

# Ritka szóreprésentációk



# Szemantika

- Mi egy szó jelentése?

# Számítógépes Szemantika

- Mi egy szó jelentése?
- Hogy lehet számítógépesen ábrázolni?

# Számítógépes Szemantika

- Mi egy szó jelentése?
- Hogy lehet számítógépesen ábrázolni?
- Disztribúciós elmélet
  - Hasonló jelentésű szavak kontextusa is hasonló
    - A **kutya** **megkergette** a **macskát**.
    - A **ragadozó** **elejtette** a **prédát**.
    - A **gepárd** **beérte** az **antilopot**.

# Számítógépes Szemantika

- Mi egy szó jelentése?
- Hogy lehet számítógépesen ábrázolni?
- Disztribúciós elmélet
  - Hasonló jelentésű szavak kontextusa is hasonló
    - A **kutya** **megkergette** a **macskát**.
    - A **ragadozó** **elejtette** a **prédát**.
    - A **gepárd** **beérte** az **antilopot**.
  - > fél évszázados elmélet

# Számítógépes Szemantika

- Mi egy szó jelentése?
- Hogy lehet számítógépesen ábrázolni?
- Disztribúciós elmélet
  - Hasonló jelentésű szavak kontextusa is hasonló
    - A **kutya** **megkergette** a **macskát**.
    - A **ragadozó** **elejtette** a **prédát**.
    - A **gepárd** **beérte** az **antilopot**.
  - > fél évszázados elmélet (< fél évtizednyi hype)

# Versengő paradigmák

- 2013: word2vec megjelenése
  - Korábbi években is voltak már NN-alapú modellek

# Versengő paradigmák

- 2013: word2vec megjelenése
  - Korábbi években is voltak már NN-alapú modellek

***Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors***

**Marco Baroni and Georgiana Dinu and Germán Kruszewski**



# Versengő paradigmák

- 2013: word2vec megjelenése
  - Korábbi években is voltak már NN-alapú modellek

***Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors***

**Marco Baroni and Georgiana Dinu and Germán Kruszewski**

**Rehabilitation of Count-based Models for Word Vector Representations**

# Szóbeágyazások (implicit) célja

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

# Szóbeágyazások (implicit) célja

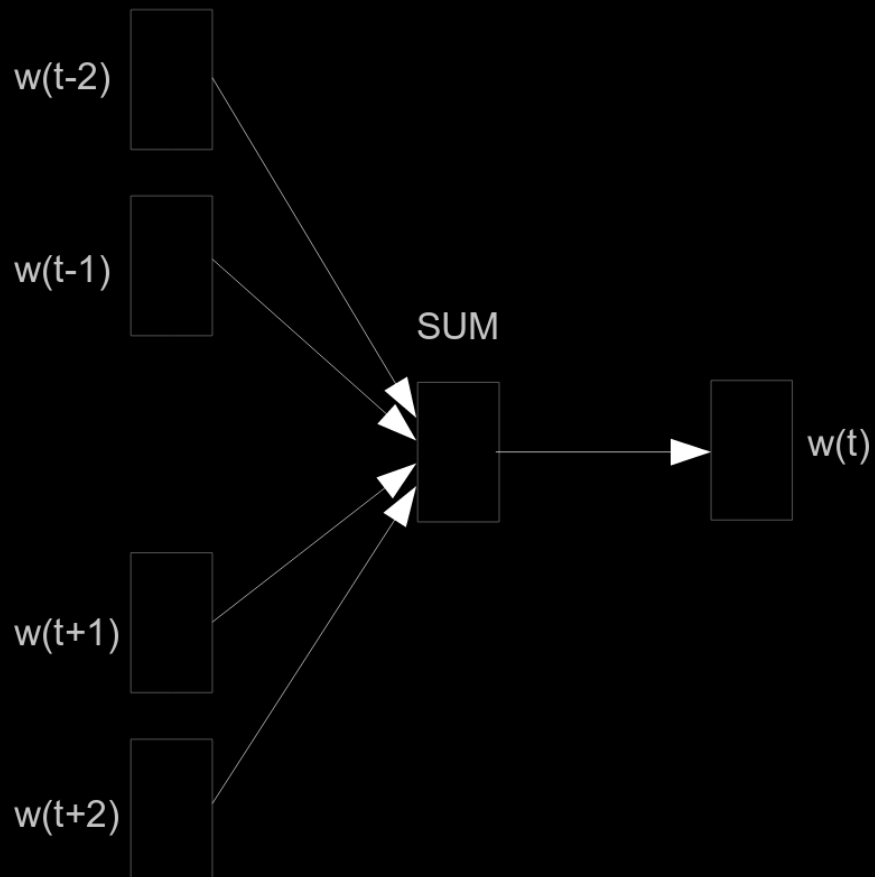
- Szavakhoz rendeljünk vektorokat, úgy, hogy azok tükrözzék a szópárok együttlőfordulási szokásait
  - Ha  $i$  és  $j$  szavak gyakran fordulnak elő, akkor a szavakhoz társított vektorok pontszorzata legyen nagy
    - Ellenkező esetben pedig kicsi

# Szóbeágyazások (implicit) célja

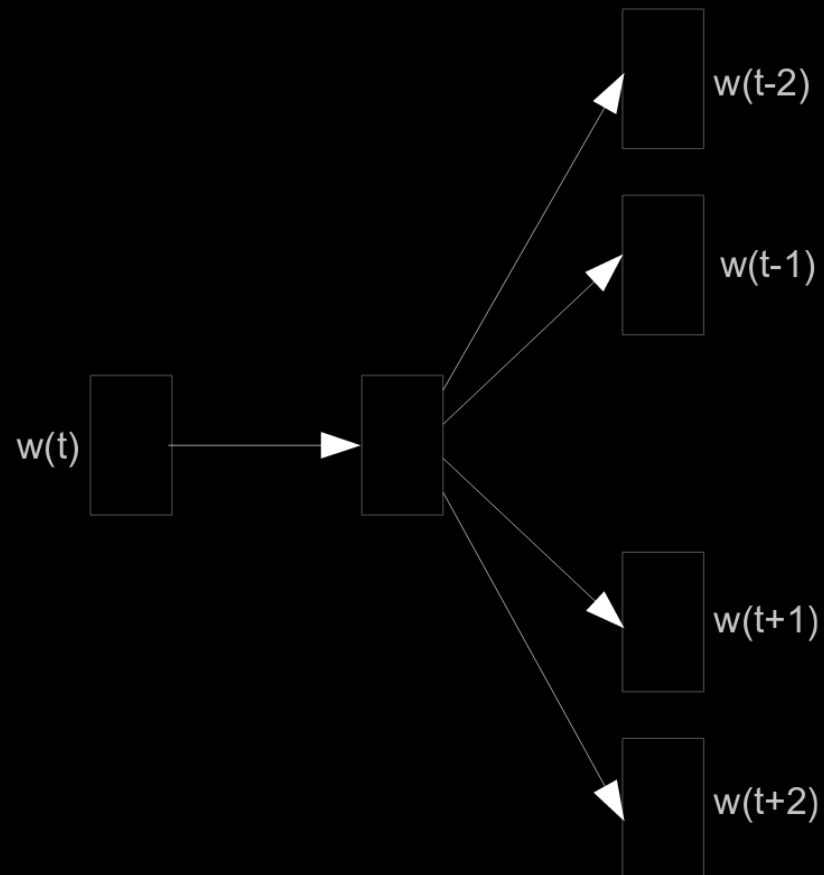
- Szavakhoz rendeljünk vektorokat, úgy, hogy azok tükrözzék a szópárok együttlőfordulási szokásait
  - Ha  $i$  és  $j$  szavak gyakran fordulnak elő, akkor a szavakhoz társított vektorok pontszorzata legyen nagy
    - Ellenkező esetben pedig kicsi
    - Pl. kutya=[3 2], macska=[2 4], gőzmozdony=[-2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

