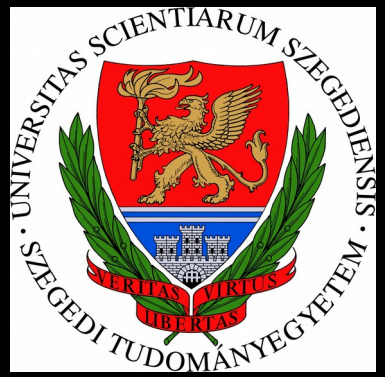
Ritka szóreprezentációk



Szóbeágyazások (implicit) célja

- Szavakhoz rendeljünk vektorokat, úgy, hogy azok tükrözzék a szópárok együttelőfordulási szokásait

Szóbeágyazások (implicit) célja

- Szavakhoz rendeljünk vektorokat, úgy, hogy azok tükrözzék a szópárok együttelőfordulási szokásait
  - Ha i és j szavak gyakran fordulnak elő, akkor a szavakhoz társított vektorok pontszorzata legyen nagy
    - Ellenkező esetben pedig kicsi

Szóbeágyazások (implicit) célja

- Szavakhoz rendeljünk vektorokat, úgy, hogy azok tükrözzék a szópárok együttelőfordulási szokásait
  - Ha i és j szavak gyakran fordulnak elő, akkor a szavakhoz társított vektorok pontszorzata legyen nagy
    - Ellenkező esetben pedig kicsi
    - Pl. kutya=[3 2], macska=[2 4], gőzmozdony=[-2 -1]

$$\overrightarrow{kutya}^T \overrightarrow{macska} \gg \overrightarrow{kutya}^T \overrightarrow{gőzmozdony}$$

word2vec célja

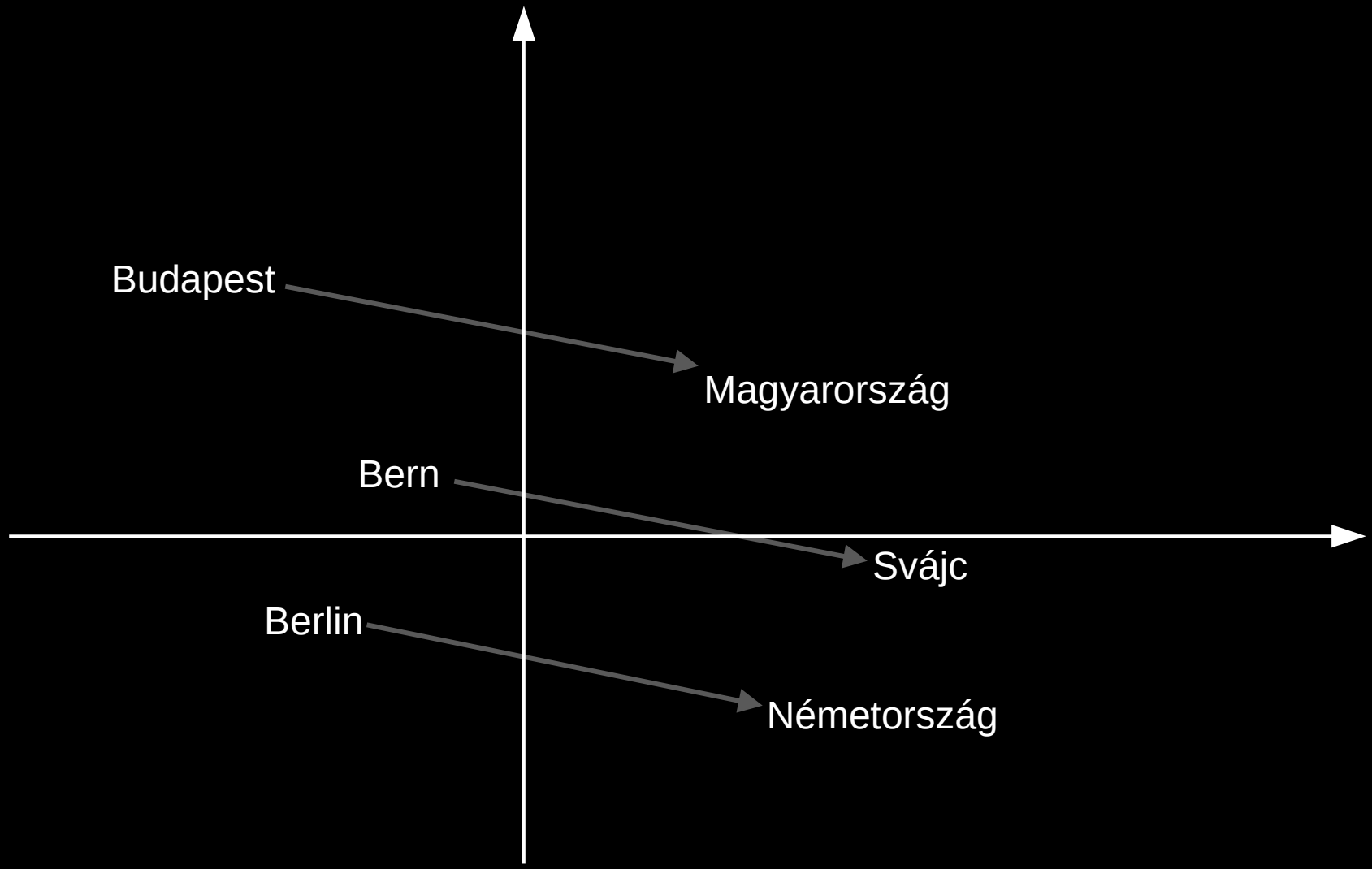
- Hasonló jelentésű input szavak kontextusukhoz illeszkedő outputot eredményezzenek

$$y(x) = softmax(V(W 1_x))$$

  - a és b szó jelentése minél hasonlóbb, y(a) és y(b) (eloszlás)vektorok annál inkább hasonlítani fognak

Szóanalógiák

- a:b::c:?



RepEval 2016

Analysis Track

- **Problems With Evaluation of Word Embeddings Using Word Similarity Tasks** [pdf]
  Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, Chris Dyer

- **Intrinsic Evaluations of Word Embeddings: What Can We Do Better?** [pdf]
  Anna Gladkova, Aleksandr Drozd

- **Issues in Evaluating Semantic Spaces Using Word Analogies** [pdf]
  Tal Linzen

- **Intrinsic Evaluation of Word Vectors Fails to Predict Extrinsic Performance** [pdf]
  Billy Chiu, Anna Korhonen, Sampo Pyysalo

- **A Critique of Word Similarity as a Method for Evaluating Distributional Semantic Models** [pdf]
  Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, David Weir

RepEval 2016

Analysis Track

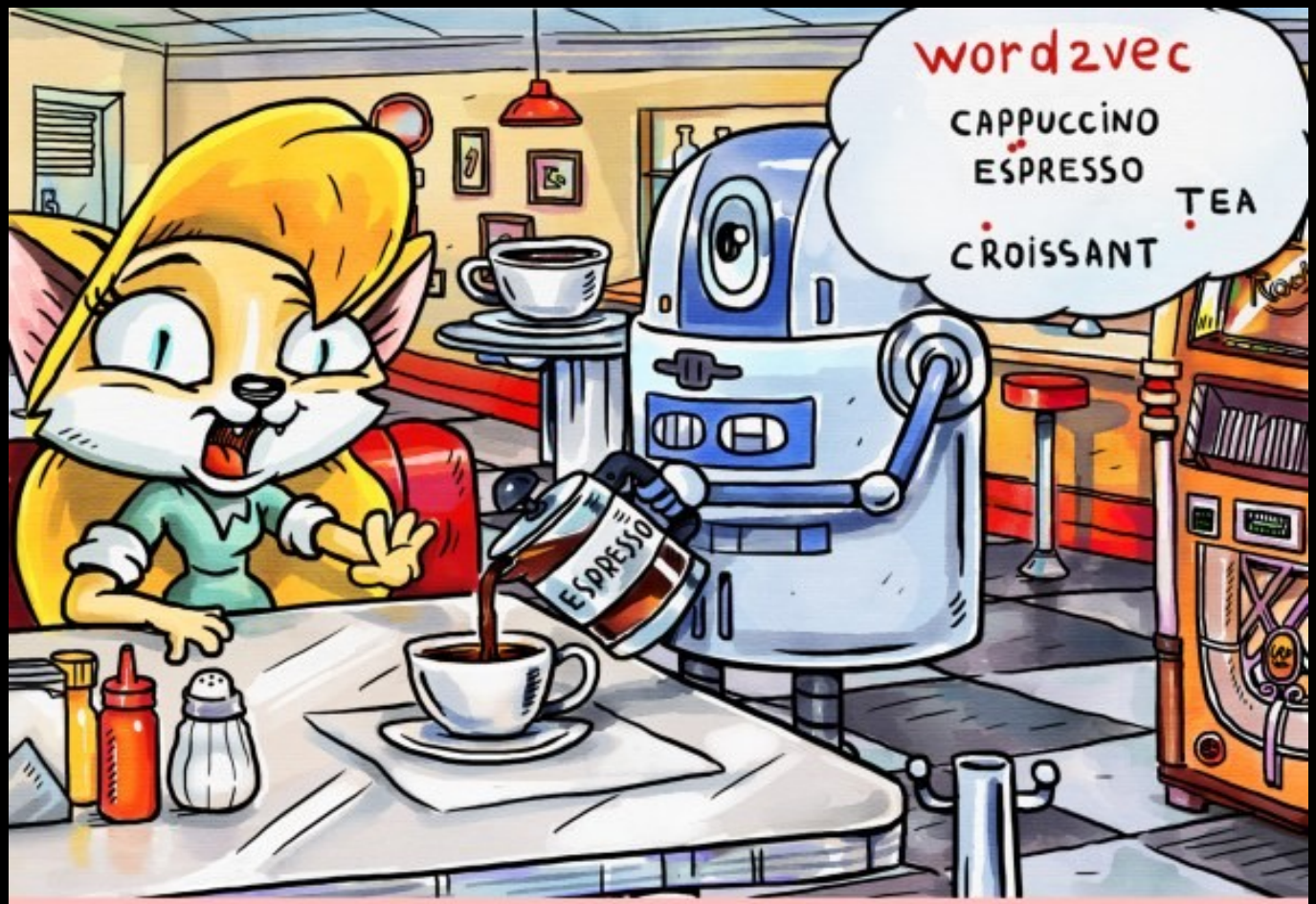
- **Problems With Evaluation of Word Embeddings Using Word Similarity Tasks** [pdf]
  Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, Chris Dyer

- **Intrinsic Evaluations of Word Embeddings: What Can We Do Better?** [pdf]
  Anna Gladkova, Aleksandr Drozd

- **Issues in Evaluating Semantic Spaces Using Word Analogies** [pdf]
  Tal Linzen

- **Intrinsic Evaluation of Word Vectors Fails to Predict Extrinsic Performance** [pdf]
  Billy Chiu, Anna Korhonen, Sampo Pyysalo

