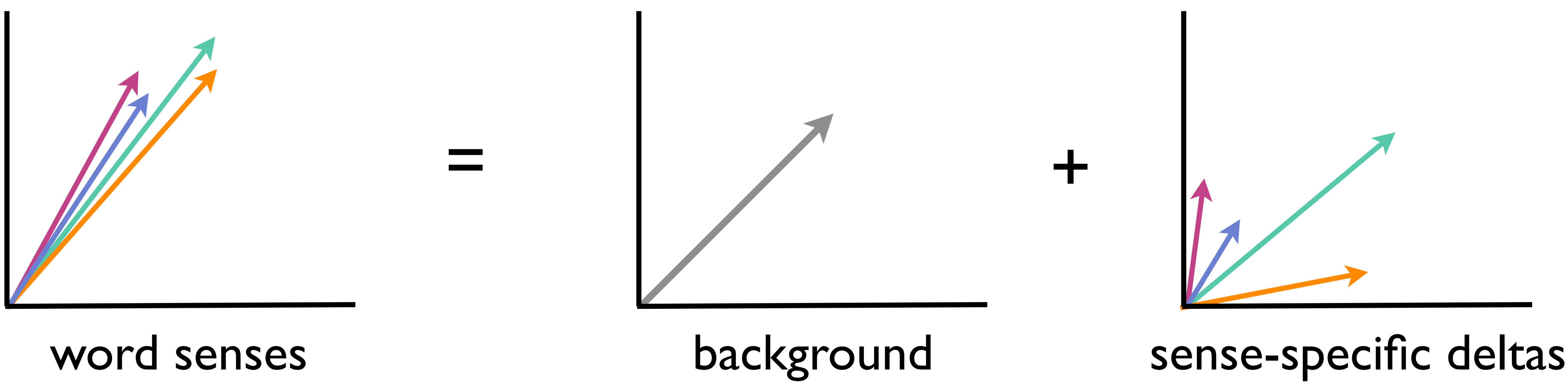
A Mixture Model with Sharing for Lexical Semantics

Joseph Reisinger and Raymond Mooney
The University of Texas at Austin

Outline

- Distributional lexical semantics
- Inadequacies of standard “centroid” representation
- Richer representations of homonymy using mixture models
- A “tiered clustering” model for polysemy:



Distributional Lexical Semantics

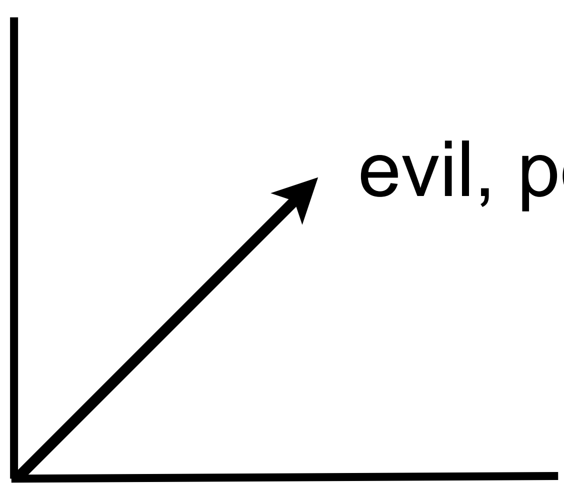
- Represent "meaning" as a point/vector in a high-dimensional space
- Word relatedness correlates with some distance metric
- Attributional / relational / resource-based, Almuhareb and Poesio (2004), Baroni and Lenci (2009), Bullinaria and Levy (2007), Erk (2007), Griffiths et al. (2007), Landauer and Dumais (1997), Moldovan (2006), Padó and Lapata (2007), Pantel and Pennacchiotti (2006), Sahlgren (2006), Turney and Pantel (2010)

(Some of) Distributional Lexical Semantics

word occurrences / context vectors

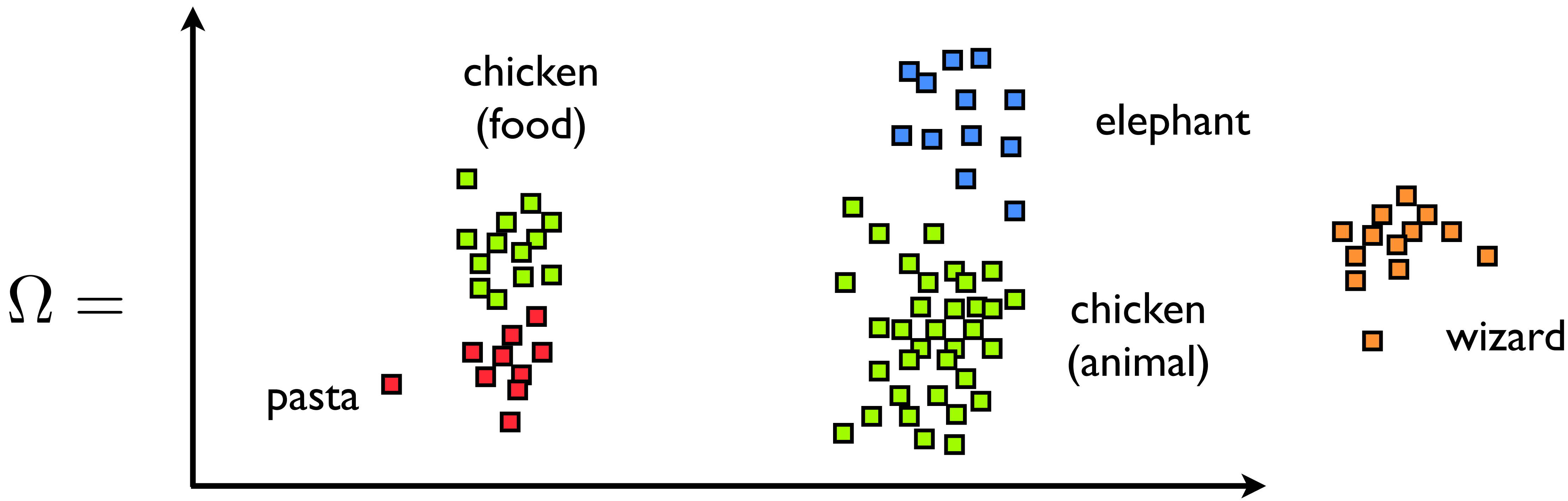
The history of **Oz** prior to The **Wonderful** Wizard of **Oz** (often **called** the
Rowling describes the **beloved** wizard **Dumbledore** as **Machiavellian** and **says**
Merlin is a **legendary** figure best known as the wizard featured in the **Arthurian legend**
A wizard **comedian** is **known** to have **survived eating** this **plant** on a bet, though he is still purple
True Image is **known** for its **simple**, wizard **driven interface**, and **received positive**
Thunder did a **cover of "Pinball** Wizard" to be featured on "Hollywood Rocks"

