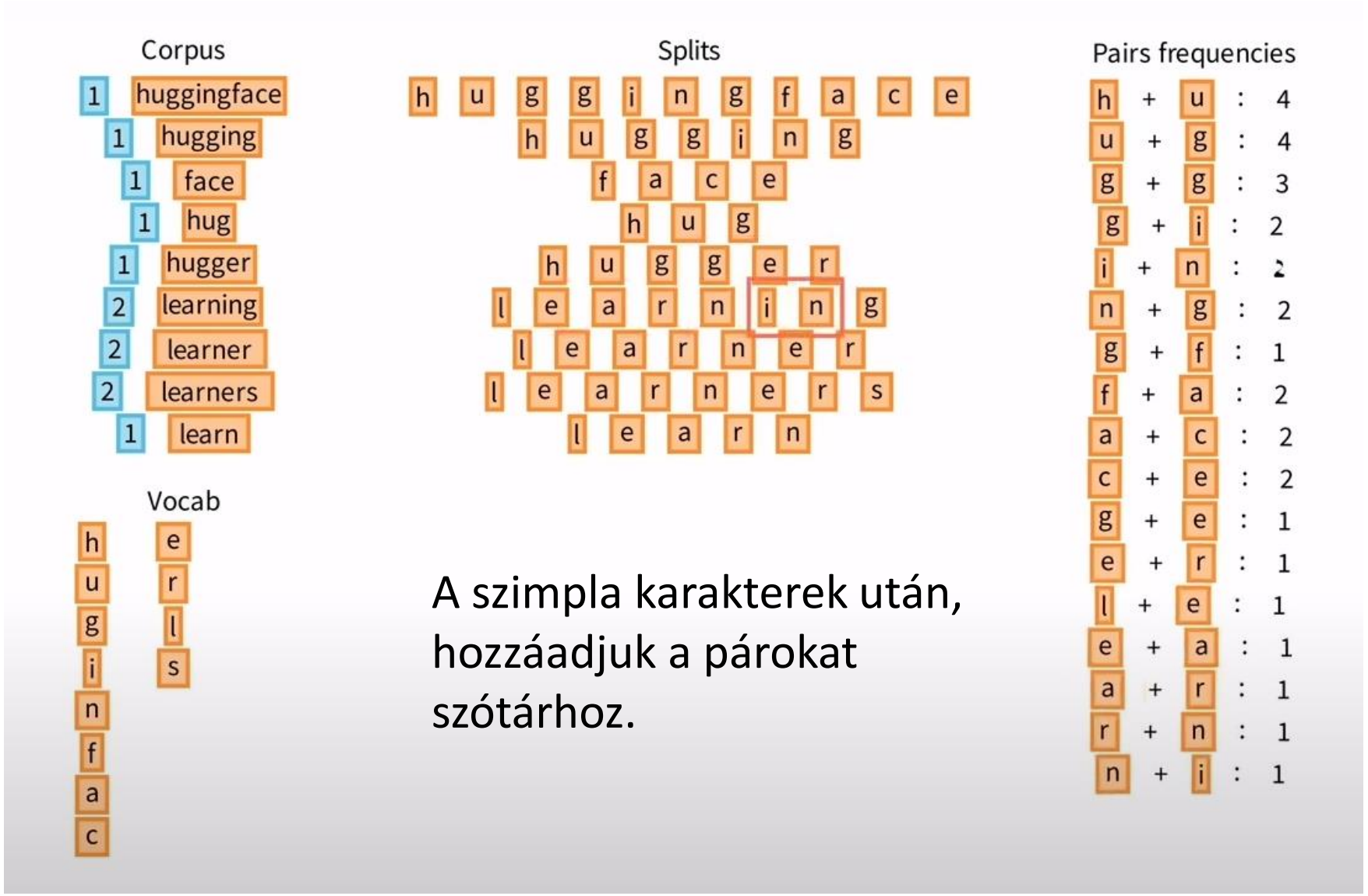
„Szótár” létrehozása, előre meghatározott mérettel



Az előforduló karaktereket felvesszük a szótárba.
Van még hely? (megadtuk, mekkora lehet a szótár)
Ha igen, akkor megyünk tovább…

Hogyan készülünk a modell tanítására?