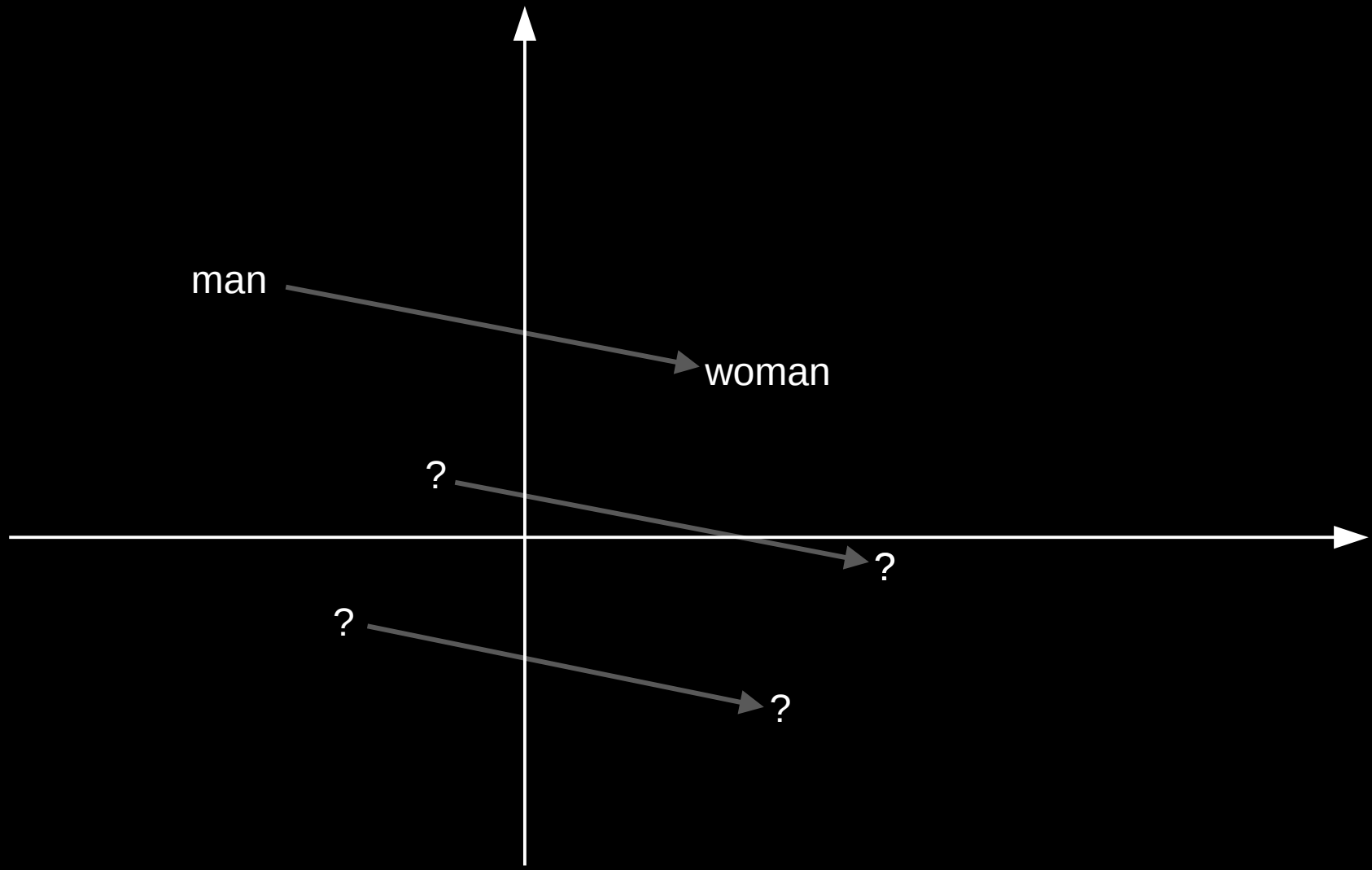
- **A Critique of Word Similarity as a Method for Evaluating Distributional Semantic Models** [pdf]
  Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, David Weir



- Espresso? But I ordered a cappuccino!
- Don't worry, the cosine distance between them is so small that they are almost the same thing.

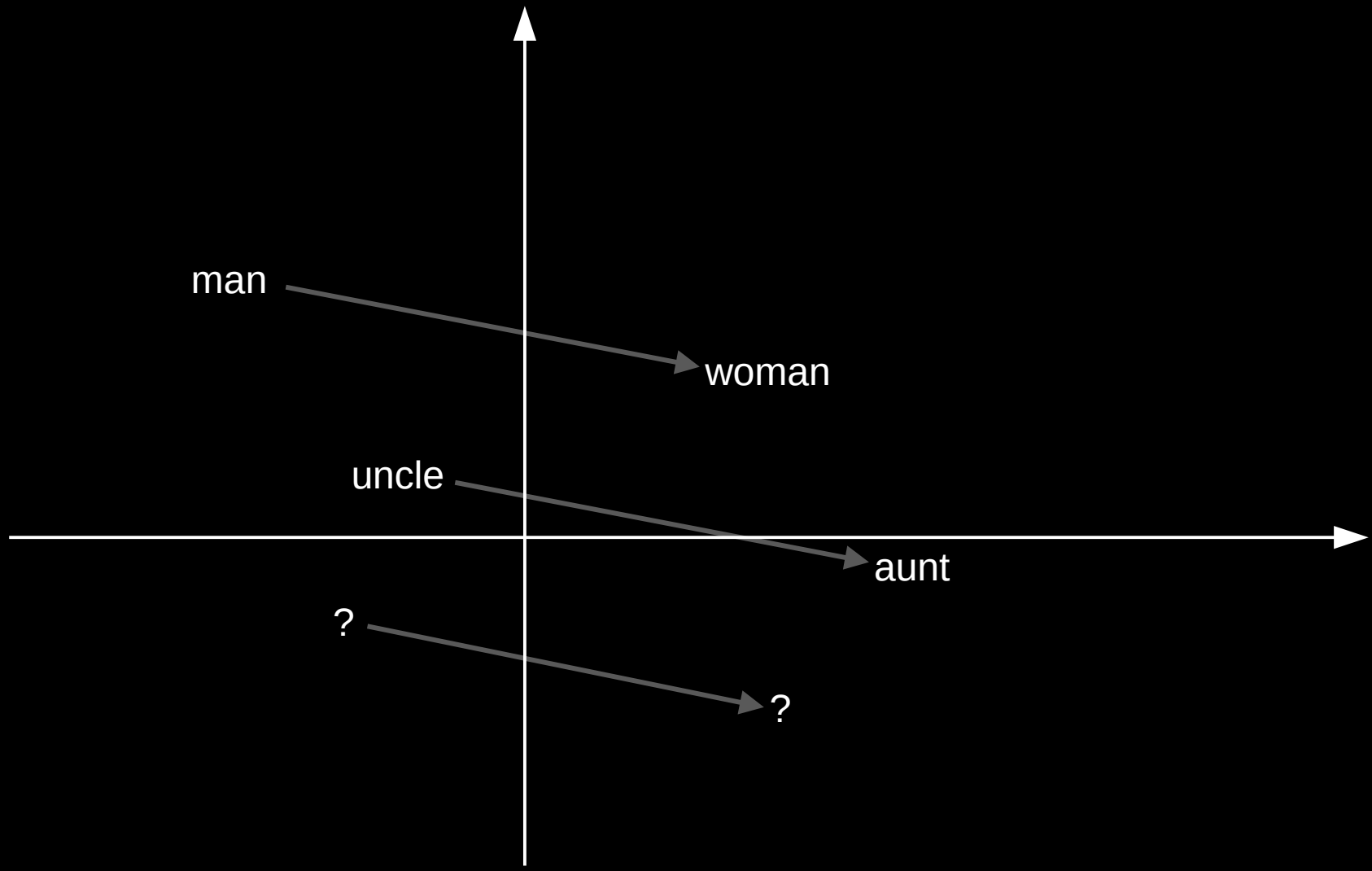
Szóanalógiák újratöltve

- a:b:?:?



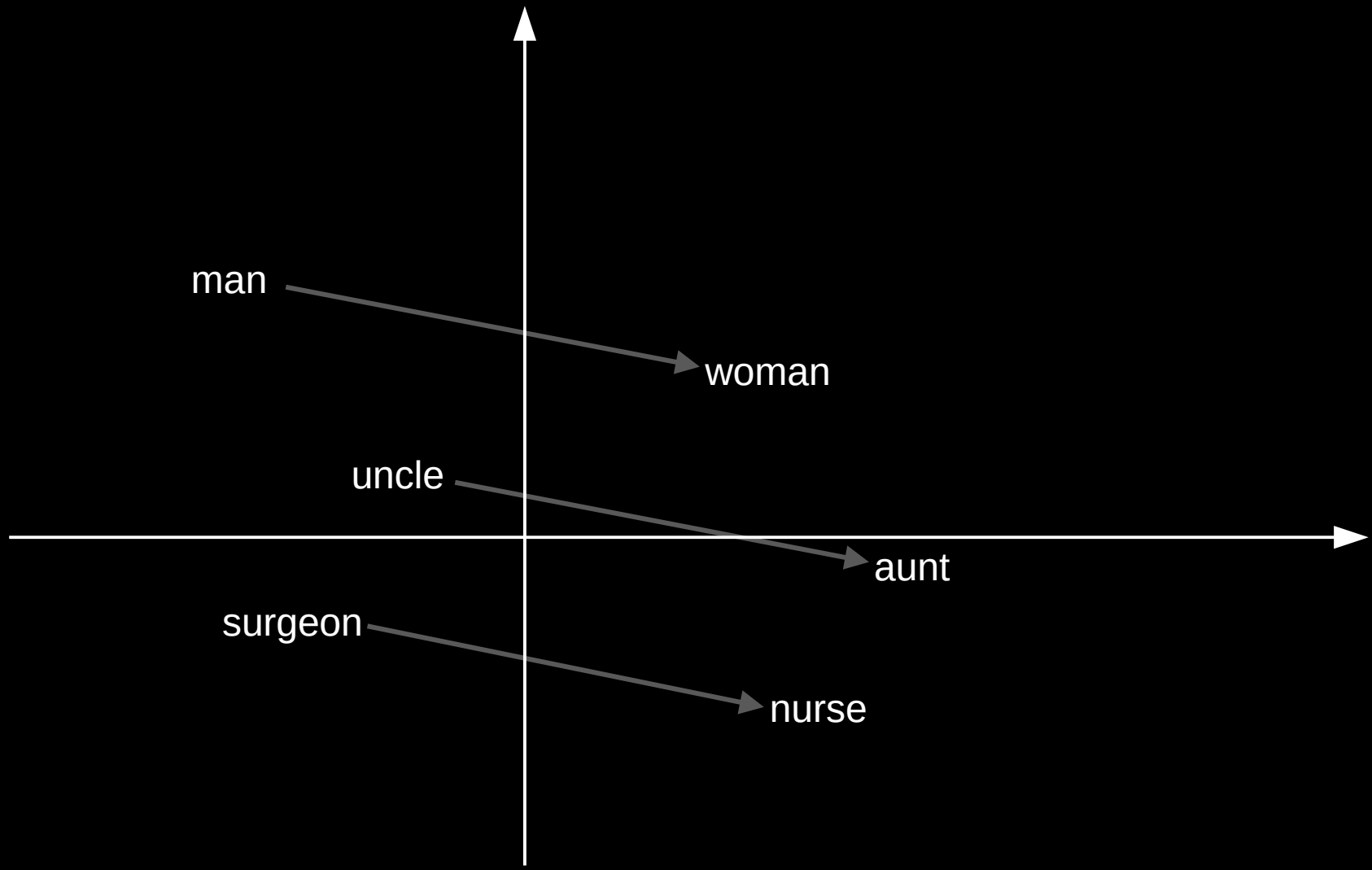
Szóanalógiák újratöltve

- a:b:?:?