compute centroid



evil, powerful, magic, wizard, Merlin, spells, Harry

One Word One Prototype

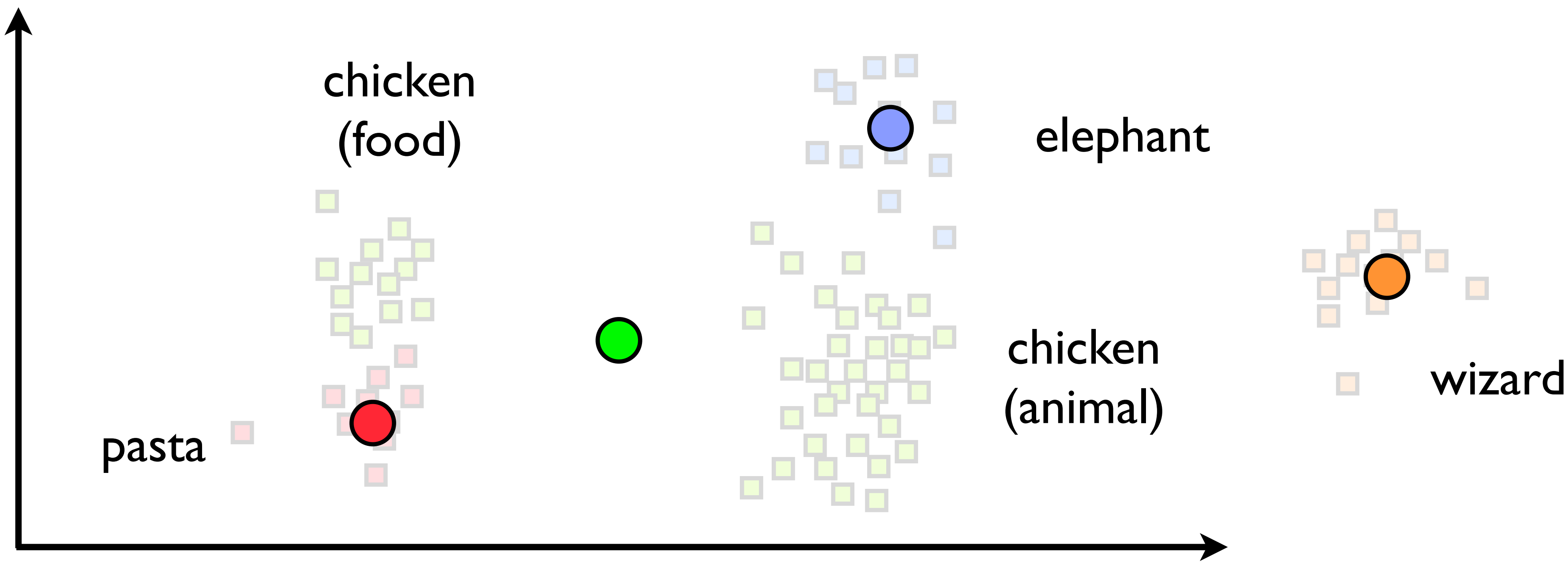


- Find the centroid of the individual word occurrence context vectors
- Conflates senses / doesn't reliably account for thematic variability in usage

e.g., Schütze (1997)

One Word One Prototype

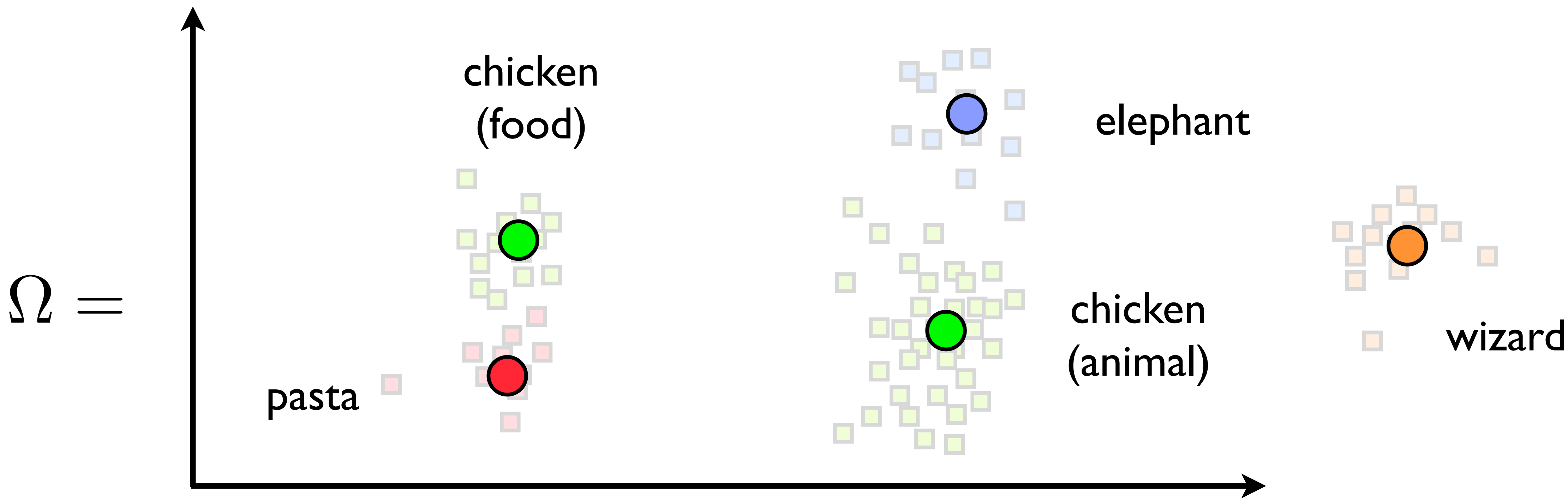
$\Omega =$



- Find the centroid of the individual word occurrence context vectors
- Conflates senses / doesn't reliably account for thematic variability in usage