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

# word2vec variánsok



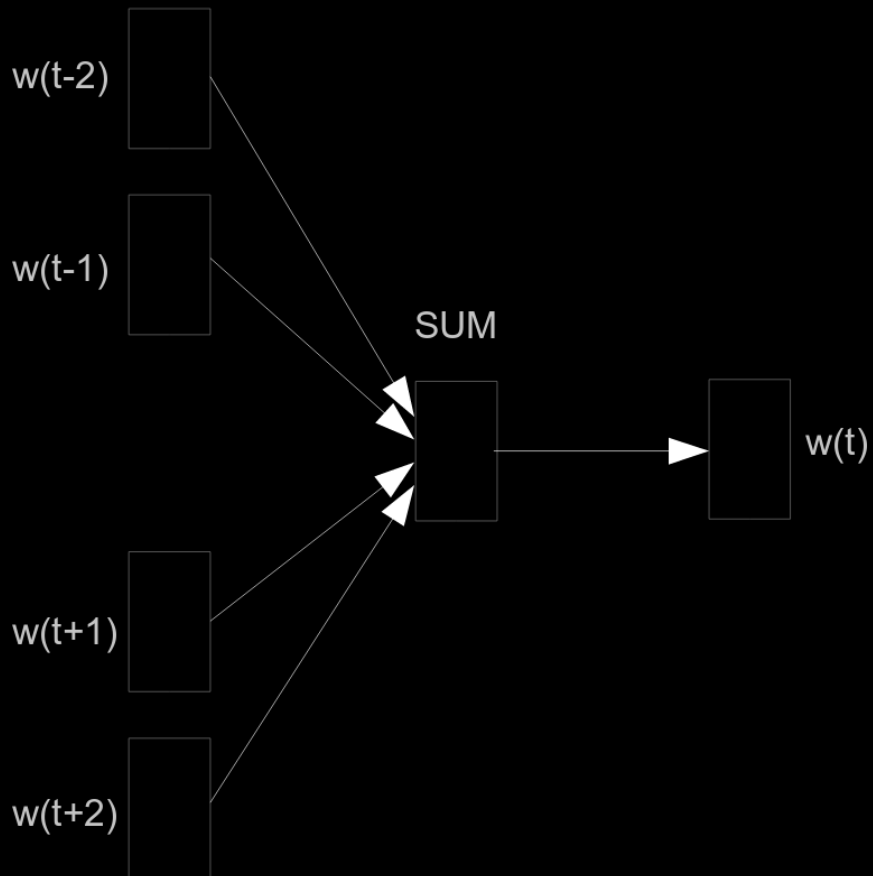
**CBOW**



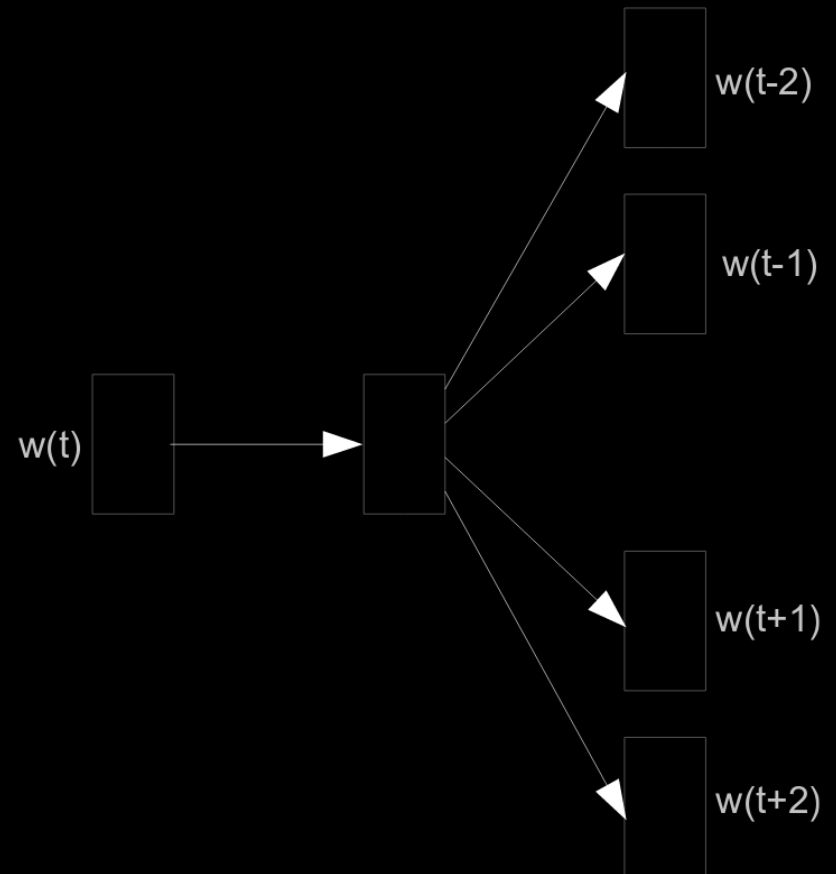
**Skip-gram**

# word2vec variánsok

*hogy az **X** záródnak !*



**CBOW**

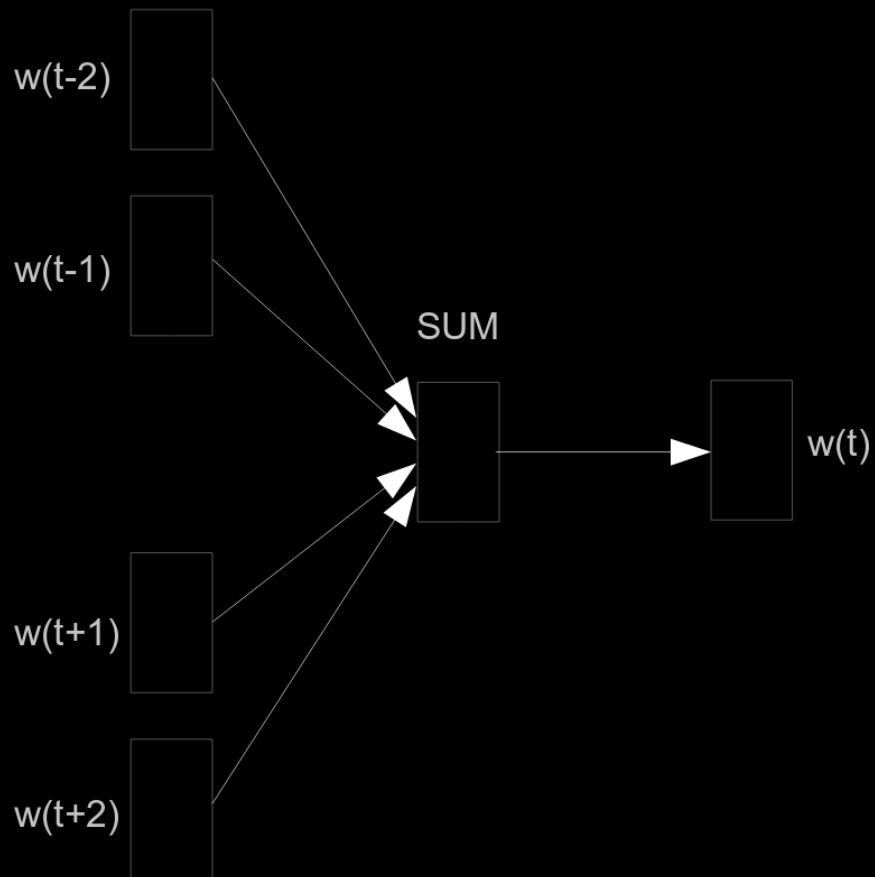


**Skip-gram**

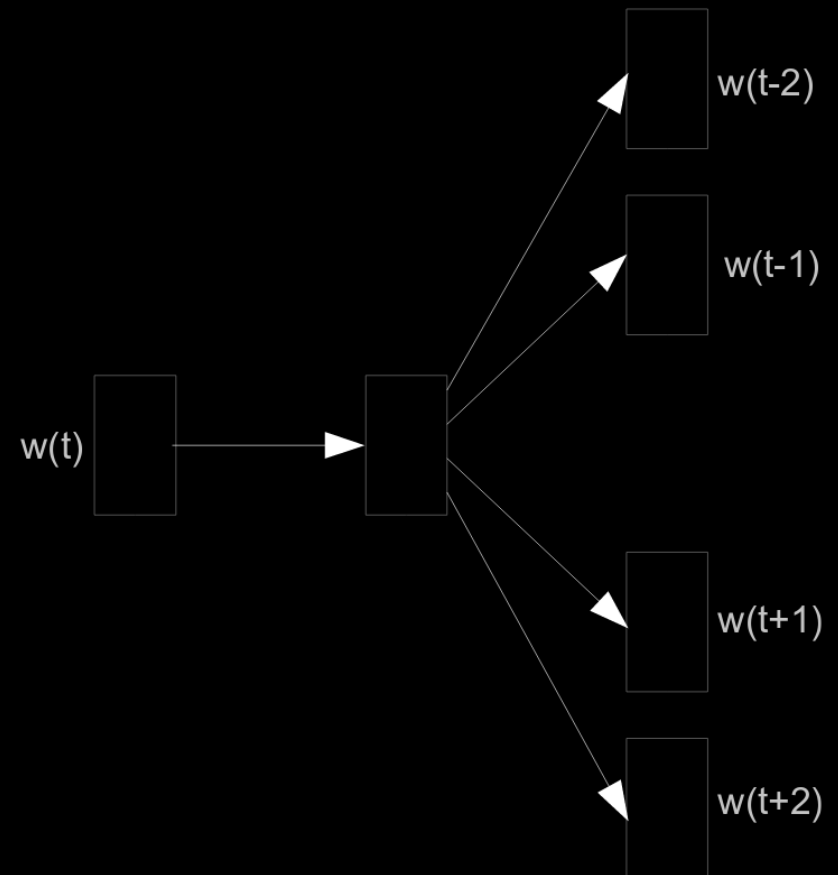
# word2vec variánsok

hogy az  $X$  záródnak !

$U$   $V$  ajtók  $Y$   $Z$



**CBOW**



**Skip-gram**

# word2vec célja

- Hasonló jelentésű input szavak kontextusukhoz illeszkedő outputot eredményezzenek