Szóanalógiák újratöltve

- a:b:?:?



Folytonos szóreprezentációk

alma [1 0 0 0 … 0 0 0 0 0 … 0] → [3.2 -1.5]

...

banán [0 0 0 0 … 1 0 0 0 0 … 0] → [2.8 -1.6]

...

lapát [0 0 0 0 … 0 0 1 0 0 … 0] → [-1.1 12.6]

…

zebra [0 0 0 0 … 0 0 0 0 0 … 1] → [0.8 0.5]

Folytonos szóreprezentációk korlátai

- A megtanult reprezentációk pont olyan jók, mint amilyen a rendelkezésre álló korpusz
  - man : programmer :: women : X
  - Alacsony fedés (agglutináció)
    - Karakterszintű (morfológia alapú) modellek
  - Poliszémia
  - Többnyelvűség
  - Korlátozott interpretálhatóság

Ritka & folytonos szóreprezentációk

alma [3.2 -1.5] → [0 1.7 0 0 -0.2 0]

...

banán [2.8 -1.6] → [0 1.1 0 0 -0.4 0]

...

lapát [-1.1 12.6] → [1.7 0 -2.1 0 0 -0.8]

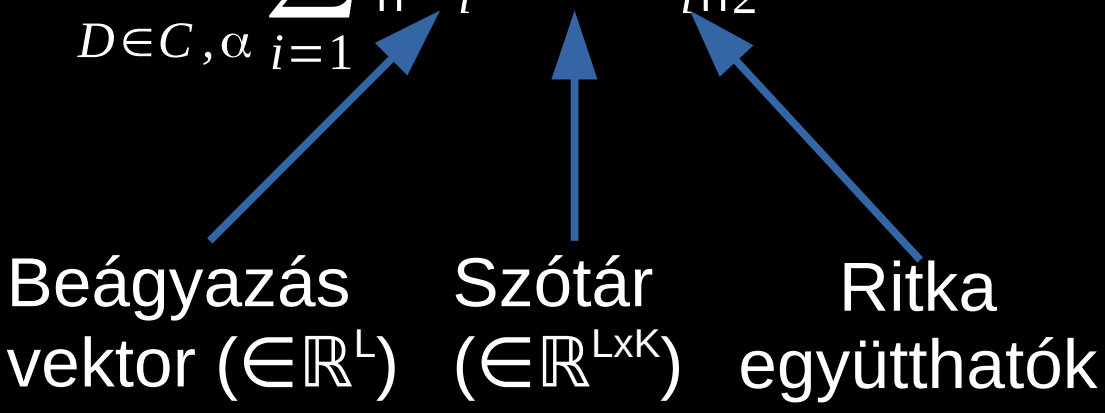
...

zebra [0.8 0.5] → [0 0 1.3 0 -1.2 0]

Ritka folytonos reprezentációk tanulása

- Adott x_i (i=1,…,|V|) szóbeágyazások esetén

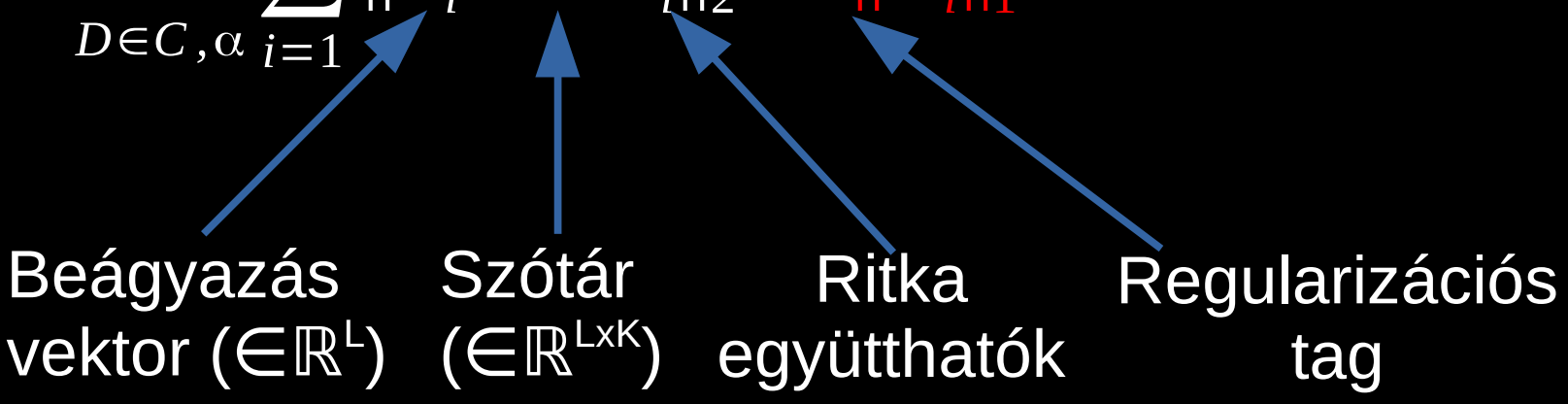
$$\min_{D \in C, \alpha} \sum_{i=1}^{|V|} \| x_i - D\,\alpha_i \|_2^2$$



Ritka folytonos reprezentációk tanulása

- Adott x_i (i=1,…,|V|) szóbeágyazások esetén

$$\min_{D \in C, \alpha} \sum_{i=1}^{|V|} \|x_i - D\,\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1$$



Ritka folytonos reprezentációk tanulása

- Adott x_i (i=1,…,|V|) szóbeágyazások esetén

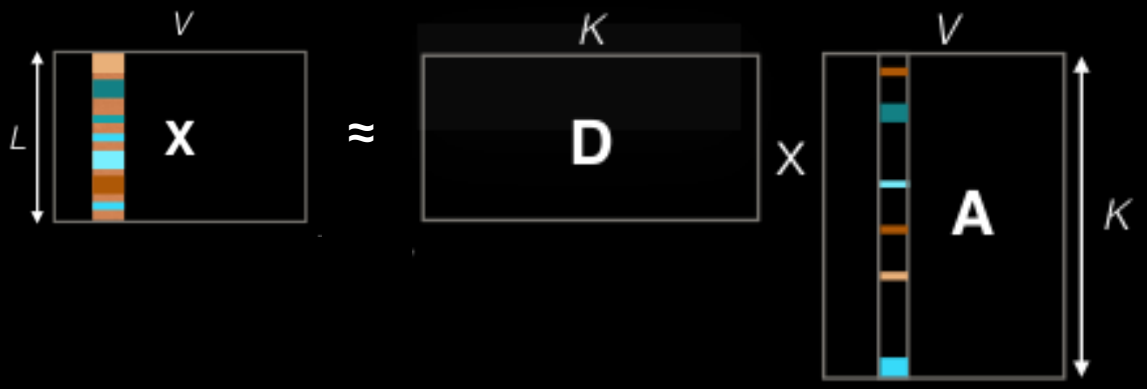
$$\min_{D \in C, \alpha} \sum_{i=1}^{|V|} \|x_i - D\,\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1$$

Beágyazás vektor ($\in \mathbb{R}^L$)

Szótár ($\in \mathbb{R}^{LxK}$)

Ritka együtthatók

Regularizációs tag



Szófaji kódolás (POS tagging) feladata

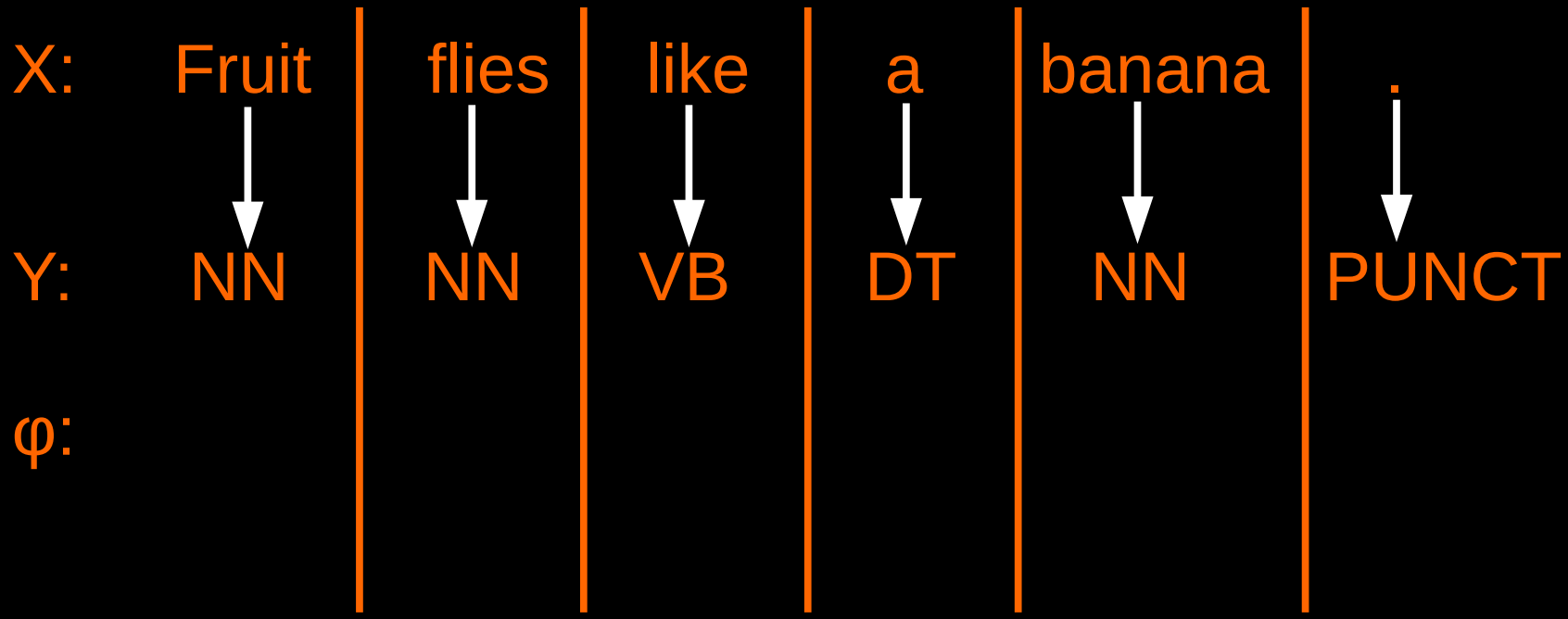
- Adott természetes nyelvű tokensorozatra határozzuk meg az egyes tokenek szófaját
  - Tipikus *in vivo* kiértékelés

Token	A	vonat	nem	vár	.
Elvárt szófaj	DET	NN	RB	VB	PUNCT
Predikált szófaj	DET	NNP	DET	VB	PUNCT
Helyes?	+1	0	0	+1	+1