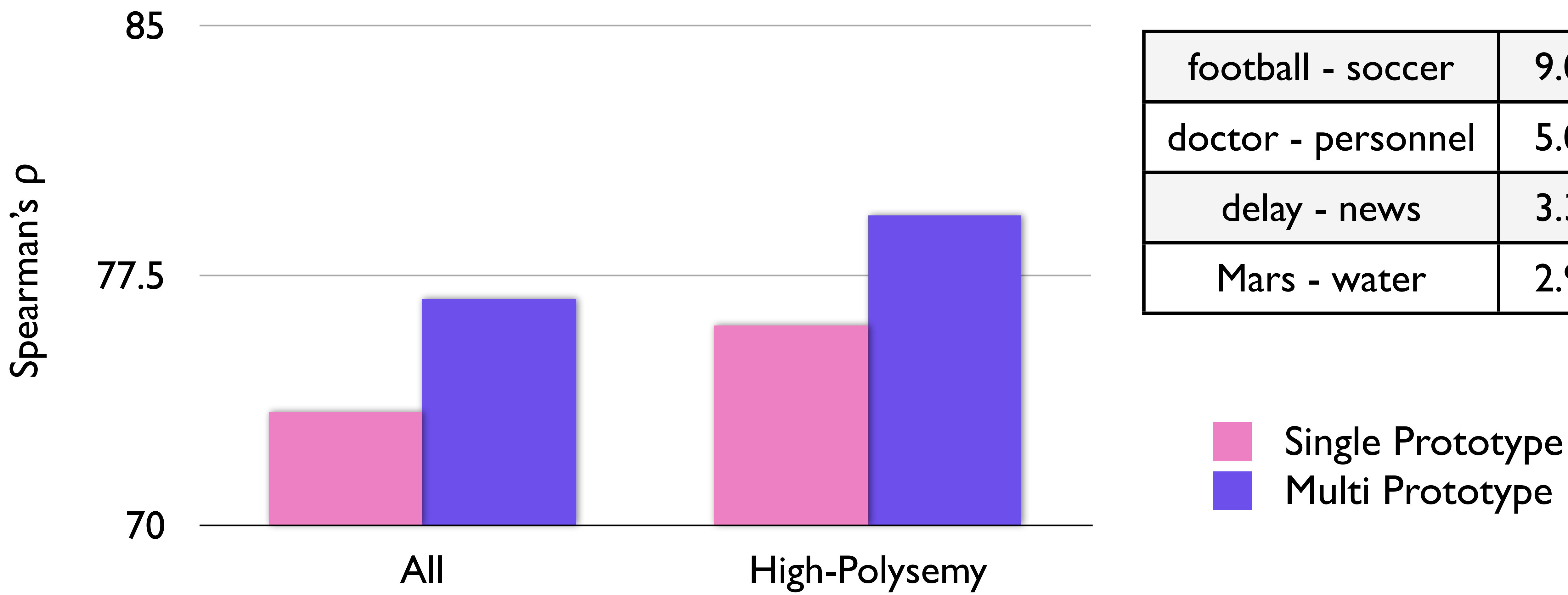
e.g., Schütze (1997)

Multi-Prototype



- Use a mixture model to cluster occurrences of each target word separately.
- Doesn't find lexicographic senses; captures contextual variance directly.

WS-353 Correlation with Human Relatedness Judgements

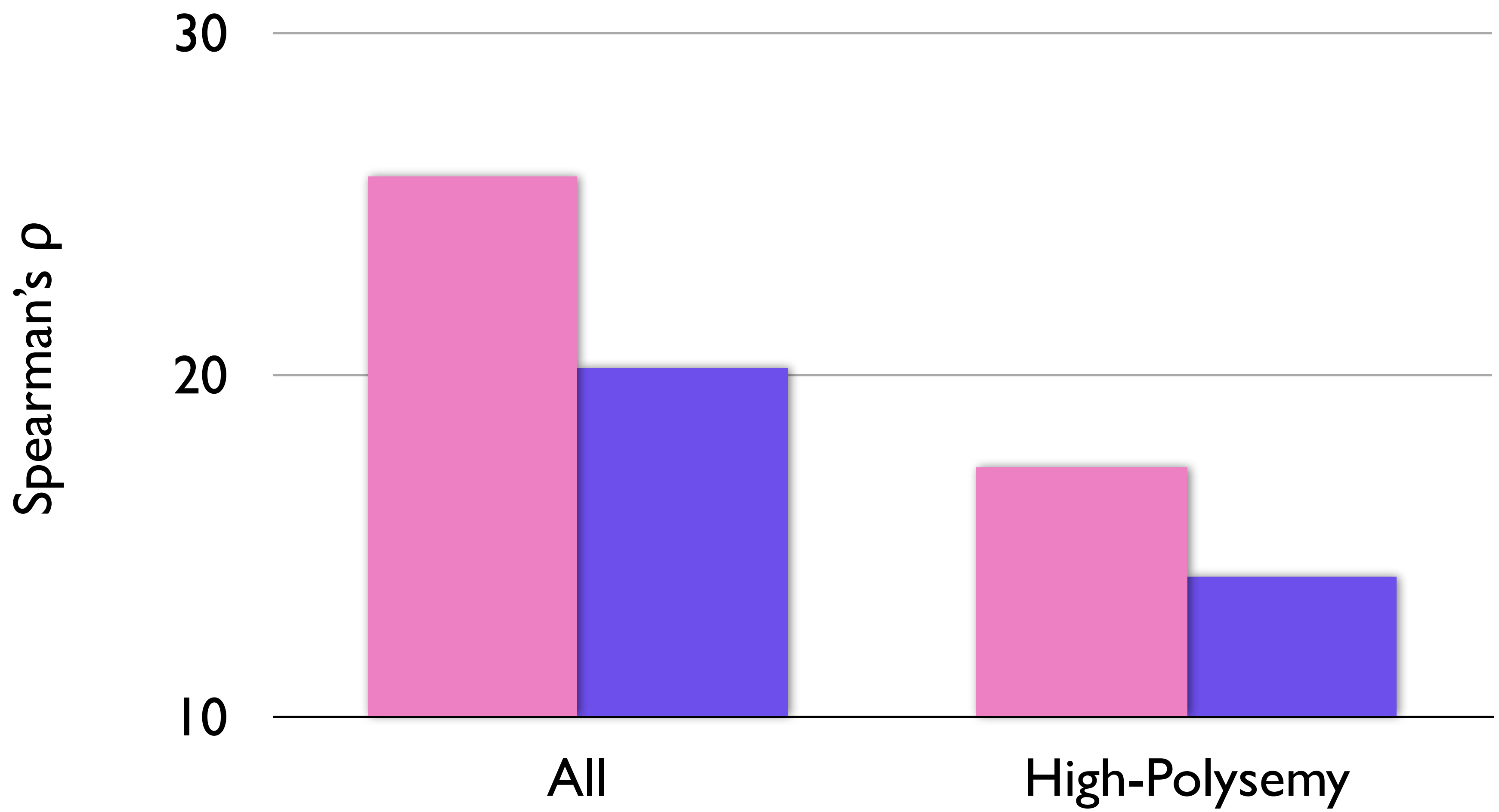


football - soccer	9.03
doctor - personnel	5.00
delay - news	3.31
Mars - water	2.94

- 353 words, biased towards pairs with high similarity
- High polysemy subset generated by counting WN senses

Dataset: Finkelstein et. al 2001

“Shared Structure Matters”



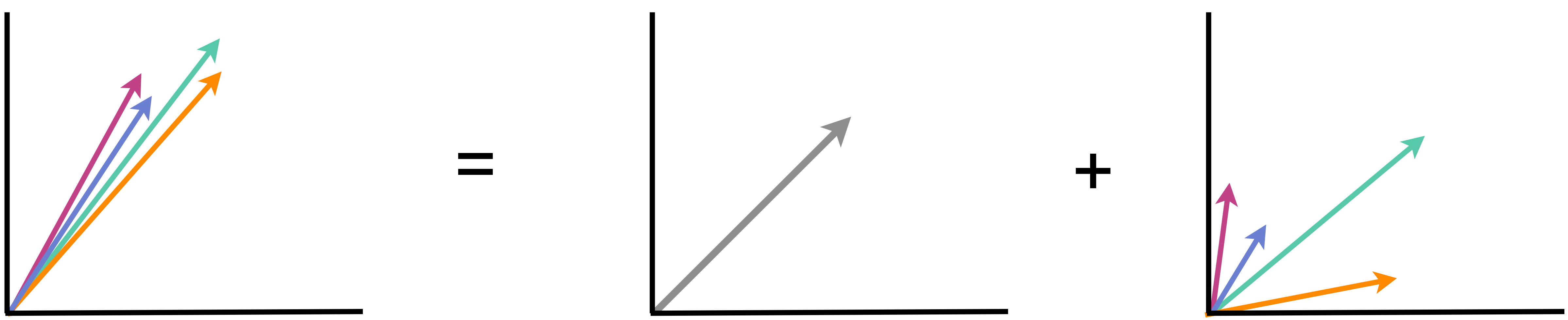
brother hit	4.7
hit brother	4.1
quake hit	3.4
hit quake	1.2

- Selectional Preference: Predict typical arguments for verbs (e.g. things that can eat or things that can be shot)
- Background cluster captures commonalities between argument fillers.