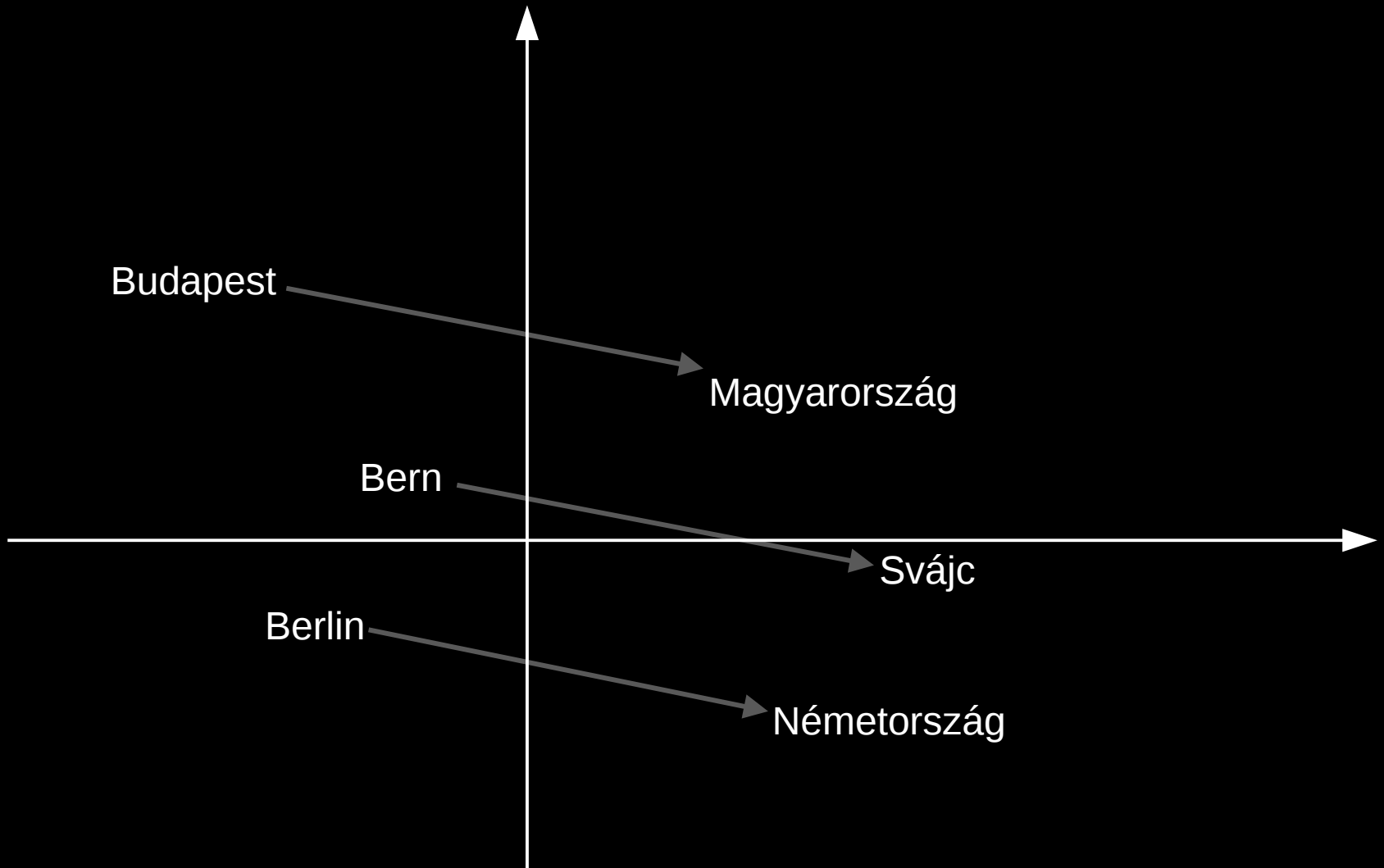
$$y(x) = \textit{softmax} \left( V \left( W 1_x \right) \right)$$

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



# Szóanalógiák

- $a:b::c:?$



# RepEval 2016

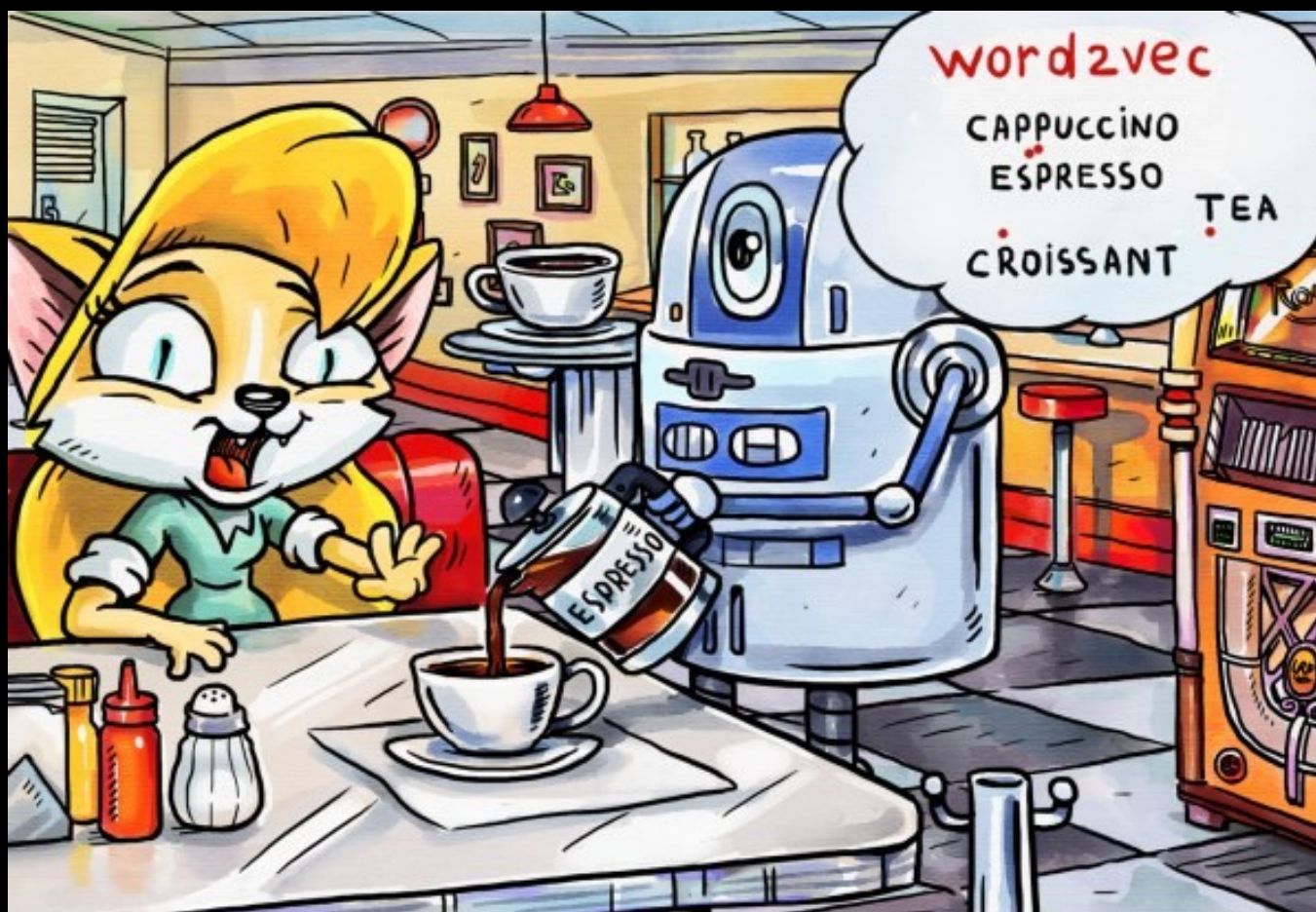
## Analysis Track

- **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

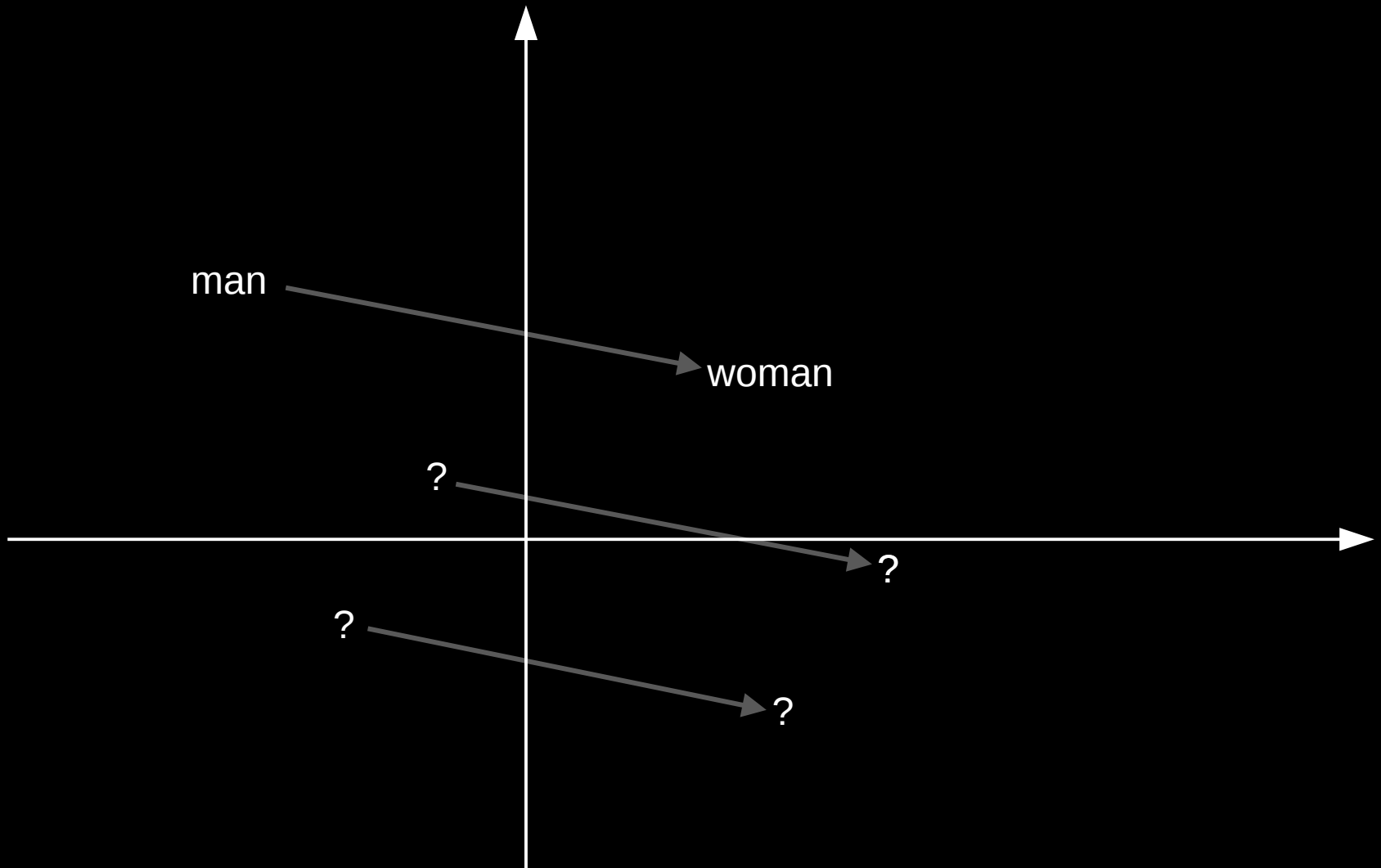
- **Problems** With Evaluation of Word Embeddings Using Word Similarity Tasks [pdf]  
*Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, Chris Dyer*
- **Intrinsic Evaluations of Word Embeddings: What Can We Do Better?** [pdf]  
*Anna Gladkova, Aleksandr Drozd*
- **Issues** in Evaluating Semantic Spaces 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.