- Kiértékelés
  - Pontosság: #eltalált szófajú szavak/#összes szó

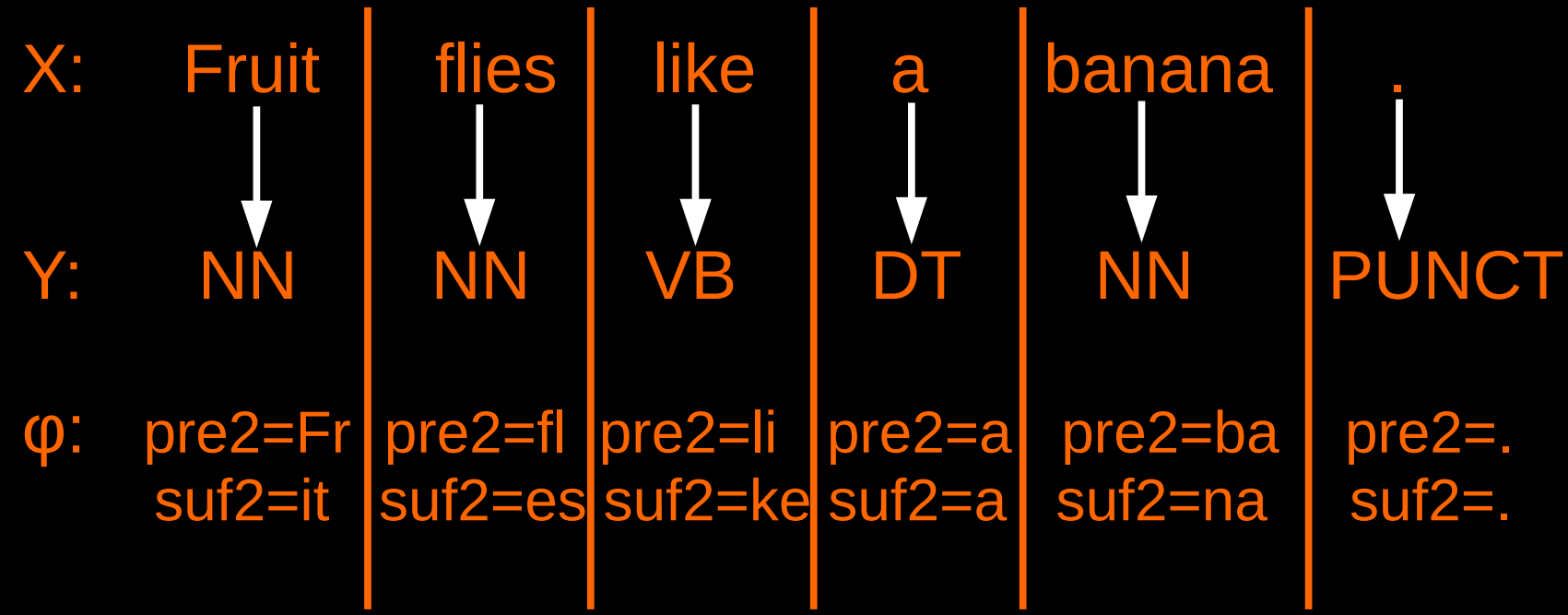
“Klasszikus” szekvenciajelölő

- Minden szóhoz számítsunk ki φ_j jellemzőket
  - φ_j vizsgálhatja pl. egy szó felszíni jegyeit (prefixum/szuffixum), de a szókörnyezetét is



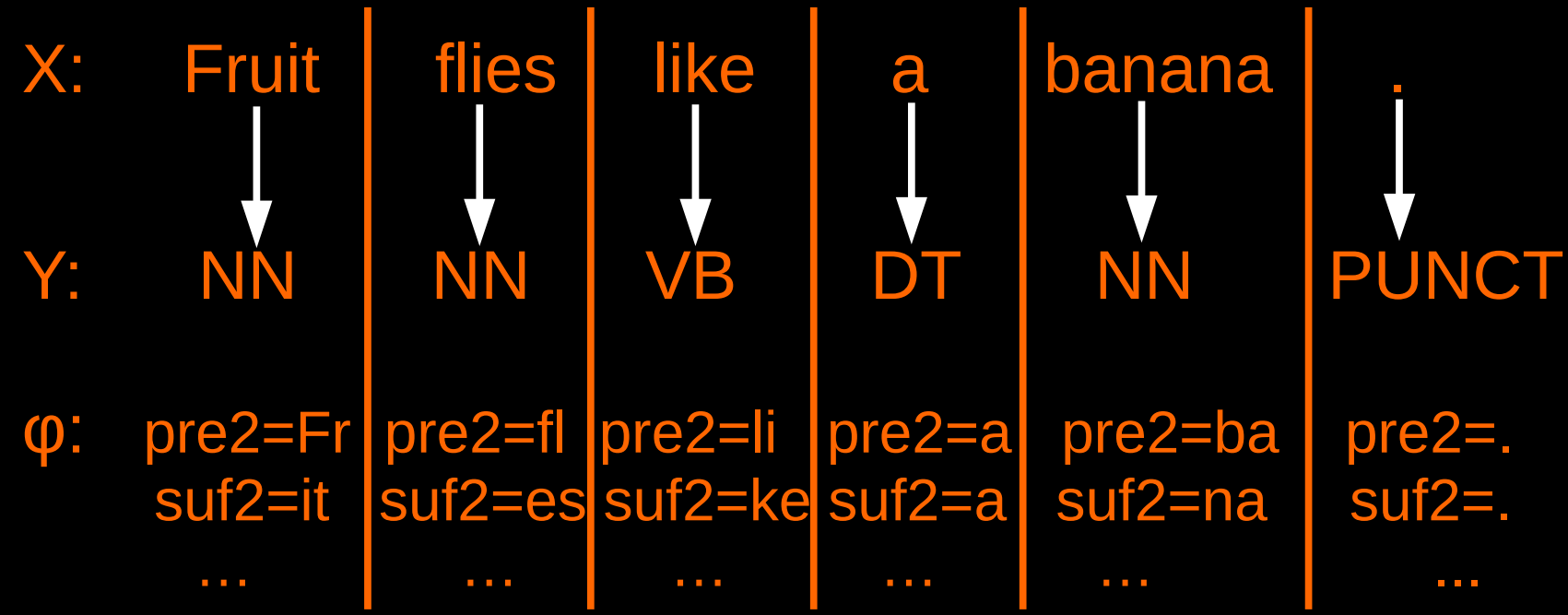
“Klasszikus” szekvenciajelölő

- Minden szóhoz számítsunk ki φ_j jellemzőket
  - φ_j vizsgálhatja pl. egy szó felszíni jegyeit (prefixum/szuffixum), de a szókörnyezetét is



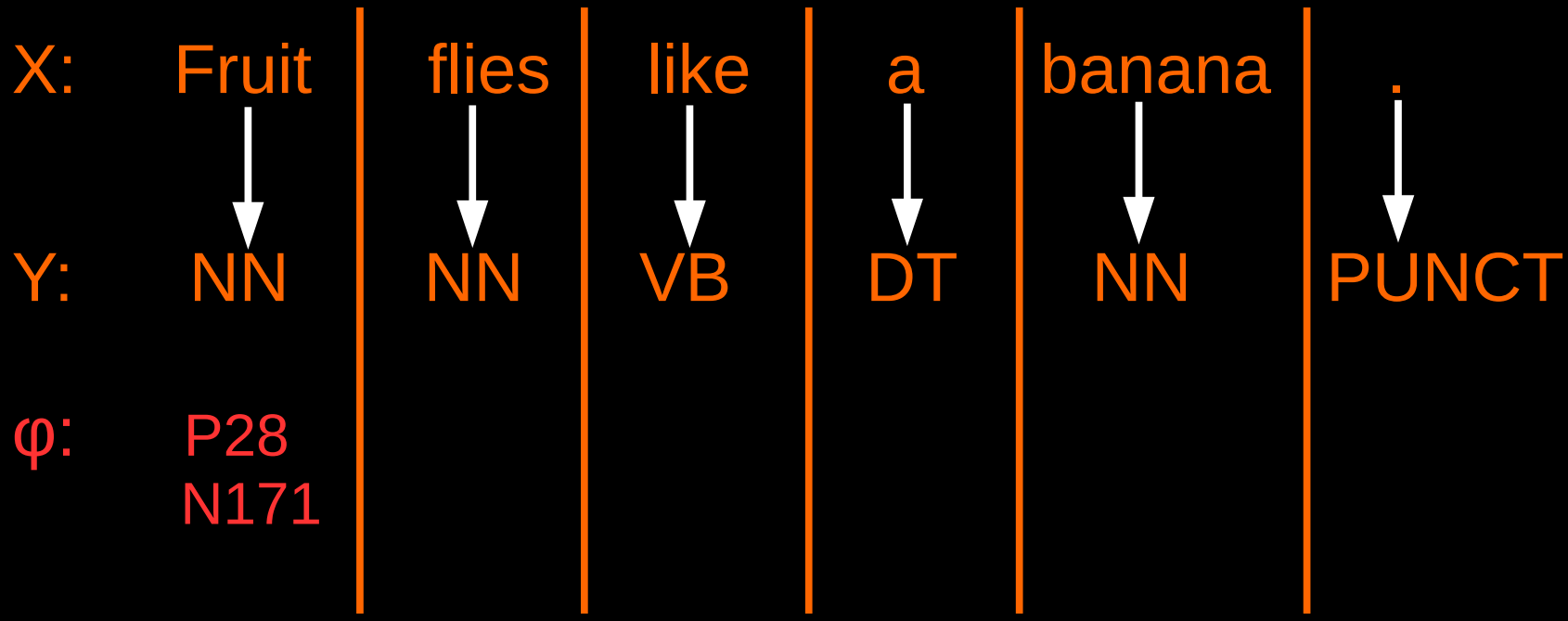
“Klasszikus” szekvenciajelölő

- Minden szóhoz számítsunk ki φ_j jellemzőket
  - φ_j vizsgálhatja pl. egy szó felszíni jegyeit (prefixum/szuffixum), de a szókörnyezetét is



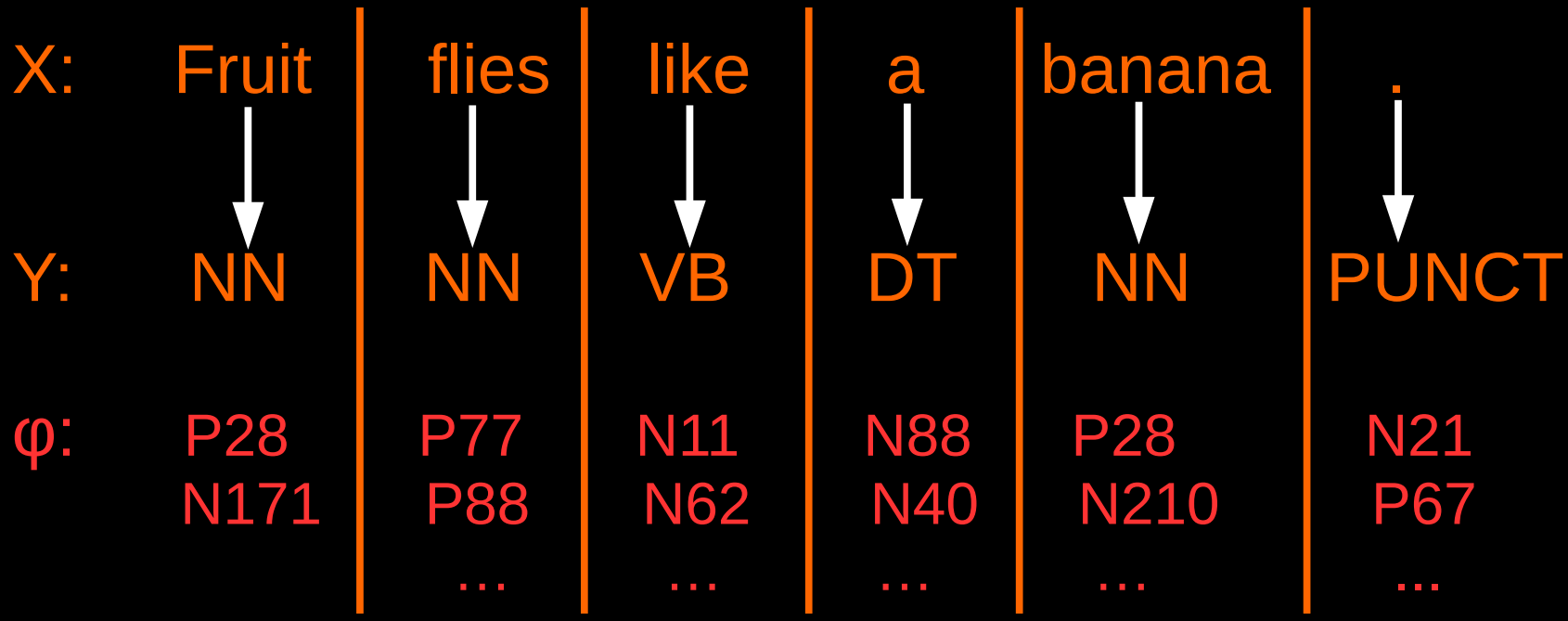
Szekvenciajelölő ritka jellemzőkkel

- Használjuk a dekompozícióból jövő α-t
  - Pl. $\overrightarrow{Fruit} \approx 1.1 \cdot \overrightarrow{d_{28}} - 0.4 \cdot \overrightarrow{d_{171}}$