Dataset: Padó et. al 2007

Tiered Clustering

- There are cases where is a high degree of overlap between "senses"
  - e.g. polysemous words like line or raise
  - or i.e. in selectional preference where argument fillers may have some common structure
- MP models homonymy but not polysemy- it cannot account for "shared" structure
- Tiered Clustering introduces a background component to account for this



word senses = background + sense-specific deltas

Tiered Clustering

Tiered Clustering

ϕ_1
Merlin King Arthur Arthurian

$\sim \mathrm{Dir}(\beta)$

Tiered Clustering

ϕ_1

Merlin
King
Arthur
Arthurian

ϕ_2 $\sim \mathrm{Dir}(\beta)$

fairy
wicked
scene
tale

Tiered Clustering

ϕ_1	ϕ_2	ϕ_3
Merlin	fairy	Harry
King	wicked	Potter
Arthur	scene	Voldemort
Arthurian	tale	Dumbledore

$\sim \mathrm{Dir}(\beta)$

Tiered Clustering

ϕ_1	ϕ_2	ϕ_3
Merlin	fairy	Harry
King	wicked	Potter
Arthur	scene	Voldemort
Arthurian	tale	Dumbledore

$\sim \mathrm{Dir}(\beta)$

ϕ_b
evil
powerful
magic
wizard

$\sim \mathrm{Dir}(\beta_{\mathrm{background}})$

Tiered Clustering

ϕ_1	ϕ_2	ϕ_3
Merlin	fairy	Harry
King	wicked	Potter
Arthur	scene	Voldemort
Arthurian	tale	Dumbledore

ϕ_b
evil
powerful
magic
wizard

Rowling describes the **beloved** **wizard** **Dumbledore** as **Machiavellian** and **says**

Tiered Clustering

ϕ_1
Merlin
King
Arthur
Arthurian

ϕ_2
fairy
wicked
scene
tale

ϕ_3
Harry
Potter
Voldemort
Dumbledore

ϕ_b
evil
powerful
magic
wizard

Rowling describes the **beloved wizard Dumbledore** as **Machiavellian** and **says**

Tiered Clustering

ϕ_1

Merlin
King
Arthur
Arthurian

ϕ_2

fairy
wicked
scene
tale

ϕ_3

Harry
Potter
Voldemort
Dumbledore

ϕ_b

evil
powerful
magic
wizard

Rowling describes the beloved <u>wizard</u> **Dumbledore** as **Machiavellian** and **says**

Tiered Clustering

ϕ_1

Merlin
King
Arthur
Arthurian

ϕ_2

fairy
wicked
scene
tale

ϕ_3

Harry
Potter
Voldemort
Dumbledore

ϕ_b

evil
powerful
magic
wizard

ϕ_3

Rowling describes the beloved wizard **Dumbledore** as **Machiavellian** and **says**

Tiered Clustering

ϕ_1
Merlin King Arthur Arthurian

ϕ_2
fairy wicked scene tale

ϕ_3
Harry Potter Voldemort Dumbledore

ϕ_{b}
evil powerful magic wizard

ϕ_3 ϕ_{b}

Rowling describes the beloved wizard **Dumbledore** as **Machiavellian** and **says**

Tiered Clustering

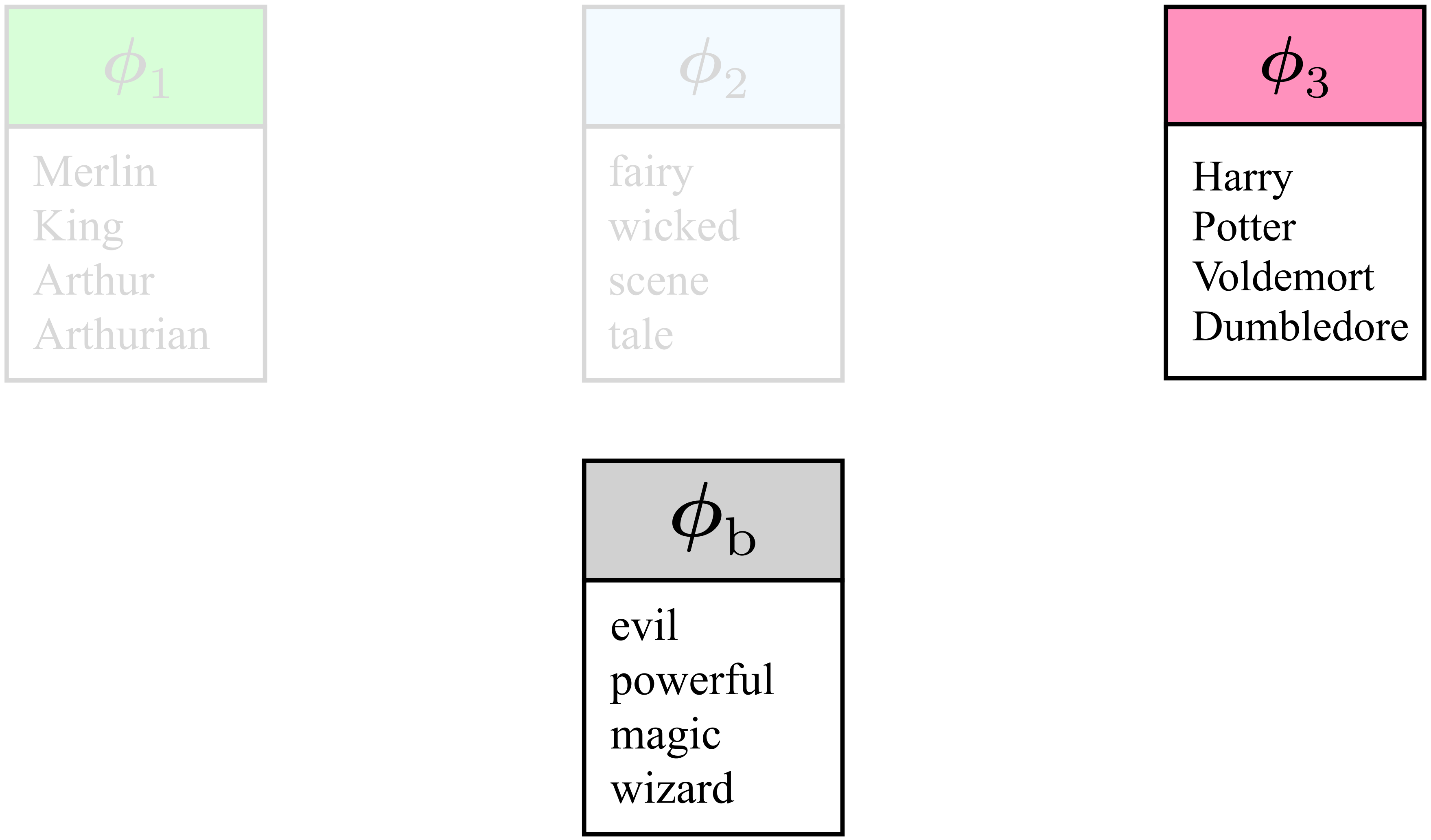
ϕ_1	ϕ_2	ϕ_3
Merlin	fairy	Harry
King	wicked	Potter
Arthur	scene	Voldemort
Arthurian	tale	Dumbledore