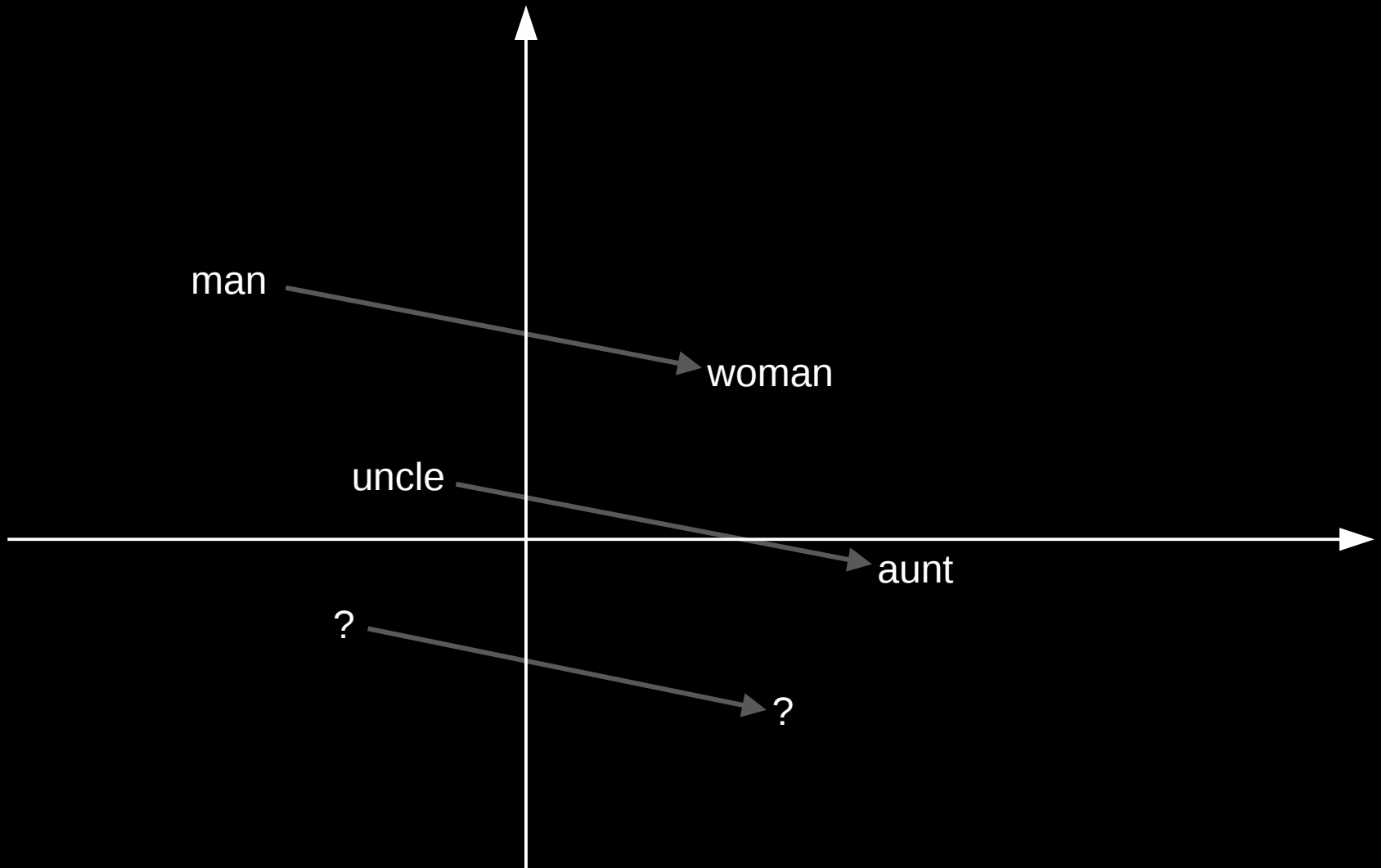
# Szóanalógiák újratöltve

- a:b:?:?



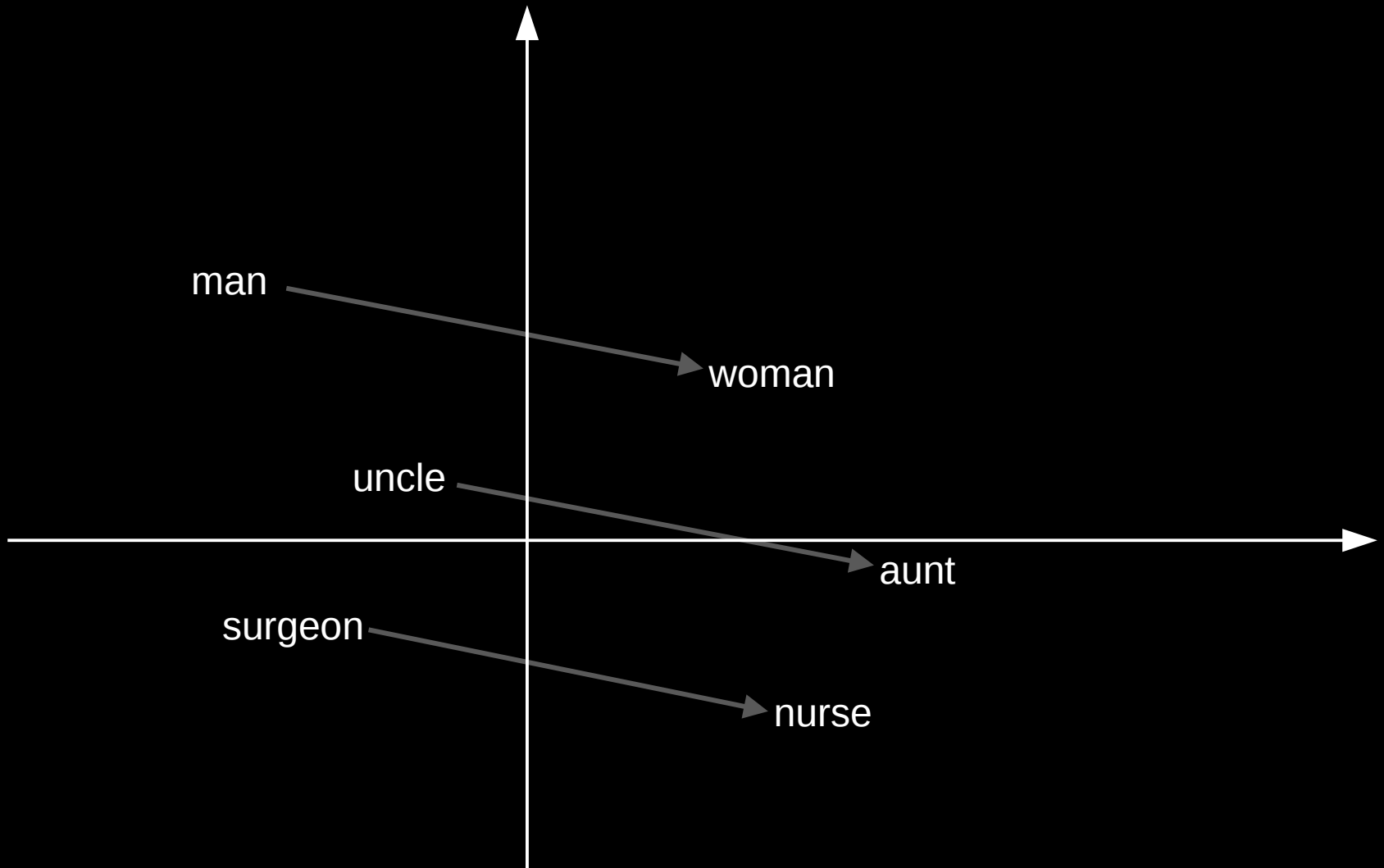
# Szóanalógiák újratöltve

- a:b:?:?



# Szóanalógiák újratöltve

- a:b:?:?



# Folytonos szóreprzentációk

alma  $[1\ 0\ 0\ 0\ \dots\ 0\ 0\ 0\ 0\ 0\ \dots\ 0] \longrightarrow [3.2\ -1.5]$

...

banán  $[0\ 0\ 0\ 0\ \dots\ 1\ 0\ 0\ 0\ 0\ \dots\ 0] \longrightarrow [2.8\ -1.6]$

...

lapát  $[0\ 0\ 0\ 0\ \dots\ 0\ 0\ 1\ 0\ 0\ \dots\ 0] \longrightarrow [-1.1\ 12.6]$

...

zebra  $[0\ 0\ 0\ 0\ \dots\ 0\ 0\ 0\ 0\ 0\ \dots\ 1] \longrightarrow [0.8\ 0.5]$



# Folytonos szóreprésentációk 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
  - In vivo vs. in vitro
  - Korlátozott interpretálhatóság

# Ritka & folytonos szóreprzentációk

alma [3.2 -1.5]  $\longrightarrow$  [ 0 1.7 0 0 -0.2 0 ]

...

banán [2.8 -1.6]  $\longrightarrow$  [ 0 1.1 0 0 -0.4 0 ]

...

lapát [-1.1 12.6]  $\longrightarrow$  [1.7 0 -2.1 0 0 -0.8]

...

zebra [0.8 0.5]  $\longrightarrow$  [ 0 0 1.3 0 -1.2 0 ]

# Ritka folytonos reprezentációk tanulása

- Adott  $x_i$  ( $i=1, \dots, |V|$ ) szóbeágyazások esetén

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

Beágyazás  
vektor ( $\in \mathbb{R}^L$ )