Szekvenciajelölő ritka jellemzőkkel

- Használjuk a dekompozícióból jövő α-t
  - Pl. $\overrightarrow{Fruit} \approx 1.1 \cdot \overrightarrow{d_{28}} - 0.4 \cdot \overrightarrow{d_{171}}$



Kísérleti körülmények

- Lineáris CRF (CRFsuite implementáció)
- Szófaji kódolás 12 nyelvre (CoNLL-X shared task)
  - Google Universal Tag Set (12 szófaj)

Kísérleti körülmények

- Lineáris CRF (CRFsuite implementáció)
- Szófaji kódolás 12 nyelvre (CoNLL-X shared task)
  - Google Universal Tag Set (12 szófaj)

- Hiperparaméterek
  - polyglot/w2v/Glove
  - L=64
  - K=1024
  - Különböző λ-k

$$\min_{D \in C, \alpha} \sum_{i=1}^{|V|} \|w_i - D\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1$$

Beágyazás vektor ($\in \mathbb{R}^L$)

Szótár ($\in \mathbb{R}^{LxK}$)

Ritka együtthatók

Baselineok

- Gazdag jellemzőkészlet (FR)
  - Hagyományos jellemzők a CRFsuite-ből kölcsönözve
    - Megelőző, rákövetkező szóalak, szókombinációk, …
  - 2 változat:
    - Karakter és szószintű modell (FR_{w+c})
    - Szószintű modell (FR_{w})

Baselineok

- Gazdag jellemzőkészlet (FR)
  - Hagyományos jellemzők a CRFsuite-ből kölcsönözve
    - Megelőző, rákövetkező szóalak, szókombinációk, …
  - 2 változat:
    - Karakter és szószintű modell (FR_{w+c})
    - Szószintű modell (FR_{w})

$$FR_{w+c} \supset FR_{w}$$

Baselineok

- Gazdag jellemzőkészlet (FR)
  - Hagyományos jellemzők a CRFsuite-ból kölcsönözve
    - Megelőző, rákövetkező szóalak, szókombinációk, …
  - 2 változat:
    - Karakter és szószintű modell (FR_{w+c})
    - Szószintű modell (FR_{w})
- Brown klaszterezés
  - Jellemzők a szavak Brown klaszterazonosítóiból

Baselineok

- Gazdag jellemzőkészlet (FR)
  - Hagyományos jellemzők a CRFsuite-ból kölcsönözve
    - Megelőző, rákövetkező szóalak, szókombinációk, …
  - 2 változat:
    - Karakter és szószintű modell (FR_{w+c})
    - Szószintű modell (FR_{w})
- Brown klaszterezés
  - Jellemzők a szavak Brown klaszterazonosítóiból
- Ritkítatlan (aka. sűrű) szóbeágyazások alapján
  - $\phi(w_i) = \{ j : \alpha_i[j] \mid \forall\, j \in 1, \ldots, 64 \}$

Többnyelvű eredmények

- Eredmények kiátlagolva 12 nyelvre

	Sűrű
polyglot	91.17%
CBOW	88.30%
SG	86.89%
Glove	81.53%

- Fő észrevételek
  - polyglot > CBOW > SG > Glove

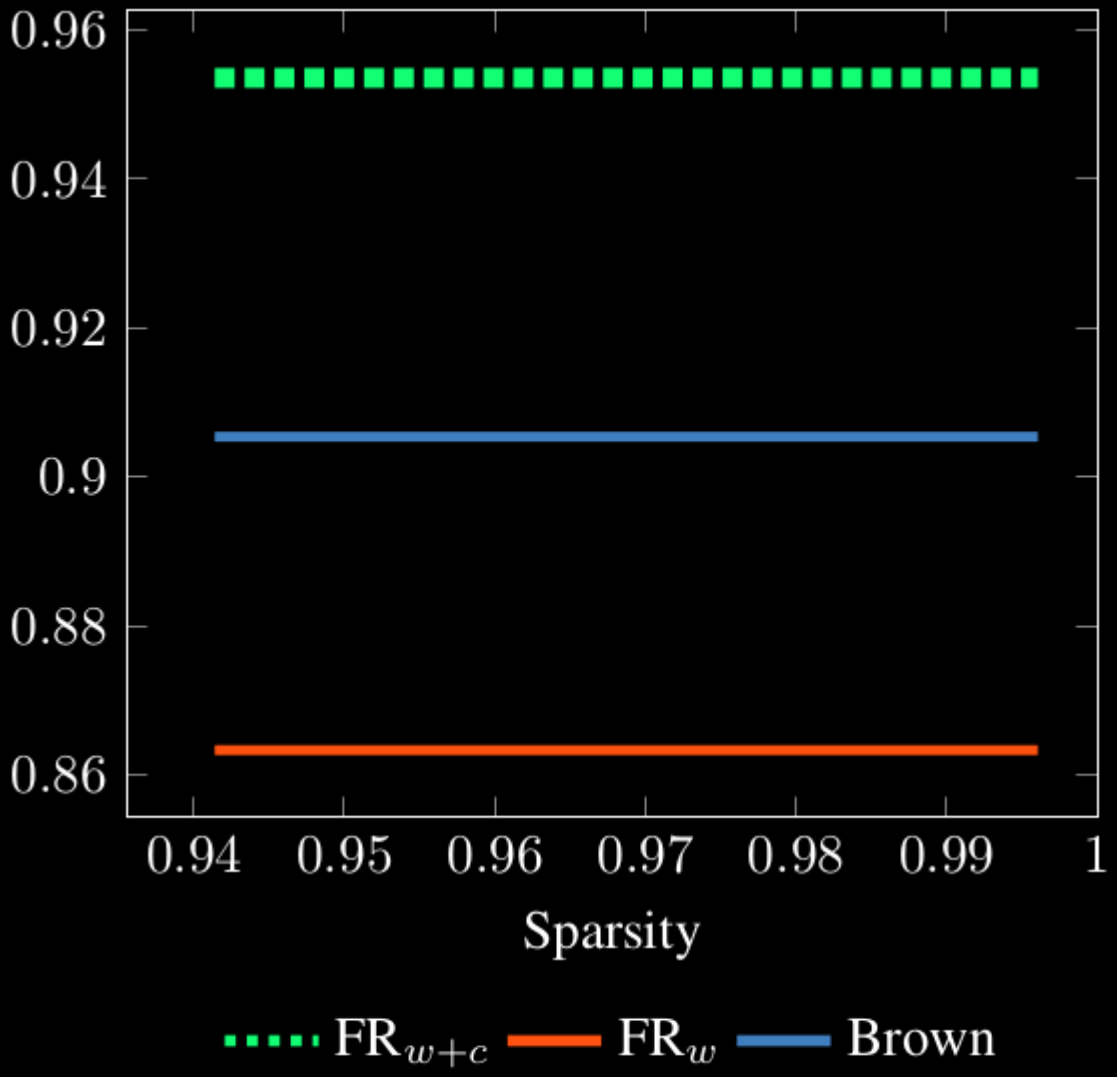
Többnyelvű eredmények

- Eredmények kiátlagolva 12 nyelvre

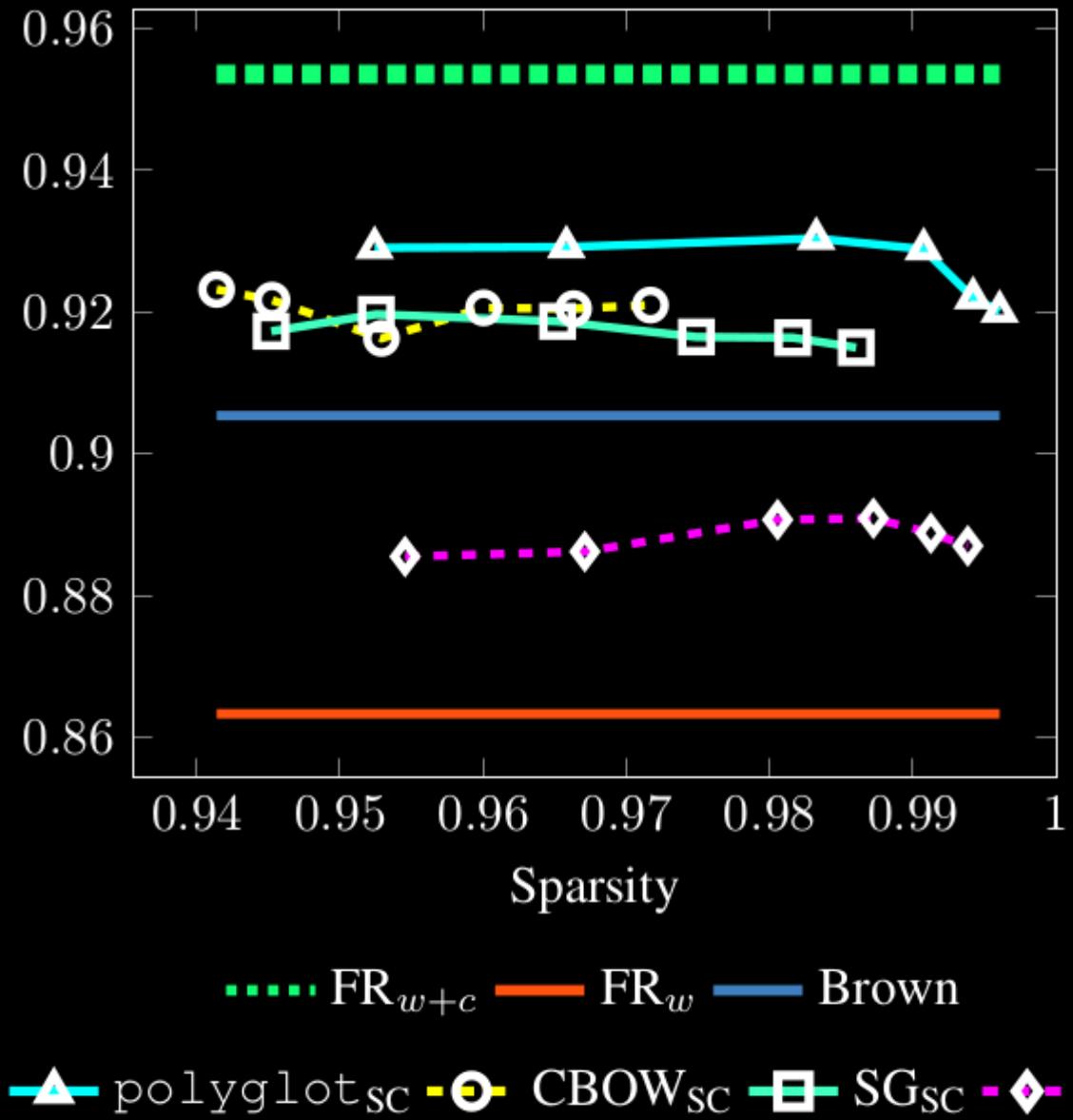
	Sűrű	R i t k a	Javulás
polyglot	91.17%	94.44%	+3.3
CBOW	88.30%	93.74%	+5.4
SG	86.89%	93.63%	+6.7
Glove	81.53%	91.92%	+10.4