ϕ_b
evil
powerful
magic
wizard

ϕ_3 ϕ_b ϕ_b

Rowling **describes** the **beloved** **wizard** **Dumbledore** as **Machiavellian** and **says**

Tiered Clustering



ϕ_3 ϕ_b ϕ_b ϕ_3 ϕ_3 ϕ_b

Rowling describes the beloved <u>wizard</u> **Dumbledore** as **Machiavellian** and says

Tiered Clustering

ϕ_1	ϕ_2	ϕ_3
Merlin King Arthur Arthurian	fairy wicked scene tale	Harry Potter Voldemort Dumbledore

ϕ_b
evil powerful magic wizard

ϕ_3 ϕ_b ϕ_b ϕ_3 ϕ_3 ϕ_b

Rowling describes the beloved wizard **Dumbledore** as **Machiavellian** and says

Tiered Clustering

ϕ_1

Merlin
King
Arthur
Arthurian

ϕ_2

fairy
wicked
scene
tale

ϕ_3 $\sim \mathrm{Dir}(\beta)$

Harry
Potter
Voldemort
Dumbledore

ϕ_{b} $\sim \mathrm{Dir}(\beta_{\mathrm{background}})$

evil
powerful
magic
wizard

Merlin is a **legendary** figure best known as the **wizard** featured in the **Arthurian legend**

Tiered Clustering

ϕ_1	ϕ_2	ϕ_3	ϕ_{b}
Merlin	fairy	Harry	evil
King	wicked	Potter	powerful
Arthur	scene	Voldemort	magic
Arthurian	tale	Dumbledore	wizard

$$
\begin{array}{rcllr}
\theta_d|\alpha & \sim & \mathrm{Beta}(\alpha) & d \in D, & \text{(cluster proportions)} \\
\phi_d|\boldsymbol{\beta}, G_0 & \sim & \mathrm{DP}(\boldsymbol{\beta}, G_0) & d \in D, & \text{(clusters)} \\
\phi_{\mathrm{back}}|\boldsymbol{\beta}_{\mathrm{back}} & \sim & \mathrm{Dirichlet}(\boldsymbol{\beta}_{\mathrm{back}}) & & \text{(background / shared)} \\
z_{i,d}|\theta_d & \sim & \mathrm{Bernoulli}(\theta_d) & i \in |\mathbf{w}_d|, & \text{(tier indicator)} \\
w_{i,d}|\phi_d, z_{i,d} & \sim & \begin{cases} \mathrm{Mult}(\phi_{\mathrm{back}}) \\ \quad (z_{i,d} = 1) \\ \mathrm{Mult}(\phi_d) \\ \quad (\text{otherwise}) \end{cases} & i \in |\mathbf{w}_d|, & \text{(features)}
\end{array}
$$

Latent Dirichlet Allocation

$$
\begin{array}{rclll}
\boldsymbol{\theta}_d|\boldsymbol{\alpha} & \sim & \text{Dirichlet}(\boldsymbol{\alpha}), & d \in D, & \text{(topic proportions)} \\
\boldsymbol{\phi}_t|\boldsymbol{\beta} & \sim & \text{Dirichlet}(\boldsymbol{\beta}), & t \in T, & \text{(topics)} \\
z_{id}|\boldsymbol{\theta}_d & \sim & \text{Mult}(\boldsymbol{\theta}_d), & i \in |\mathbf{w}_d|, & \text{(topic indicators)} \\
w_{id}|\boldsymbol{\phi}_{z_{id}} & \sim & \text{Mult}(\boldsymbol{\phi}_{z_{id}}), & i \in |\mathbf{w}_d|, & \text{(words)}
\end{array}
$$

Tiered Clustering

$$
\begin{array}{rclll}
\theta_d|\alpha & \sim & \text{Beta}(\alpha) & d \in D, & \text{(cluster proportions)} \\
\boldsymbol{\phi}_d|\boldsymbol{\beta}, G_0 & \sim & \text{DP}(\boldsymbol{\beta}, G_0) & d \in D, & \text{(clusters)} \\
\boldsymbol{\phi}_{\text{back}}|\boldsymbol{\beta}_{\text{back}} & \sim & \text{Dirichlet}(\boldsymbol{\beta}_{\text{back}}) & & \text{(background / shared)} \\
z_{i,d}|\theta_d & \sim & \text{Bernoulli}(\theta_d) & i \in |\mathbf{w}_d|, & \text{(tier indicator)} \\
w_{i,d}|\boldsymbol{\phi}_d, z_{i,d} & \sim & \begin{cases} \text{Mult}(\boldsymbol{\phi}_{\text{back}}) \\ \quad (z_{i,d} = 1) \\ \text{Mult}(\boldsymbol{\phi}_d) \\ \quad \text{(otherwise)} \end{cases} & i \in |\mathbf{w}_d|, & \text{(features)}
\end{array}
$$

Prototype Comparison

multi-prototype

LIFE	STOCK
my, you, real, about, your, would	market, price, stock, company, value, crash
years, spent, rest, lived, last	housing, breeding, all, large, stock, many
sentenced, imprisonment, sentence, prison	car, racing, company, cars, summer, NASCAR
years, cycle, life, all, expectancy, other	stock, extended, folded, card, barrel, cards
all, life, way, people, human, social, many	rolling, locomotives, new, character, line

tiered

LIFE	STOCK
all, about, life, would, death	stock, all, other, company, new
my, you, real, your, about	market, crash, markets, price, prices
spent, years, rest, lived, last	housing, breeding, fish, water, horses
sentenced, imprisonment, sentence, prison	car, racing, cars, NASCAR, race, engine
insurance, peer, Baron, member, company	card, cards, player, pile, game, paper
Guru, Rabbi, Baba, la, teachings	rolling, locomotives, line, new, railway

(more polysemous) (more homonymous)