Szótár  
( $\in \mathbb{R}^{L \times K}$ )

Ritka  
együtthetők

# Ritka folytonos reprezentációk tanulása

- Adott  $x_i$  ( $i=1, \dots, |V|$ ) 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  
( $\in \mathbb{R}^{L \times K}$ )

Ritka  
együtthatók

Regularizációs  
tag

# Ritka folytonos reprezentációk tanulása

- Adott  $x_i$  ( $i=1, \dots, |V|$ ) szóbeágyazások esetén

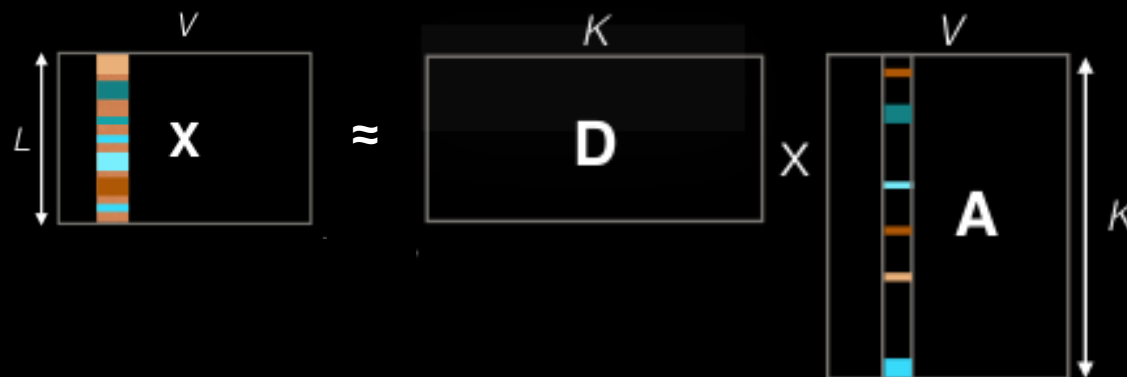
$$\min_{D \in C, \alpha} \sum_{i=1}^{|V|} \|x_i - D \alpha_i\|_2^2 + \lambda \|\alpha_i\|_1$$

Beágyazás  
vektor ( $\in \mathbb{R}^L$ )

Szótár  
( $\in \mathbb{R}^{L \times K}$ )

Ritka  
együtthetők

Regularizációs  
tag



# Szófaji kódolás (POS tagging) feladata

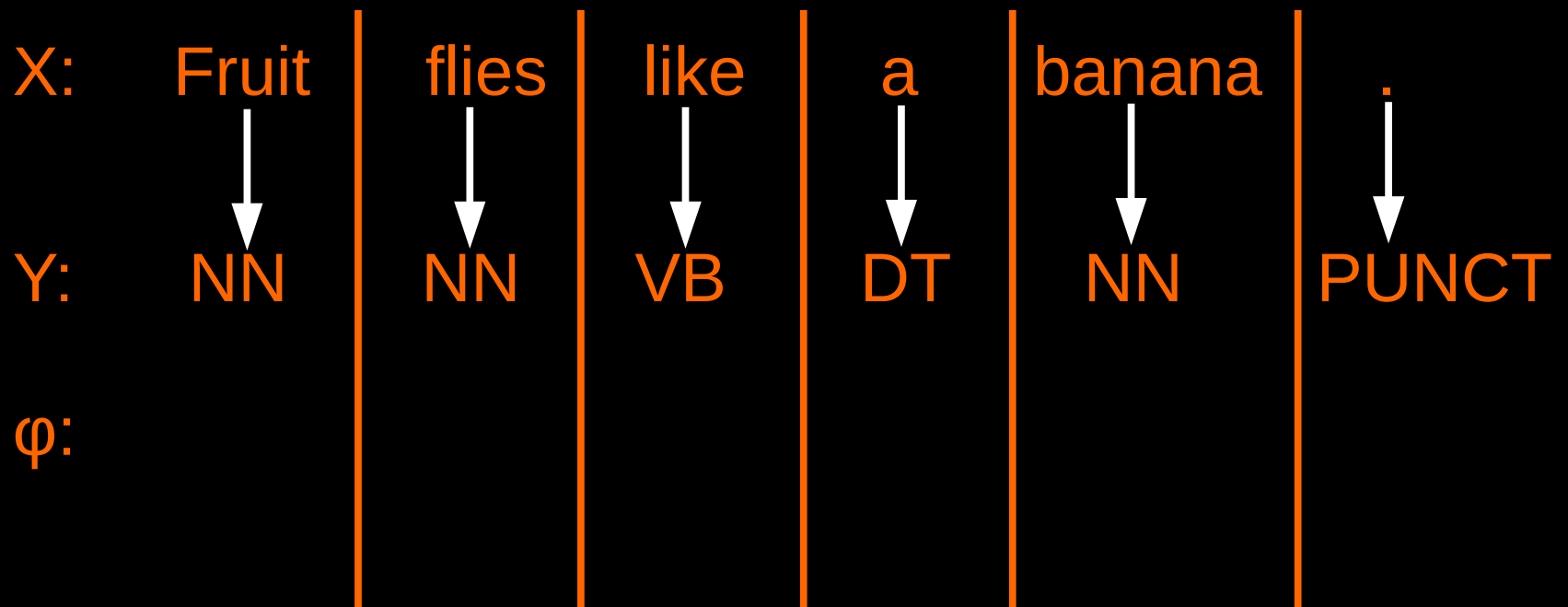
- Adott természetes nyelvű tokensorozatra határozzuk meg az egyes tokenek szófaját
  - Tipikus *in vivo* kiértékelés

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  $\phi_j$  jellemzőket
  - $\phi_j$  vizsgálhatja pl. egy szó felszíni jegyeit (prefixum/suffixum), de a szókörnyezetét is



# “Klasszikus” szekvenciajelölő

- Minden szóhoz számítsunk ki  $\varphi_j$  jellemzőket
  - $\varphi_j$  vizsgálhatja pl. egy szó felszíni jegyeit (prefixum/suffixum), de a szókörnyezetét is

X:	Fruit	flies	like	a	banana	.
	↓	↓	↓	↓	↓	↓
Y:	NN	NN	VB	DT	NN	PUNCT
$\varphi$ :	pre2=Fr suf2=it	pre2=fl suf2=es	pre2=li suf2=ke	pre2=a suf2=a	pre2=ba suf2=na	pre2=. suf2=.



# “Klasszikus” szekvenciajelölő

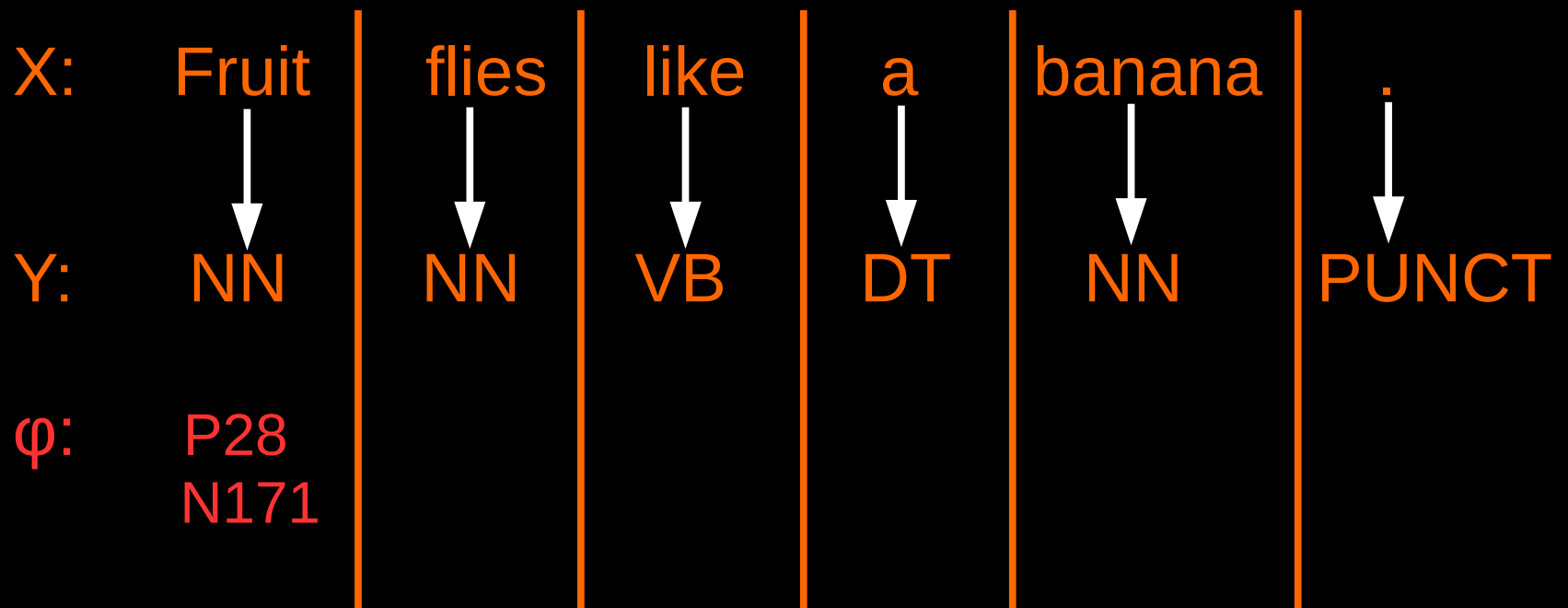
- Minden szóhoz számítsunk ki  $\varphi_j$  jellemzőket
  - $\varphi_j$  vizsgálhatja pl. egy szó felszíni jegyeit (prefixum/suffixum), de a szókörnyezetét is

X:	Fruit	flies	like	a	banana	.
	↓	↓	↓	↓	↓	↓
Y:	NN	NN	VB	DT	NN	PUNCT
$\varphi$ :	pre2=Fr suf2=it ...	pre2=fl suf2=es ...	pre2=li suf2=ke ...	pre2=a suf2=a ...	pre2=ba suf2=na ...	pre2=. suf2=. ...