- Fő észrevételek
  - polyglot > CBOW > SG > Glove
  - Ritka reprezentáció >> sűrű reprezentáció

Eredmények magyarra

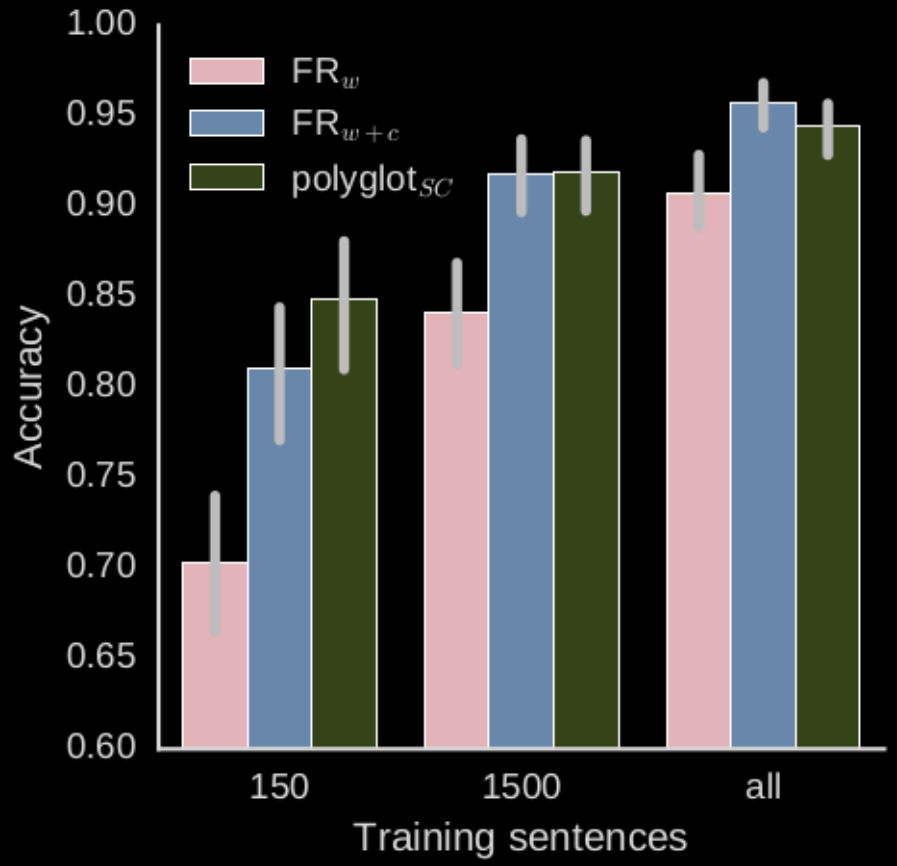


Eredmények magyarra



Általánosítóképesség

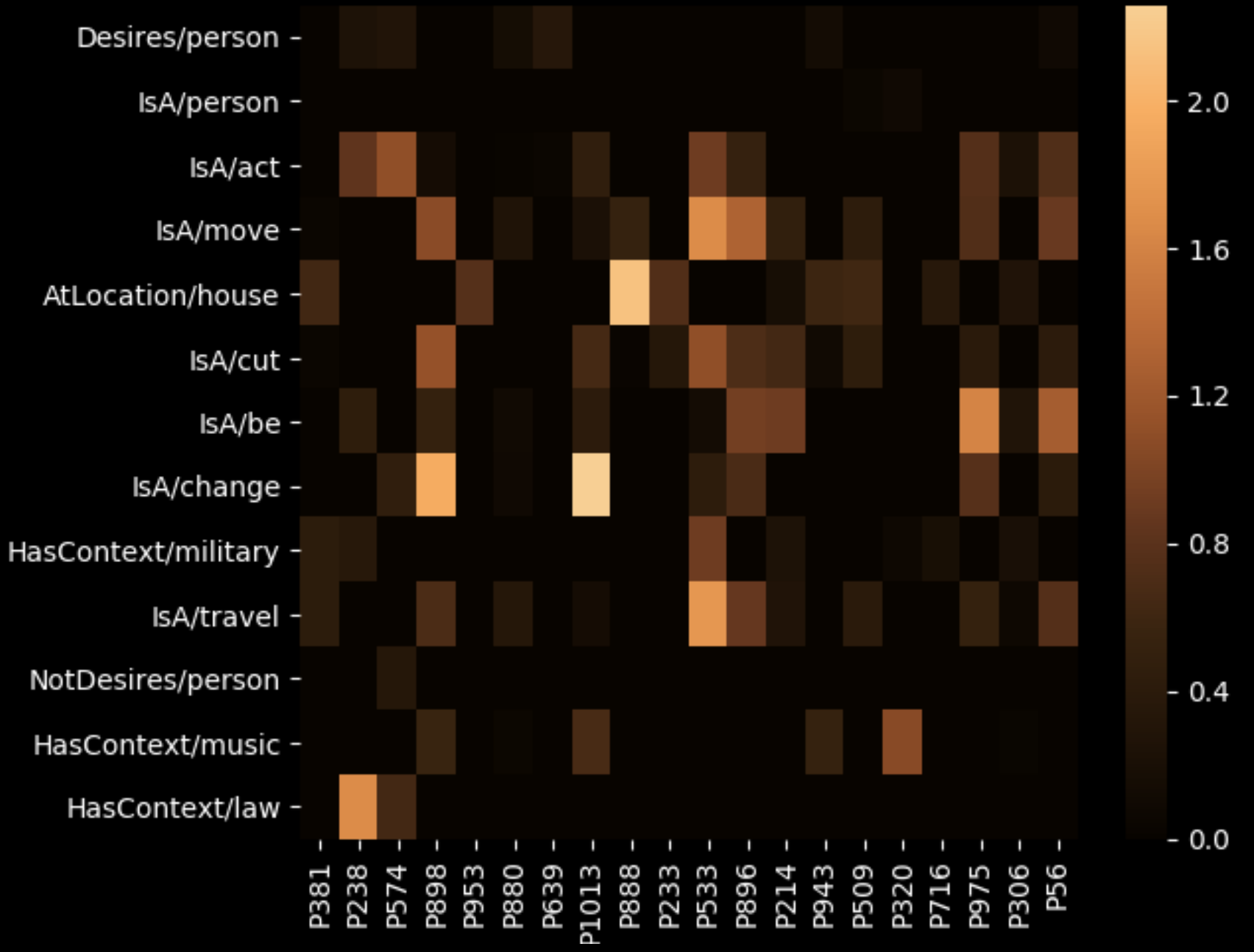
- A tanítóadatbázist mesterségesen lecsökkentettük
  - Első 150, illetve 1500 mondat fölhasználása



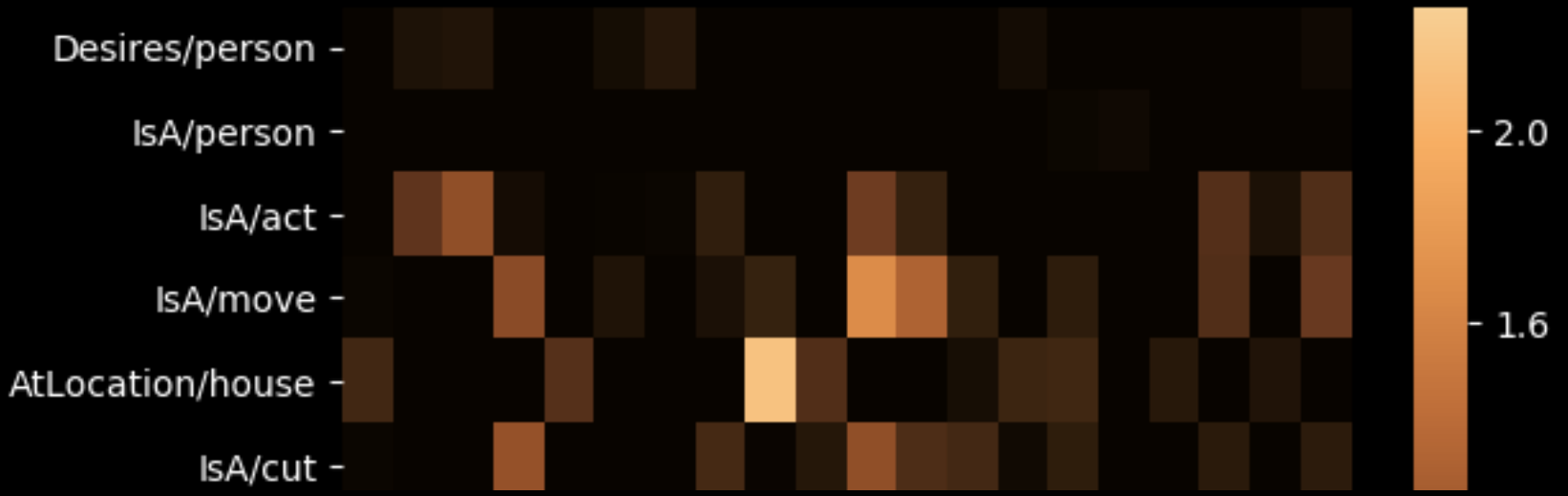
Ritka == interpretálható reprezentáció?

- ConceptNet (CN) tudásbázis alapján
  - A bázisok és CN attribútumok együttes előfordulásának vizsgálata
  - Pontonkénti kölcsönös információtartalom

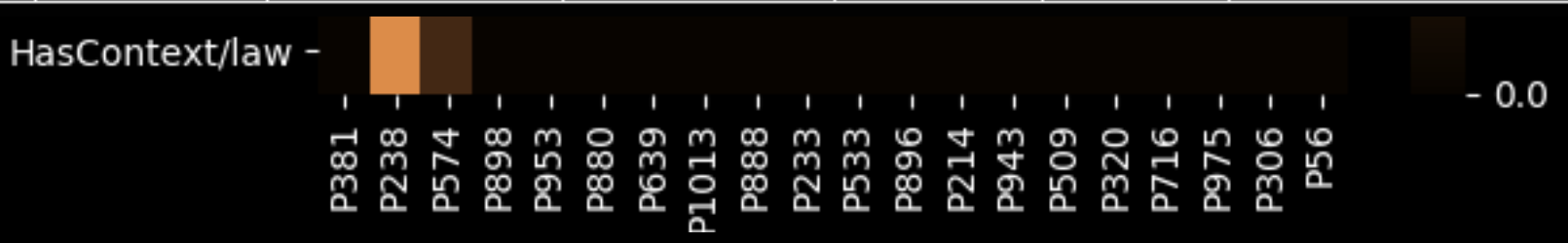
Ritka == interpretálható reprezentáció?