Prototype Comparison

multi-prototype

LIFE	STOCK
my, you, real, about, your, would	market, price, stock, company, value, crash
years, spent, rest, lived, last	housing, breeding, all, large, stock, many
sentenced, imprisonment, sentence, prison	car, racing, company, cars, summer, NASCAR
years, cycle, life, all, expectancy, other	stock, extended, folded, card, barrel, cards
all, life, way, people, human, social, many	rolling, locomotives, new, character, line

tiered

LIFE	STOCK
all, about, life, would, death	stock, all, other, company, new
my, you, real, your, about	market, crash, markets, price, prices
spent, years, rest, lived, last	housing, breeding, fish, water, horses
sentenced, imprisonment, sentence, prison	car, racing, cars, NASCAR, race, engine
insurance, peer, Baron, member, company	card, cards, player, pile, game, paper
Guru, Rabbi, Baba, la, teachings	rolling, locomotives, line, new, railway
(more polysemous)	(more homonymous)

Prototype Comparison

multi-prototype

LIFE

my, you, real, about, your, would
years, spent, rest, lived, last
sentenced, imprisonment, sentence, prison
years, cycle, life, all, expectancy, other
all, life, way, people, human, social, many

STOCK

market, price, stock, company, value, crash
housing, breeding, all, large, stock, many
car, racing, company, cars, summer, NASCAR
stock, extended, folded, card, barrel, cards
rolling, locomotives, new, character, line

tiered

LIFE

all, about, life, would, death
my, you, real, your, about
spent, years, rest, lived, last
sentenced, imprisonment, sentence, prison
insurance, peer, Baron, member, company
Guru, Rabbi, Baba, la, teachings

(more polysemous)

STOCK

stock, all, other, company, new
market, crash, markets, price, prices
housing, breeding, fish, water, horses
car, racing, cars, NASCAR, race, engine
card, cards, player, pile, game, paper
rolling, locomotives, line, new, railway

(more homonymous)

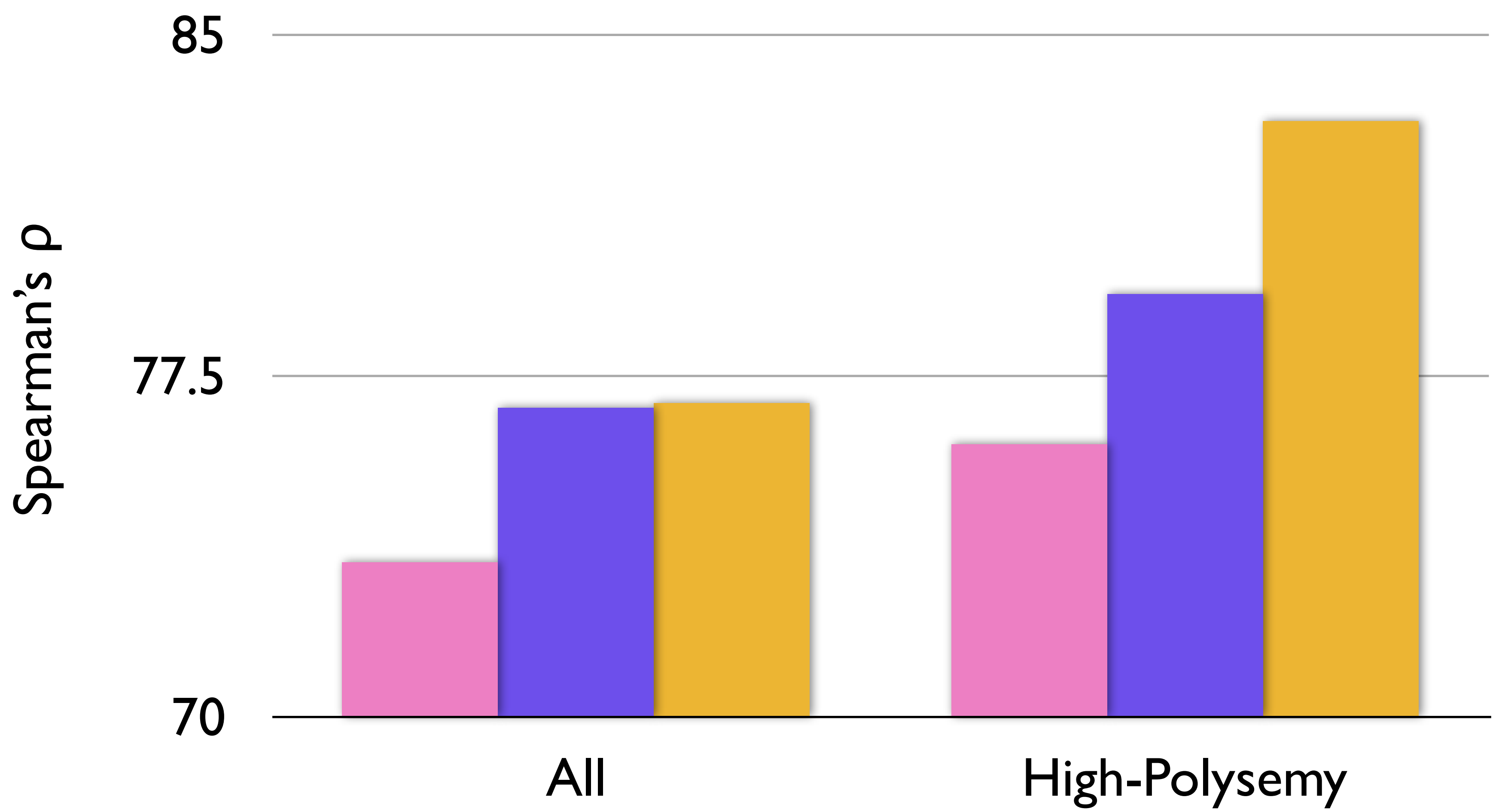
Experimental Setup

- Features: Unigram context features, 10 word windows; frequency-based pruning
- Model Training: Gibbs sampling; convergence when # of cluster swaps falls below a fixed threshold
- Evaluation: Spearman's ρ with human similarity judgements
- Inferred similarity is cosine using tf-idf features, adapted for tiered clustering:

$$\mathrm{AvgSim}(w, w') = \frac{1}{K_1 K_2} \sum_{j=1}^{K_1} \sum_{k=1}^{K_2} d(\phi_k^w, \phi_j^{w'})$$

$$\mathrm{TieredAvgSim}(w, w') = \mathrm{AvgSim}(w, w') + d(\phi_{\mathrm{back}}^w, \phi_{\mathrm{back}}^{w'})$$