# Szekvenciajelölő **ritka** jellemzőkkel

- Használjuk a dekompozícióból jövő  $\alpha$ -t

– Pl.  $\vec{Fruit} \approx 1.1 \cdot \vec{d}_{28} - 0.4 \cdot \vec{d}_{171}$



# Szekvenciajelölő **ritka** jellemzőkkel

- Használjuk a dekompozícióból jövő  $\alpha$ -t

– Pl.  $\overrightarrow{Fruit} \approx 1.1 \cdot \overrightarrow{d_{28}} - 0.4 \cdot \overrightarrow{d_{171}}$

X:	Fruit	flies	like	a	banana	.
	↓	↓	↓	↓	↓	↓
Y:	NN	NN	VB	DT	NN	PUNCT
$\varphi$ :	P28 N171	P77 P88 ...	N11 N62 ...	N88 N40 ...	P28 N210 ...	N21 P67 ...

# Többnyelvű eredmények

- Eredmények kiátlagolva 12 nyelvre

	Sűrű	R i t k a
polyglot	91.17%	94.44%
CBOW	88.30%	93.74%
SG	86.89%	93.63%
Glove	81.53%	91.92%

- Fő észrevételek
  - polyglot > CBOW > SG > Glove

# Többnyelvű eredmények

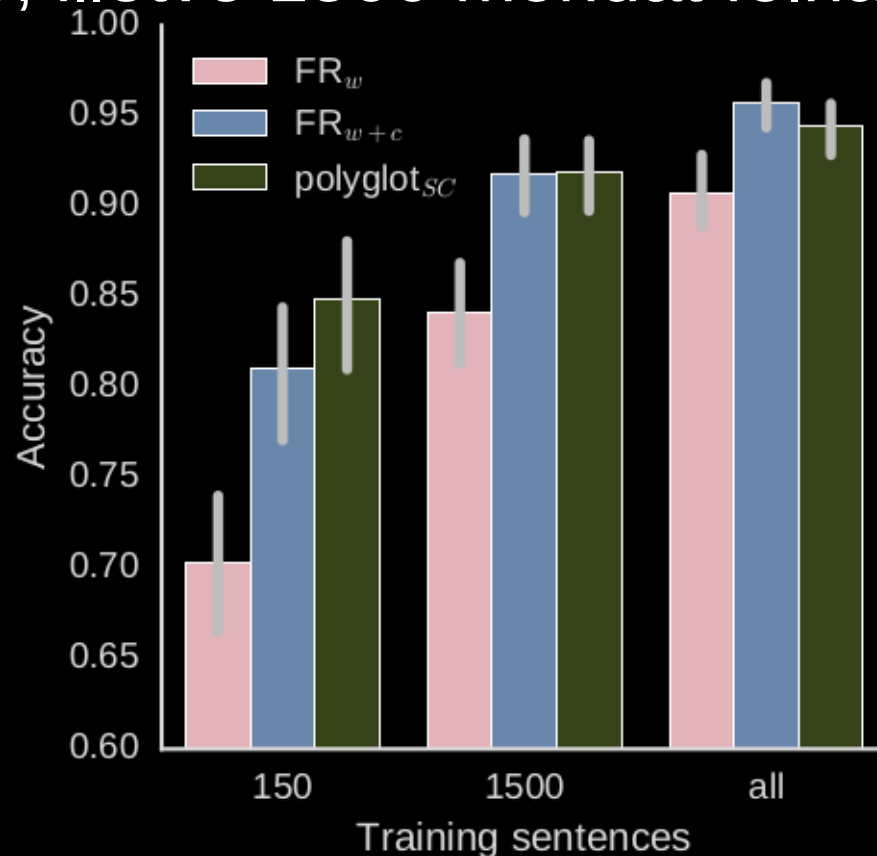
- Eredmények kiátlagolva 12 nyelvre

	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

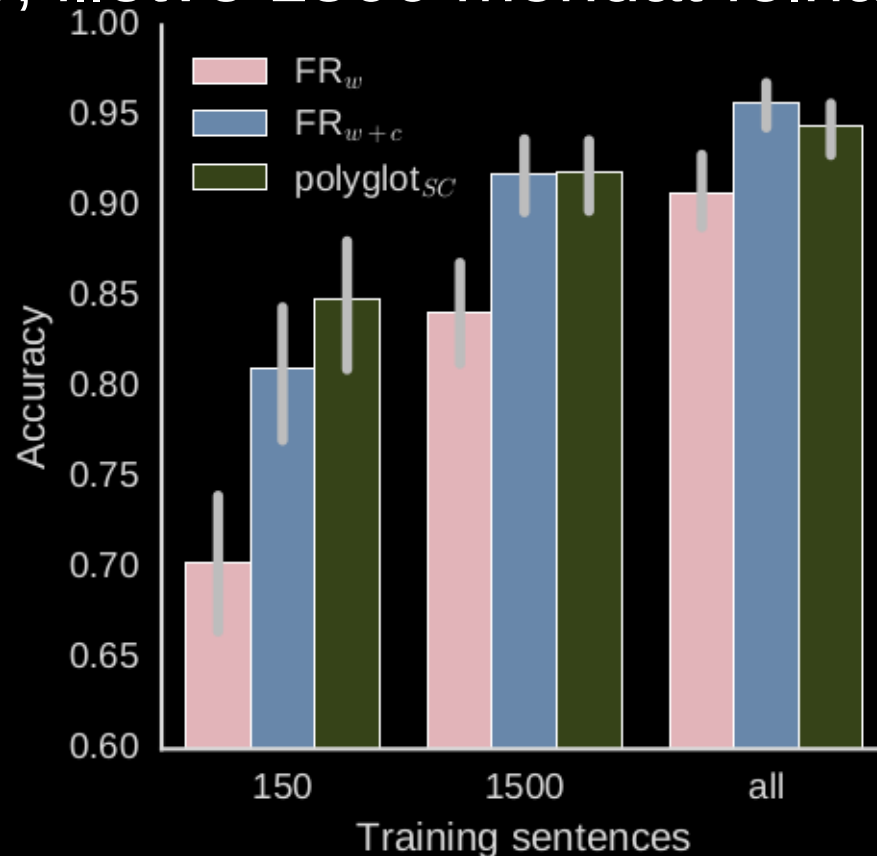
# Általánosítóképesség

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



# Általánosítóképesség

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



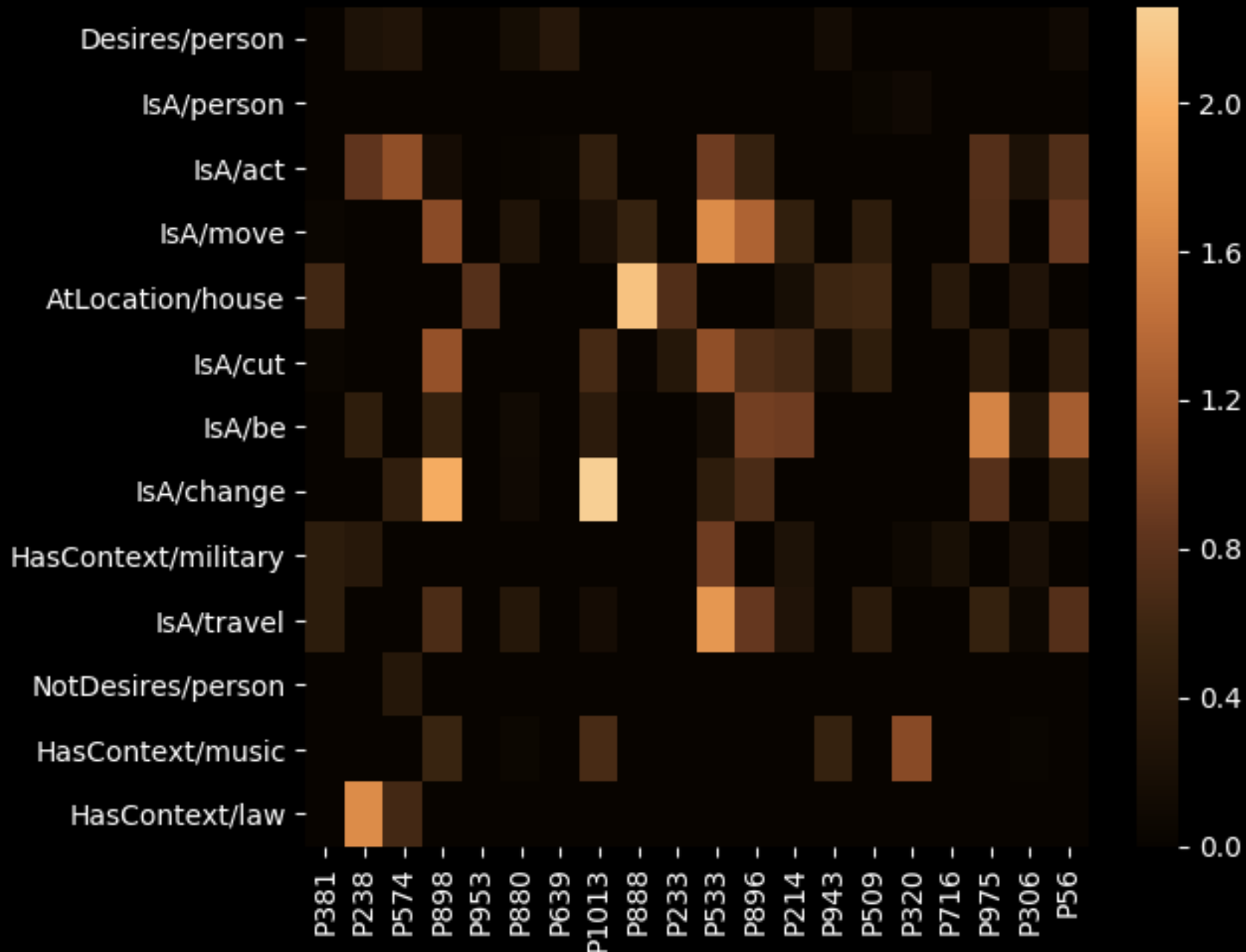
- Tulajdonnévfelismerésen is tesztelve

# 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 szóreprzentációk 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 szóreprzentációk 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 szóreprzentációk