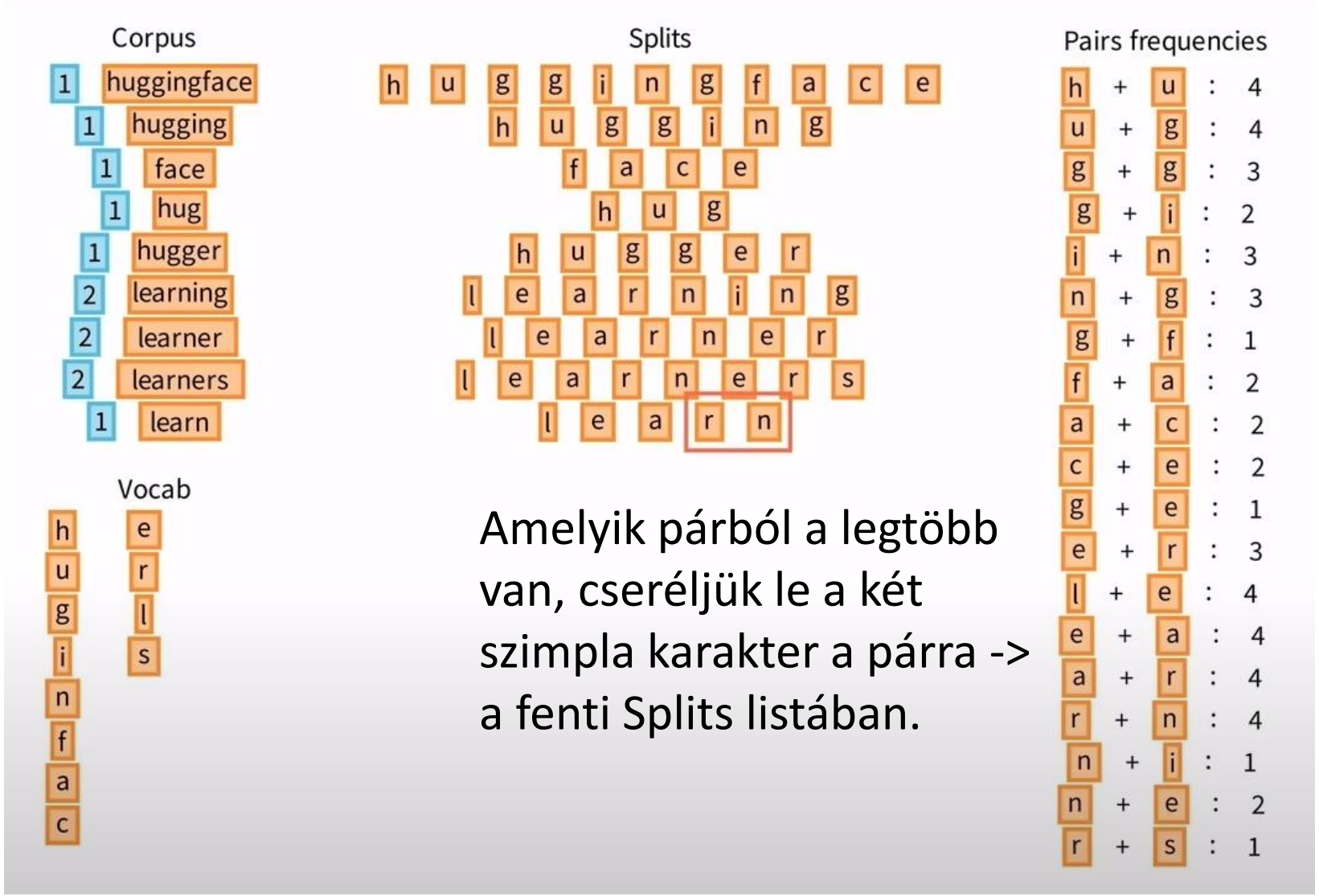
„Szótár” létrehozása, előre meghatározott mérettel



A szimpla karakterek után, hozzáadjuk a párokat szótárhoz.