WS-353 Correlation with Human Relatedness Judgements

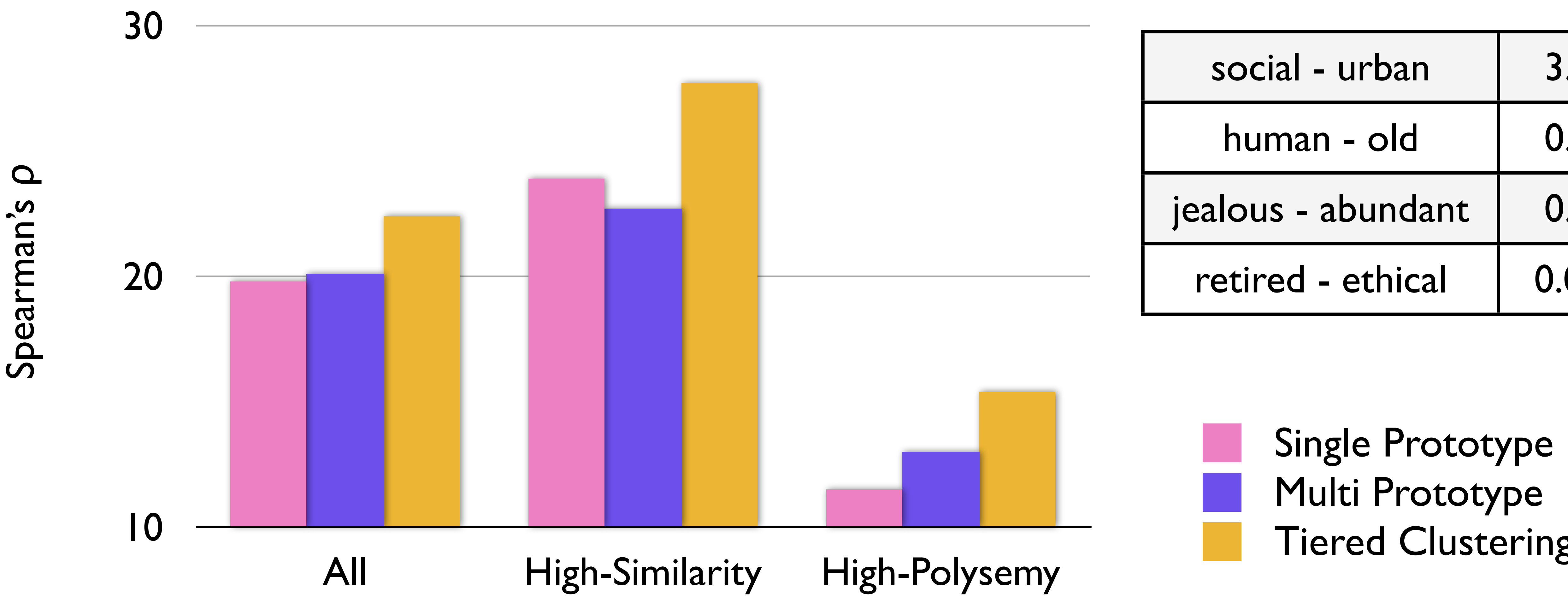


football - soccer	9.03
doctor - personnel	5.00
delay - news	3.31
Mars - water	2.94

- 353 words, biased towards pairs with high similarity
- High polysemy subset generated by counting WN senses

Dataset: Finkelstein et. al 2001

WN-Evocation Relatedness Correlation

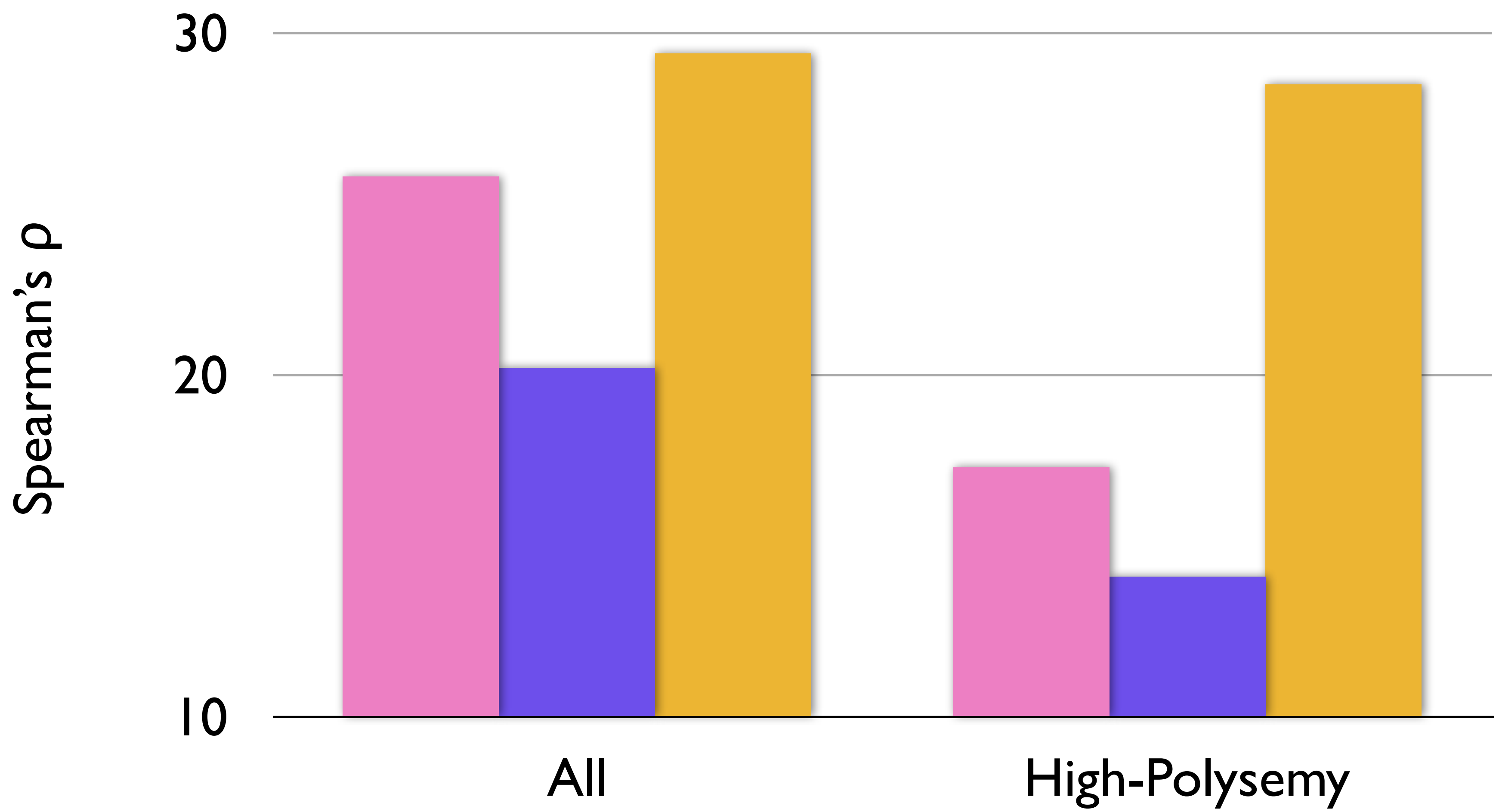


social - urban	3.3
human - old	0.9
jealous - abundant	0.2
retired - ethical	0.03

- Much higher degree of unrelated words than WS-353, more uniform sample over word pairs
- High similarity set restricts to pairs in the top quartile by human rating

Dataset: Ma et. al 2009

Padó Selectional Preference



brother hit	4.7
hit brother	4.1
quake hit	3.4
hit quake	1.2

- Predict typical arguments for verbs (e.g. things that can eat or things that can be shot)
- Separate model for each argument slot
- Background cluster captures commonalities between argument fillers.