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

- “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 szóreprzentációk

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

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

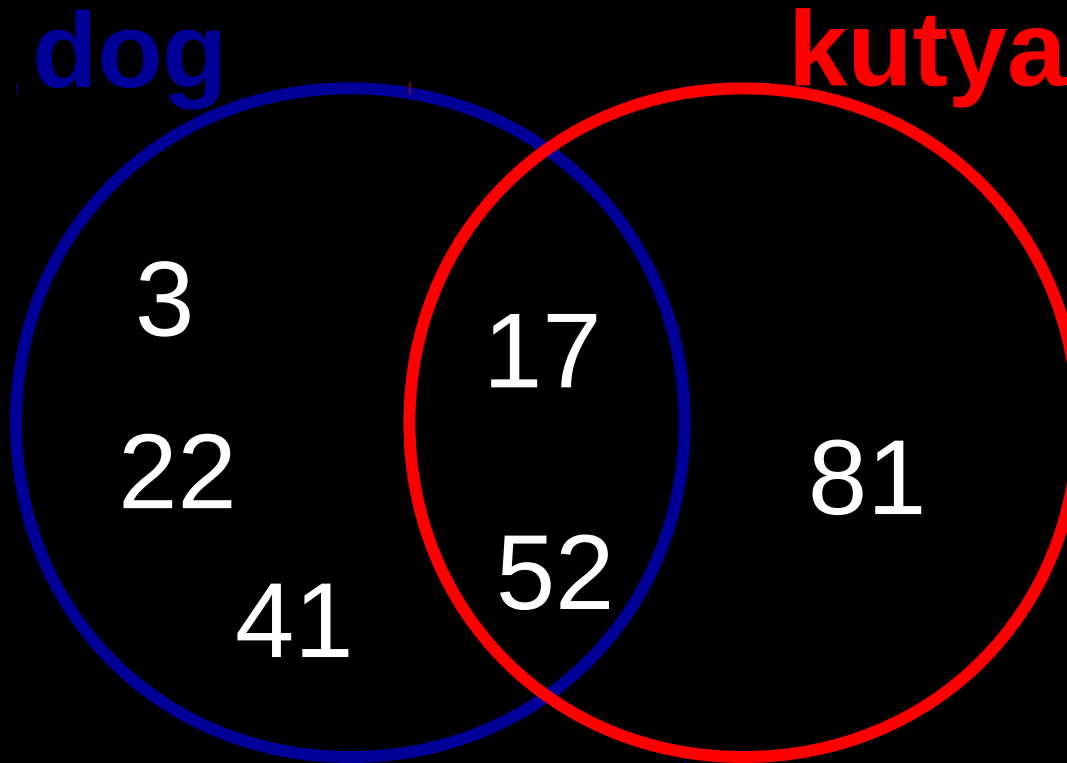
# Többnyelvű ritka szóreprzentációk

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

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

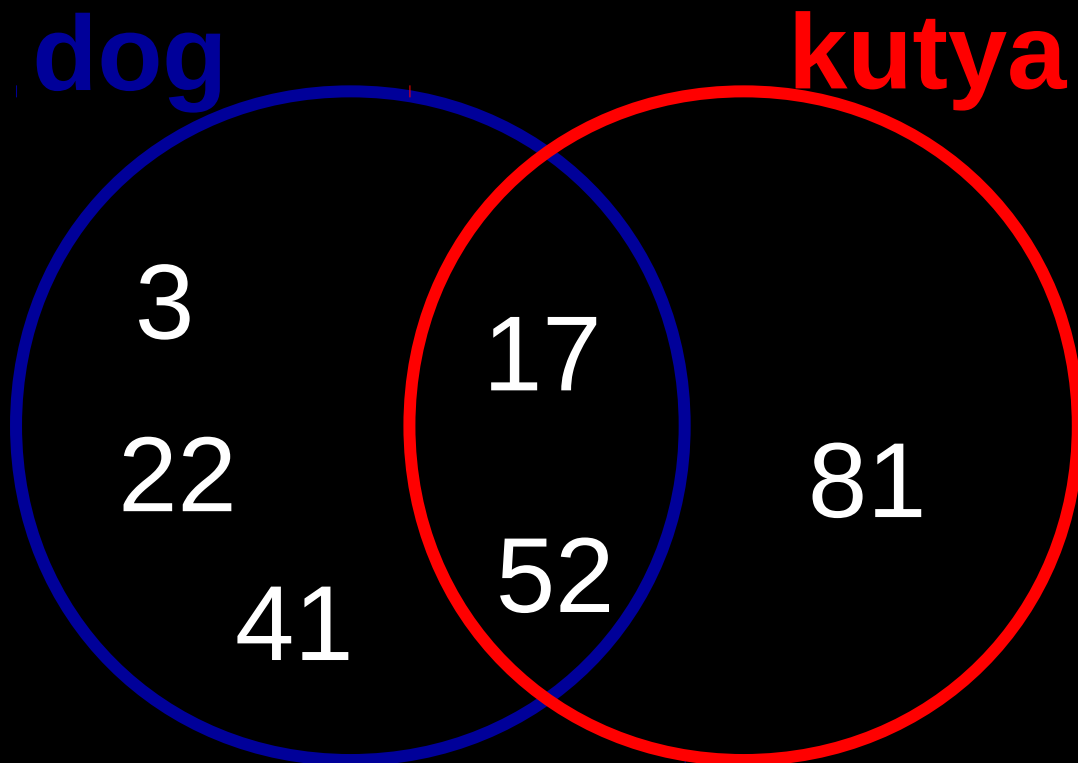
$$\alpha^{(c)} = \min_{\alpha} \sum_{i=1}^{|V|} \|M w_i^{(c)} - D^{(f)} \alpha_i\|_2^2 + \lambda \|\alpha_i\|_1$$