Hogyan készülünk a modell tanítására?

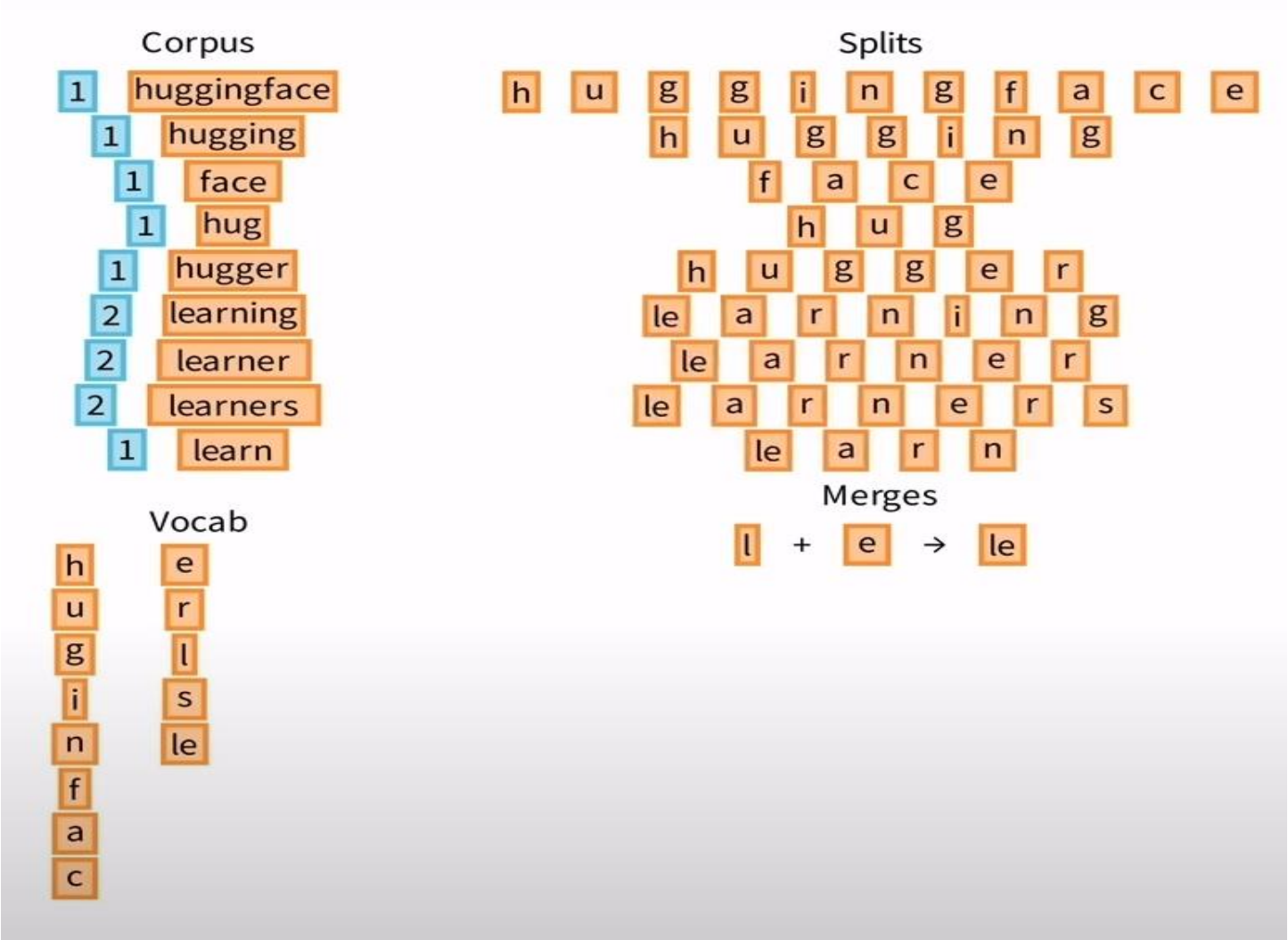
„Szótár” létrehozása, előre meghatározott mérettel



Amelyik párból a legtöbb van, cseréljük le a két szimpla karakter a párra -> a fenti Splits listában.