Ritka == interpretálható reprezentáció?



Basis	Top-1	Top-2	Top-3	Top-4	Top-5	Most associated ConcepNet relation
P381	village	neighbourhood	neighborhood	fort	township	AtLocation/house
P238	amendment	decision	inquiry	obligation	petition	HasContext/law
P574	stability	coherence	sensitivity	separation	efficiency	IsA/act
P898	harden	darken	pierce	flatten	loosen	IsA/change
P953	coal	oil	food	cotton	grain	AtLocation/house



Ritka szóreprezentációk több nyelvre

apple [3.2 -1.5] → [0 1.7 0 0 -0.2 0]

...

banana [2.8 -1.6] → [0 1.1 0 0 -0.4 0]

...

zebra [0.8 0.5] → [0 0 1.3 0 -1.2 0]

Ritka szóreprezentációk több nyelvre

apple [3.2 -1.5] → [0 1.7 0 0 -0.2 0]

...

banana [2.8 -1.6] → [0 1.1 0 0 -0.4 0]

...

zebra [0.8 0.5] → [0 0 1.3 0 -1.2 0]

---

alma [1.2 0.8] → [0.8 0 0 0.4 0 0]

...

banán [0.7 1.1] → [2.1 0 0 0.7 0 0]

...

zebra [-0.2 0.5] → [0 0 0.9 0 0 1.2]

Többnyelvű ritka szóreprezentációk

$$D^{(f)}, \alpha^{(f)} = \min_{D,\alpha} \sum_{i=1}^{|V|} \| w_i^{(f)} - D\,\alpha_i \|_2^2 + \lambda \| \alpha_i \|_1$$

- “Sűrű" szóbeágyazások ***f*** forrás,- és ***c*** célnyelv közötti leképezésére több megoldás ismert (Smith et al., 2017; Artetxe et al., 2016; Hamilton et al., 2016)

Többnyelvű ritka szóreprezentációk

$$D^{(f)}, \alpha^{(f)} = \min_{D,\alpha} \sum_{i=1}^{|V|} \| w_i^{(f)} - D\,\alpha_i \|_2^2 + \lambda \| \alpha_i \|_1$$

- “Sűrű" szóbeágyazások ***f*** forrás,- és ***c*** célnyelv közötti leképezésére több megoldás ismert (Smith et al., 2017; Artetxe et al., 2016; Hamilton et al., 2016)
  - A célnyelvre a forrásnyelv $D^{(f)}$ szótármátrixát és a forrásnyelv-célnyelv közötti M leképezést használjuk

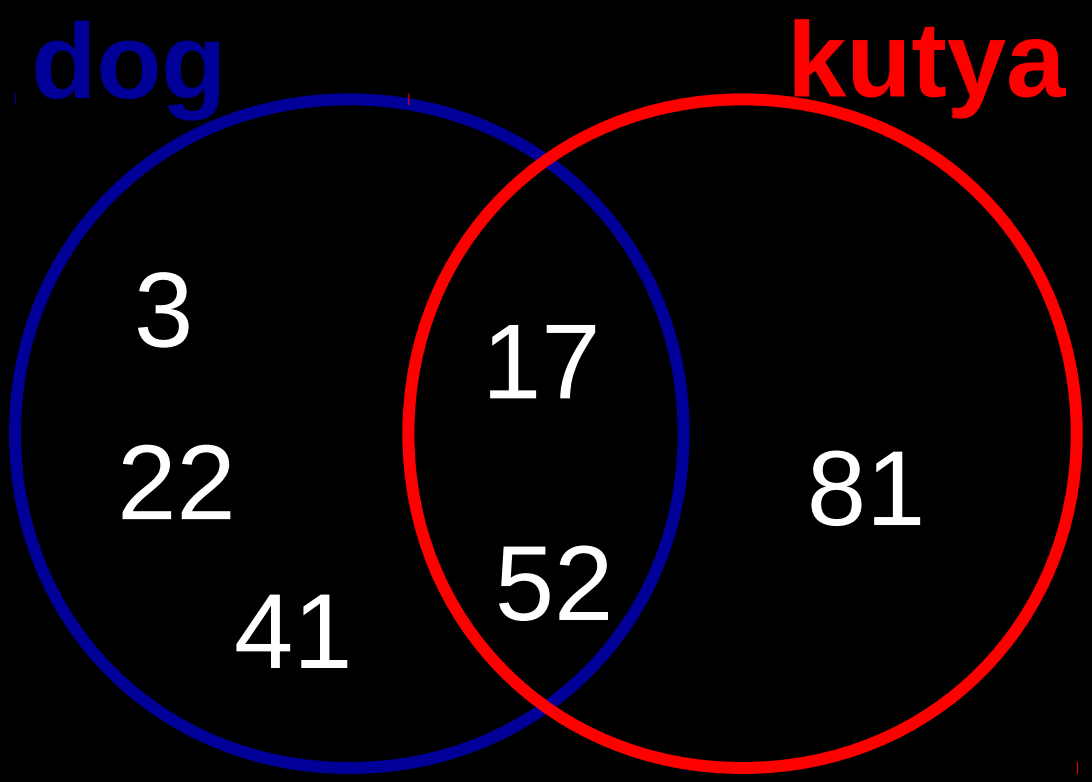
Többnyelvű ritka szóreprezentációk

$$D^{(f)}, \alpha^{(f)} = \min_{D,\alpha} \sum_{i=1}^{|V|} \| w_i^{(f)} - D\,\alpha_i \|_2^2 + \lambda \| \alpha_i \|_1$$

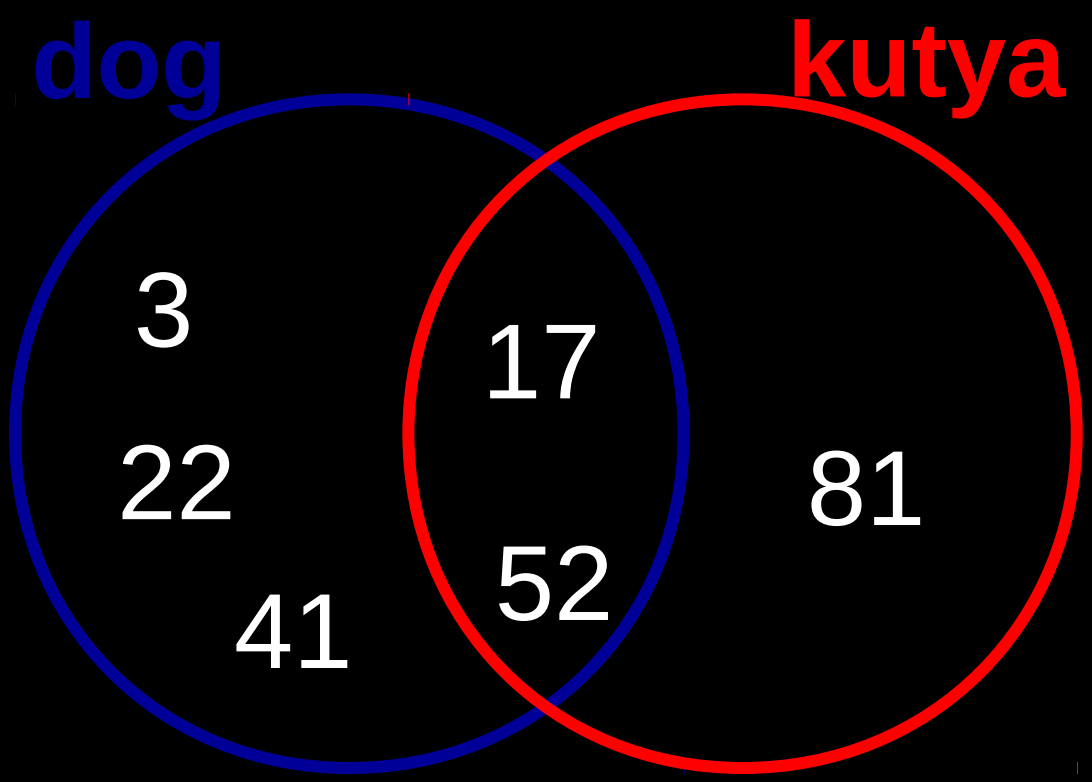
- “Sűrű" szóbeágyazások ***f*** forrás,- és ***c*** célnyelv közötti leképezésére több megoldás ismert (Smith et al., 2017; Artetxe et al., 2016; Hamilton et al., 2016)
  - A célnyelvre a forrásnyelv $D^{(f)}$ szótármátrixát és a forrásnyelv-célnyelv közötti M leképezést használjuk

$$\alpha^{(c)} = \min_{\alpha} \sum_{i=1}^{|V|} \| M\, w_i^{(c)} - D^{(f)}\, \alpha_i \|_2^2 + \lambda \| \alpha_i \|_1$$