# Ritka reprezentációk átfedése





# Ritka reprezentációk átfedése

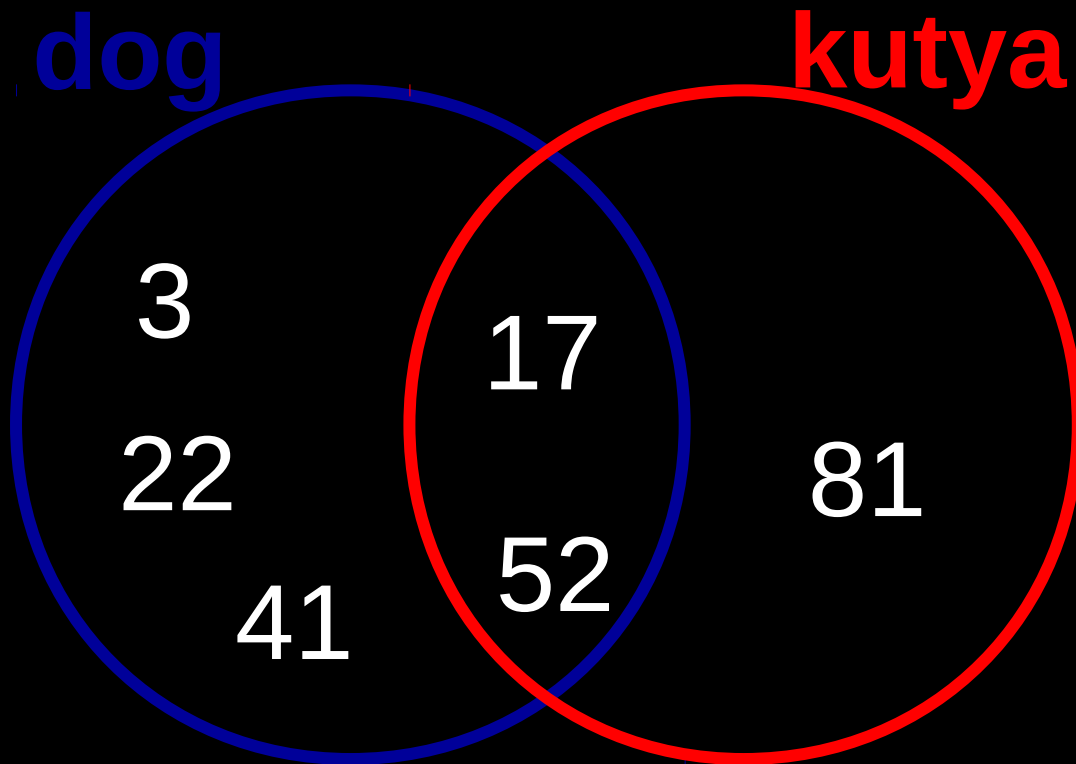


$$P=2/3$$

$$R=2/5$$

$$F=1/2$$

# Ritka reprezentációk átfedése



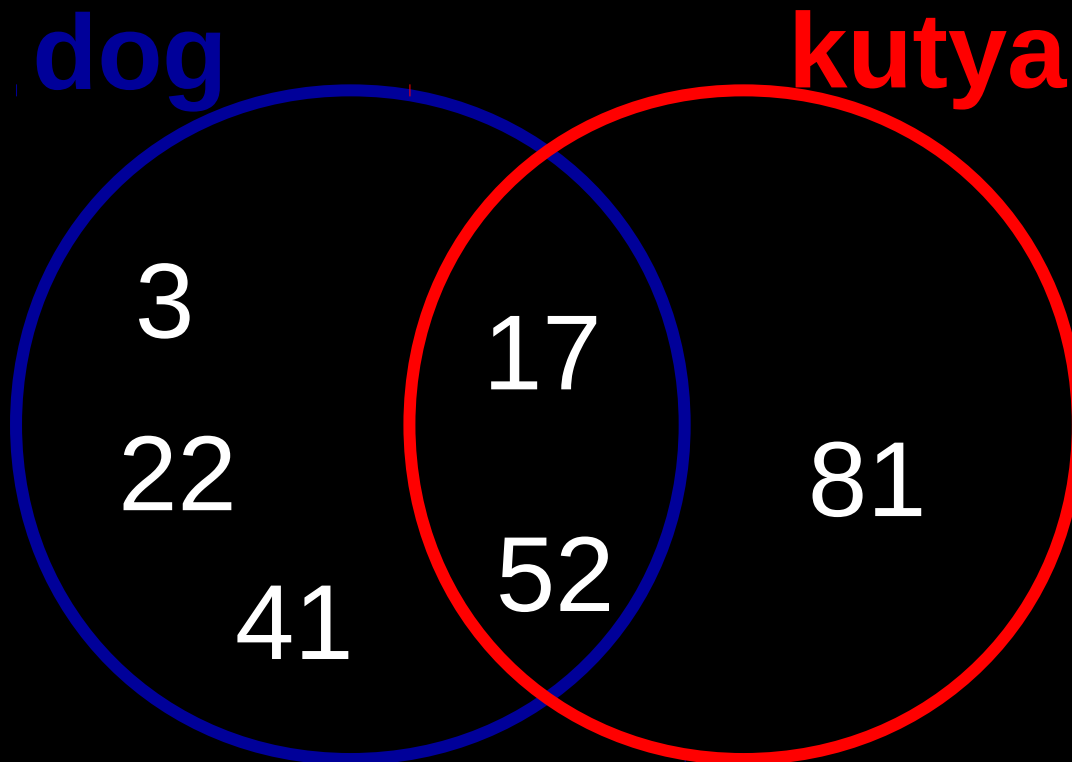
$$P=2/3$$

$$R=2/5$$

$$F=1/2$$

$\lambda$	Prec.	Rec.	F						
0.1	0.023	0.024	<b>0.024</b>						
0.3	0.001	0.001	0.001						
0.5	0.000	0.000	0.000						
	M=I								

# Ritka reprezentációk átfedése



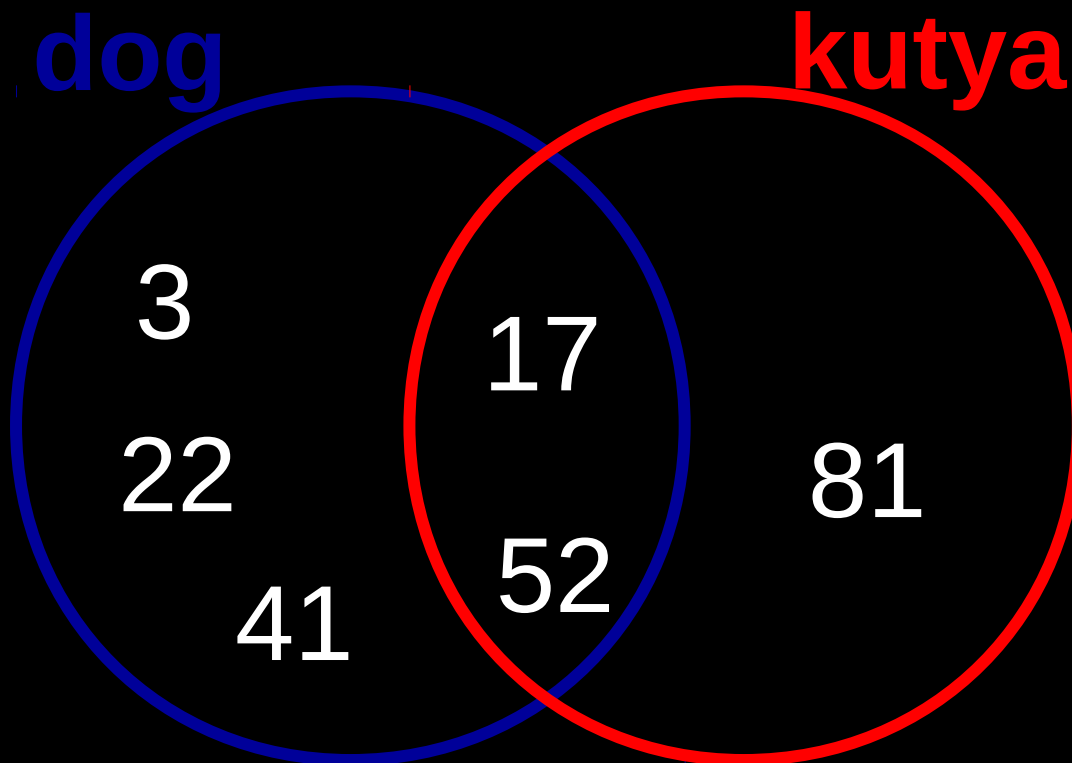
$$P=2/3$$

$$R=2/5$$

$$F=1/2$$

$\lambda$	Prec.	Rec.	F	Prec.	Rec.	F			
0.1	0.023	0.024	<b>0.024</b>	0.170	0.117	0.139			
0.3	0.001	0.001	0.001	0.345	0.118	<b>0.176</b>			
0.5	0.000	0.000	0.000	0.600	0.009	0.018			
	M=I			Tetszőleges M					

# Ritka reprezentációk átfedése



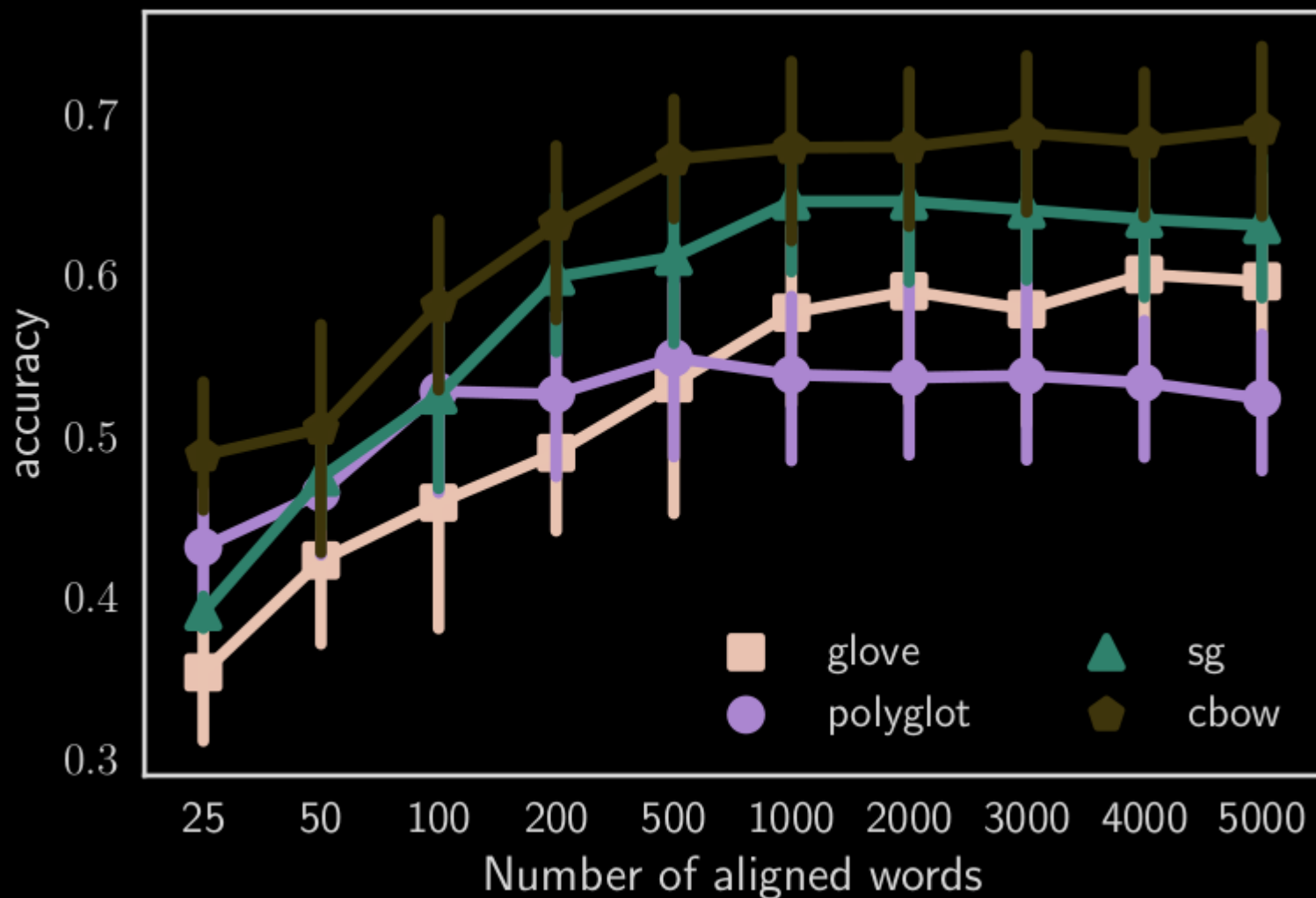
$$P=2/3$$

$$R=2/5$$

$$F=1/2$$

$\lambda$	Prec.	Rec.	F	Prec.	Rec.	F	Prec.	Rec.	F
0.1	0.023	0.024	<b>0.024</b>	0.170	0.117	0.139	0.098	0.137	0.114
0.3	0.001	0.001	0.001	0.345	0.118	<b>0.176</b>	0.167	0.208	0.185
0.5	0.000	0.000	0.000	0.600	0.009	0.018	0.271	0.202	<b>0.232</b>
	M=I			Tetszőleges M			Ortogonalis M		

# Nyelvközi szófaji kódolás eredményei



# Konklúzió

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

# Konklúzió

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