Hogyan készülünk a modell tanítására?

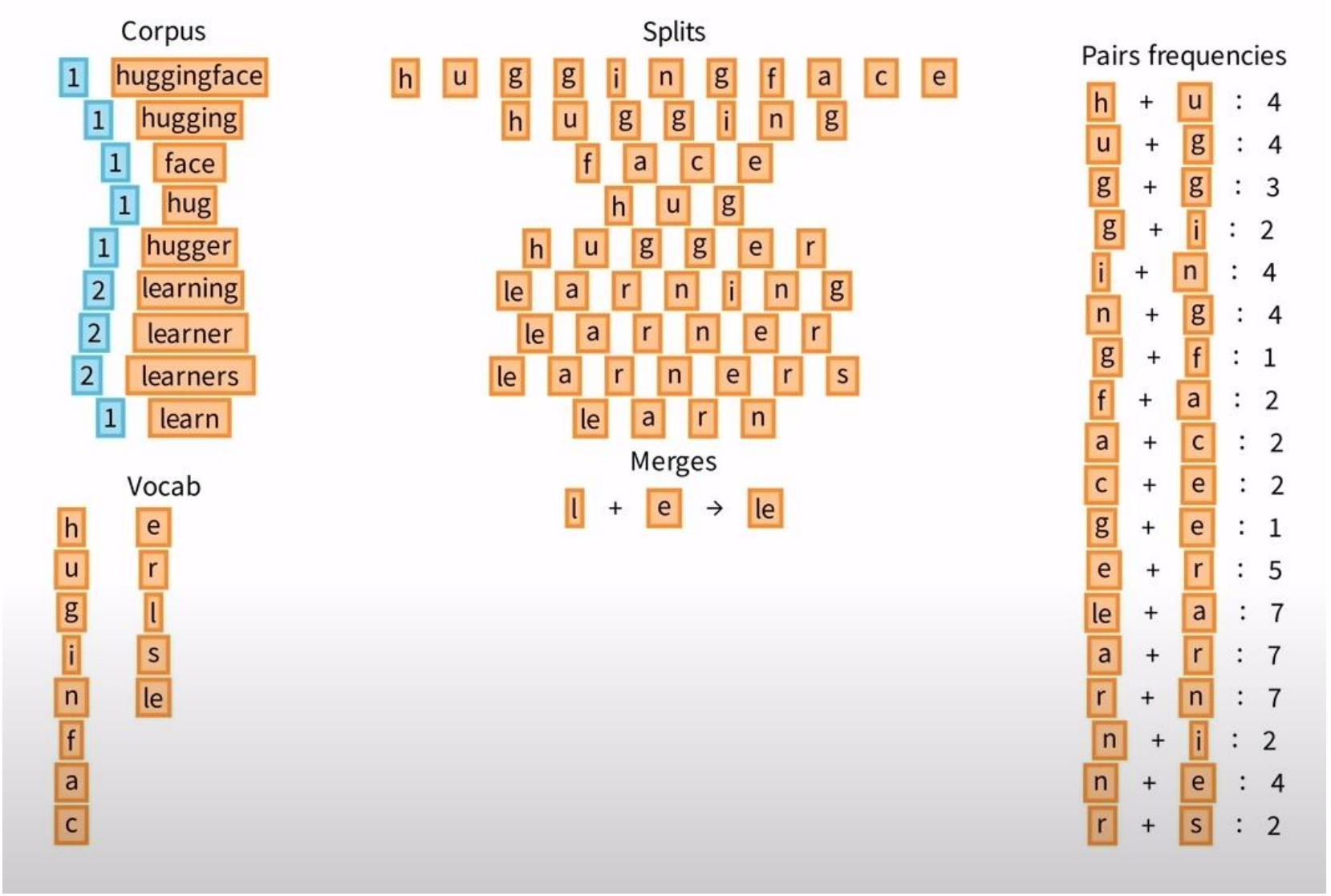
„Szótár” létrehozása, előre meghatározott mérettel



l e -> le

Hogyan készülünk a modell tanítására?