Ritka reprezentációk átfedése

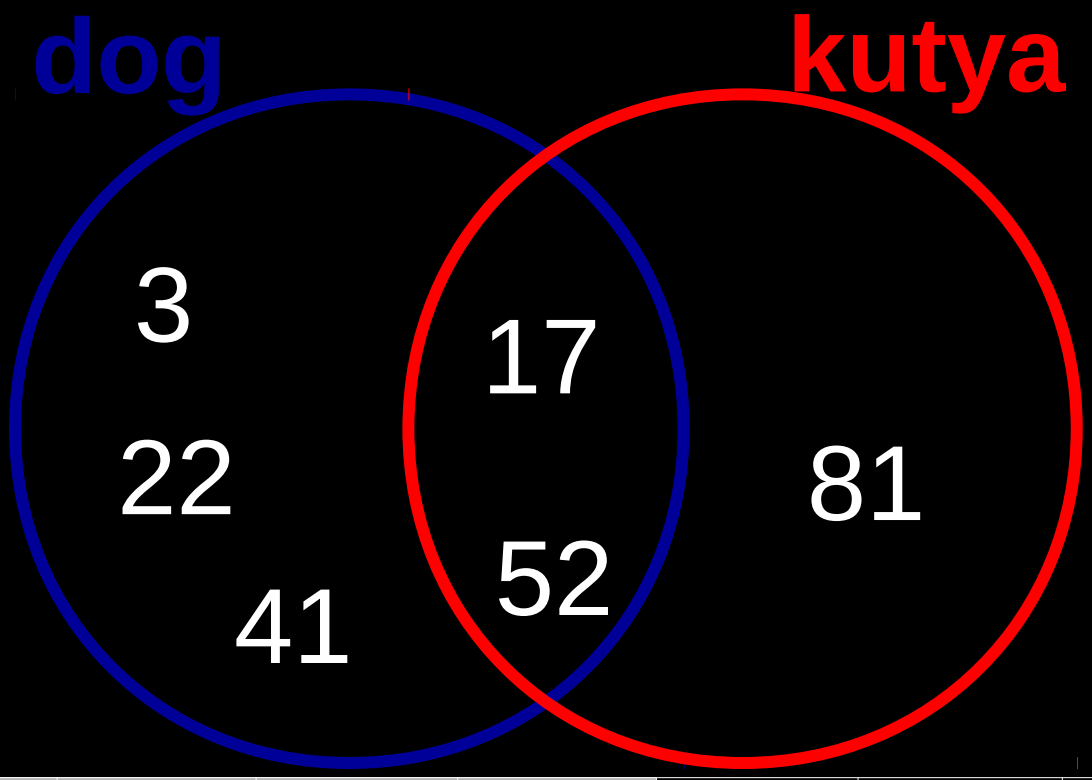


Ritka reprezentációk átfedése



P=2/3
R=2/5
F=1/2

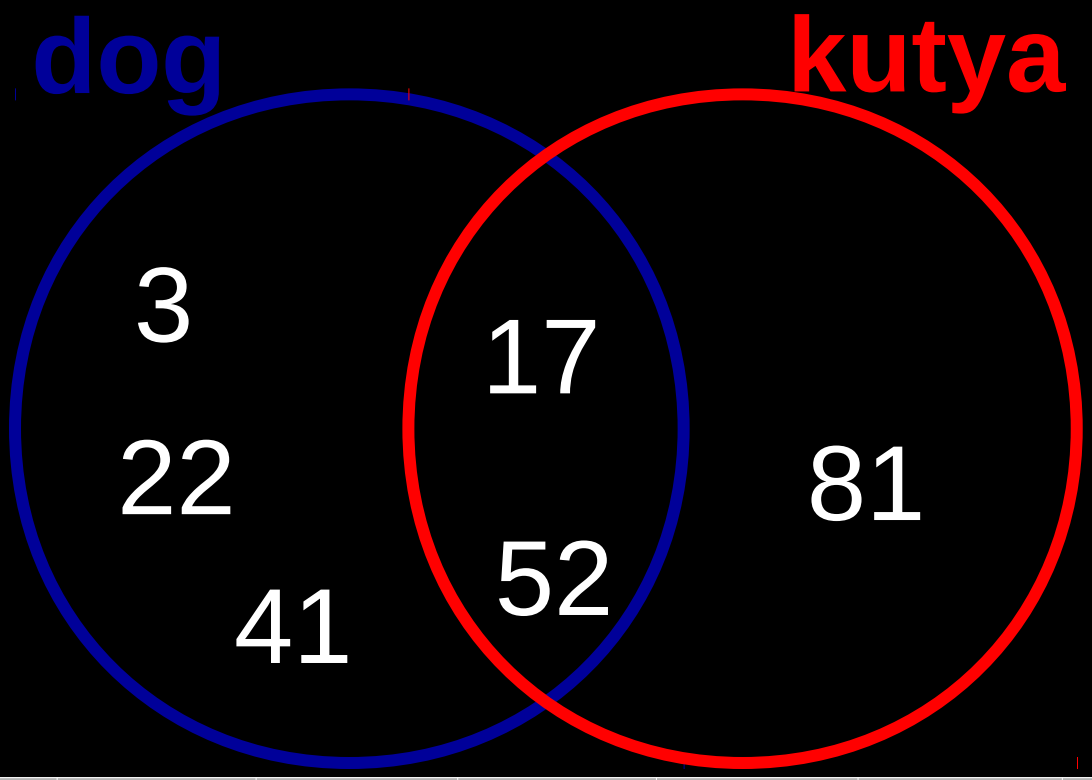
Ritka reprezentációk átfedése



P=2/3
R=2/5
F=1/2

λ	Prec.	Rec.	F
0.1	0.023	0.024	**0.024**
0.3	0.001	0.001	0.001
0.5	0.000	0.000	0.000
	M=I		

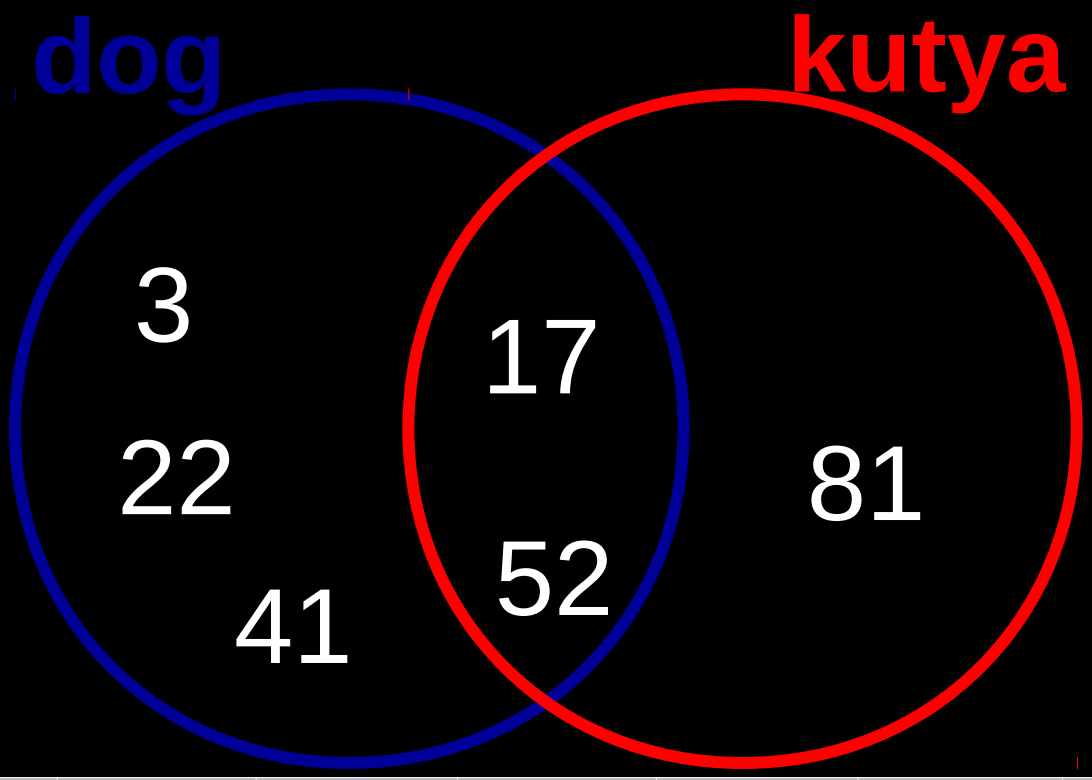
Ritka reprezentációk átfedése



P=2/3
R=2/5
F=1/2

λ	Prec.	Rec.	F	Prec.	Rec.	F			
0.1	0.023	0.024	**0.024**	0.170	0.117	0.139			
0.3	0.001	0.001	0.001	0.345	0.118	**0.176**			
0.5	0.000	0.000	0.000	0.600	0.009	0.018			
	M=I			Tetszőleges M					

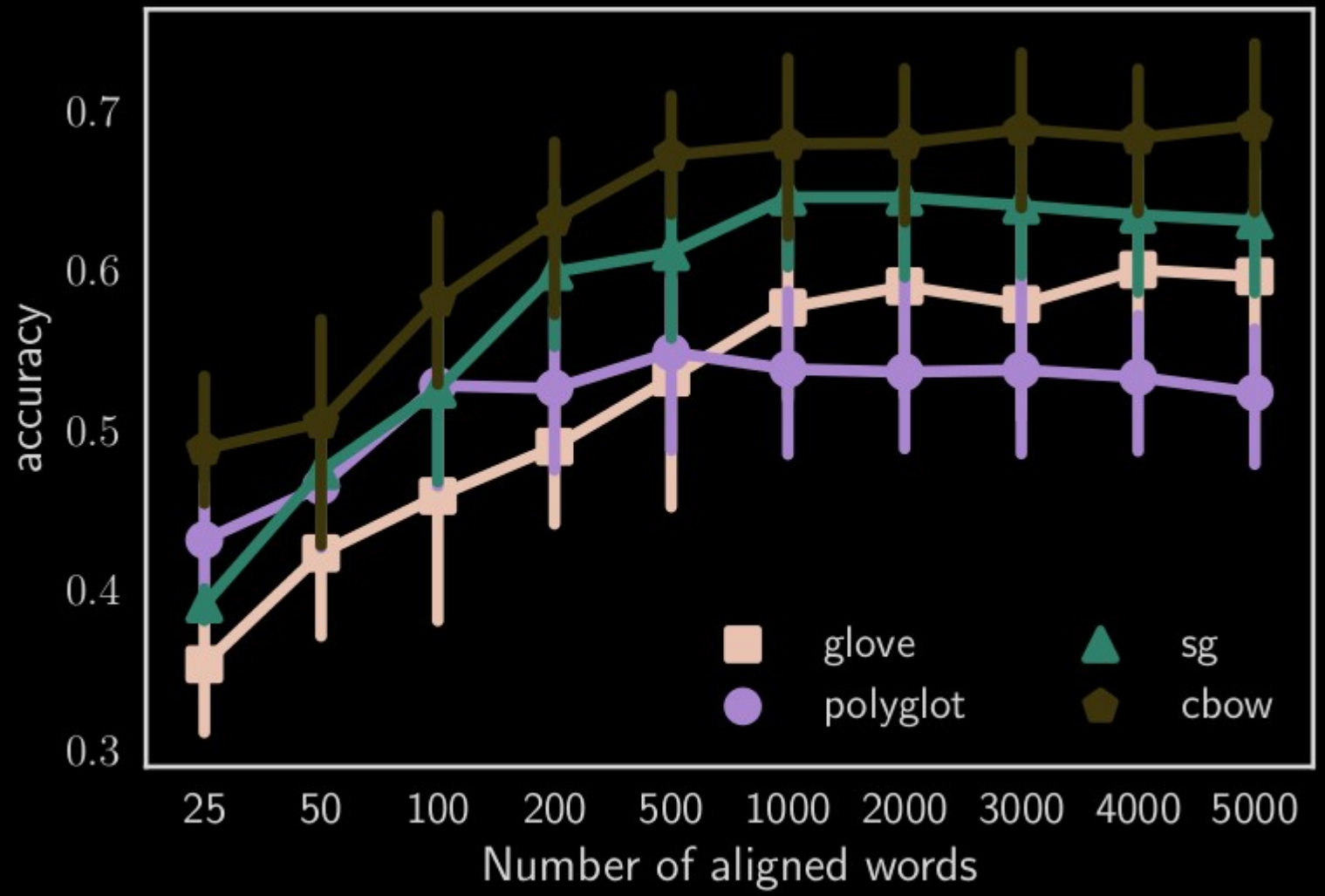
Ritka reprezentációk átfedése



P=2/3
R=2/5
F=1/2

λ	Prec.	Rec.	F	Prec.	Rec.	F	Prec.	Rec.	F
0.1	0.023	0.024	**0.024**	0.170	0.117	0.139	0.098	0.137	0.114
0.3	0.001	0.001	0.001	0.345	0.118	**0.176**	0.167	0.208	0.185
0.5	0.000	0.000	0.000	0.600	0.009	0.018	0.271	0.202	**0.232**
	M=I			Tetszőleges M			Ortogonális M		

Nyelvközi szófaji kódolás eredményei



Konklúzió

- Egyszerű modell, mégis pontos eredmények
- Nyelv,-és feladatközi robusztusság
- Jó általánosítóképesség
- Bíztató jelek az interpretálhatóságra vonatkozóan
- Nyitott kérdések?

Konklúzió

- Egyszerű modell, mégis pontos eredmények
- Nyelv,-és feladatközi robusztusság
- Jó általánosítóképesség
- Bíztató jelek az interpretálhatóságra vonatkozóan
- Nyitott kérdések?
  - Szószerkezet/mondat/bekezdés szintű reprezentációk