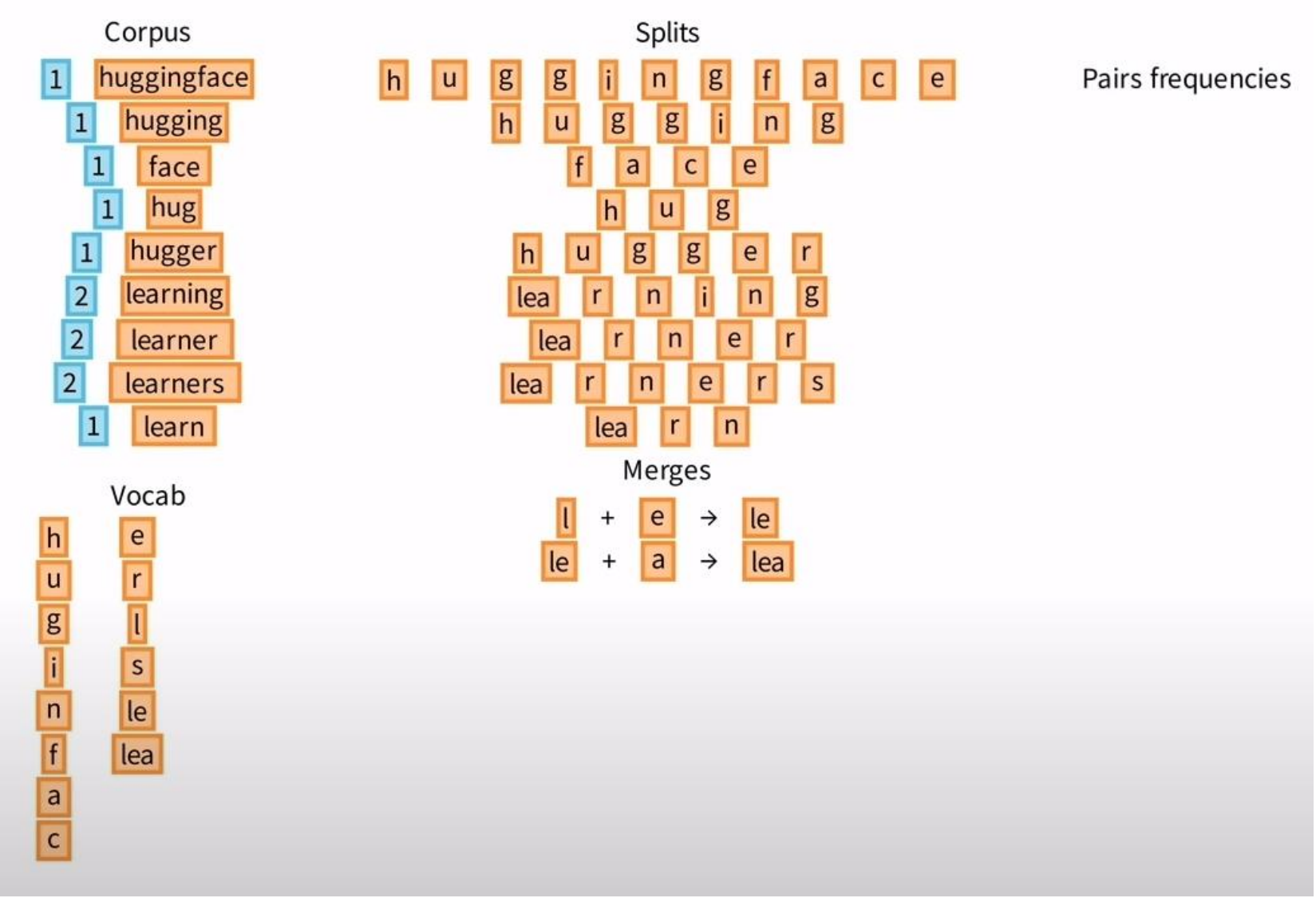
„Szótár” létrehozása, előre meghatározott mérettel



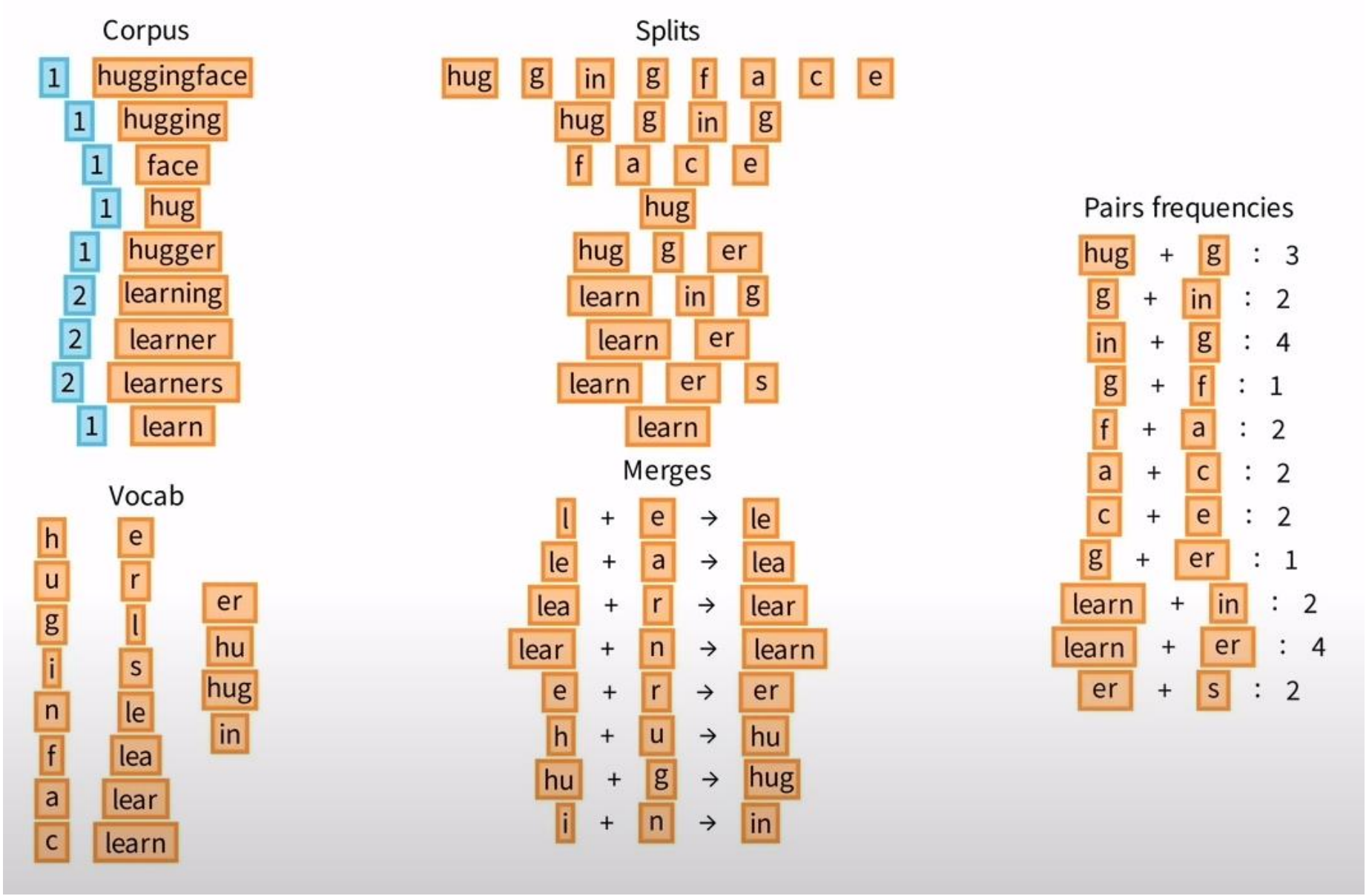
Keressünk újra token párokat és számoljuk meg az előfordulásuk gyakoriságát

Hogyan készülünk a modell tanítására?

„Szótár” létrehozása, előre meghatározott mérettel

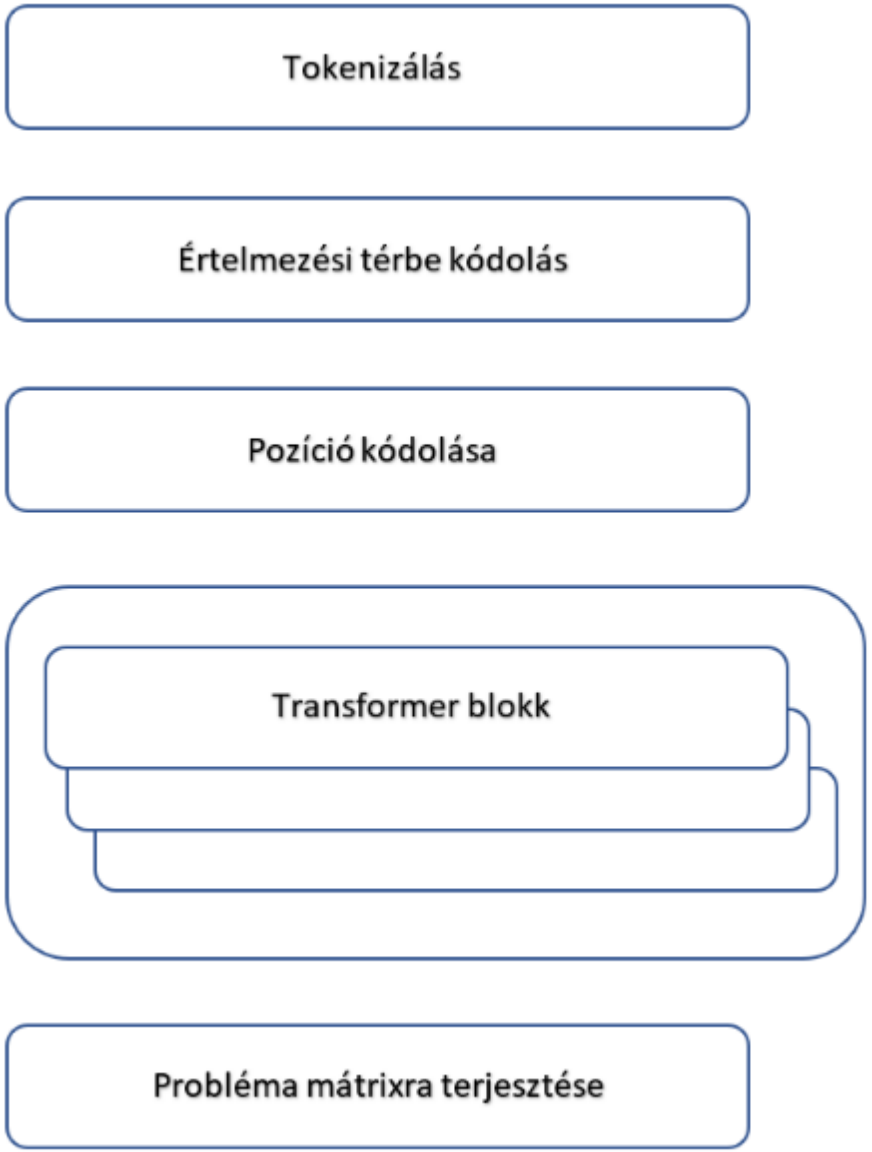


Hogyan készülünk a modell tanítására?



Addig csináljuk, amíg el nem fogy a hely a szótárban.
Mekkora legyen a magyar szótár?
Megtaláltuk, ez az egyik kutatási eredményünk

Hogyan készülünk a modell tanítására?



Következőnek átnézhetjük az értelmezési térbe kódolást.

Köszönjük a figyelmet!

Kérdések?