Dataset: Padó et. al 2007

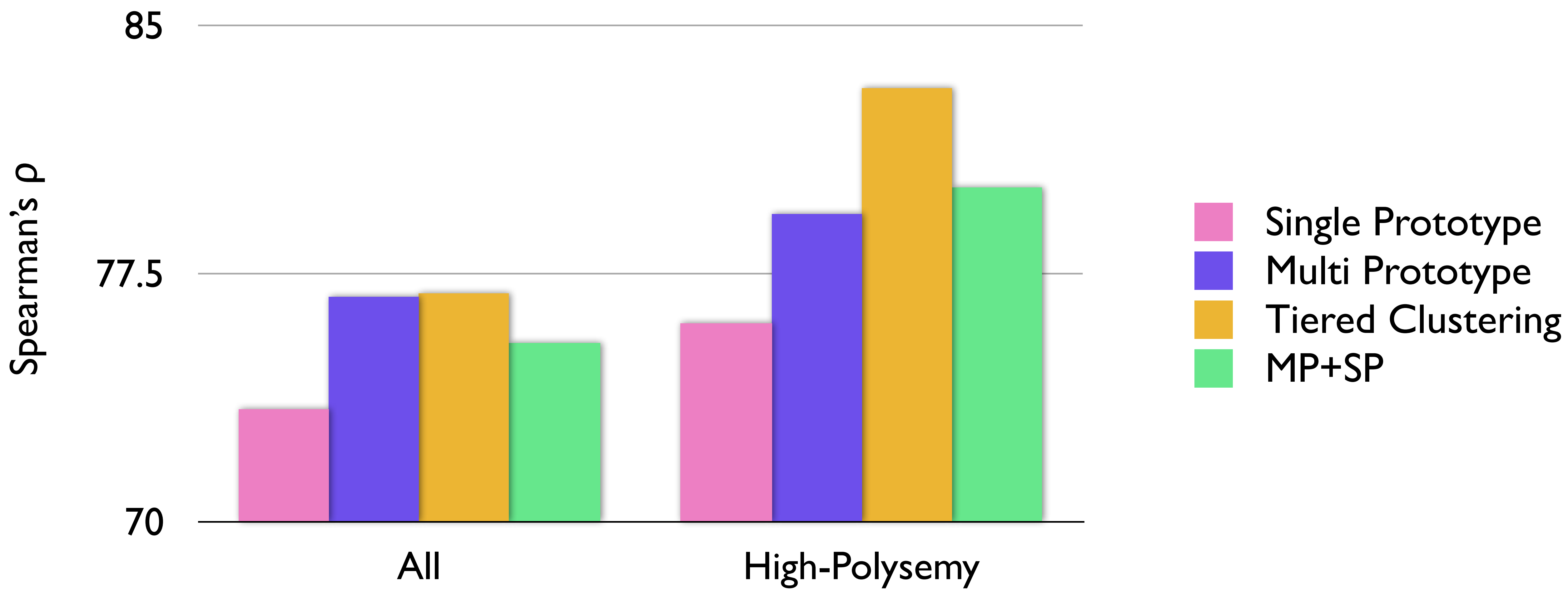
Conclusion

- Introduced a latent variable model explicitly accounting for shared structure when clustering
- *A priori* assumption that features can be separated into shared and idiosyncratic components
- Showed significant improvement over baseline for word-relatedness and selectional preference

Thanks!

Questions

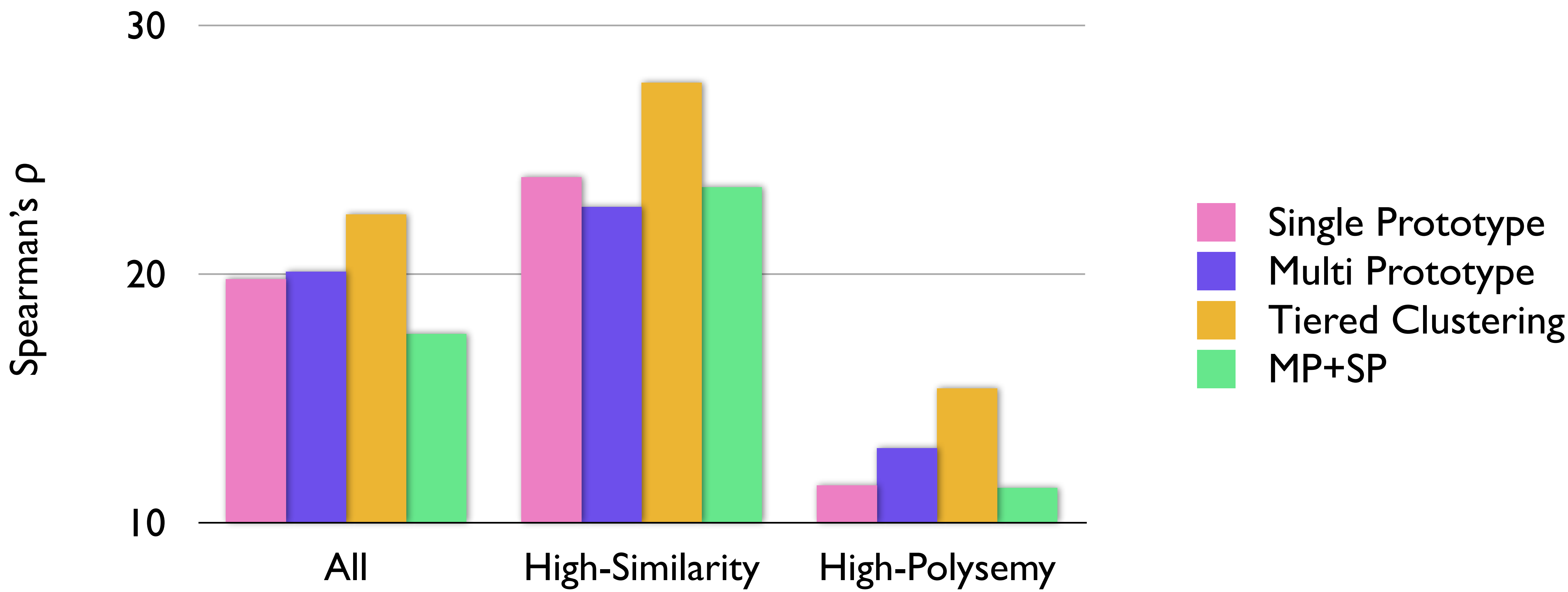
WS-353 Correlation with Human Relatedness Judgements



- 353 words, biased towards pairs with abnormally high similarity
- High polysemy subset generated by counting WN senses

(Finkelstein et. al 2001)

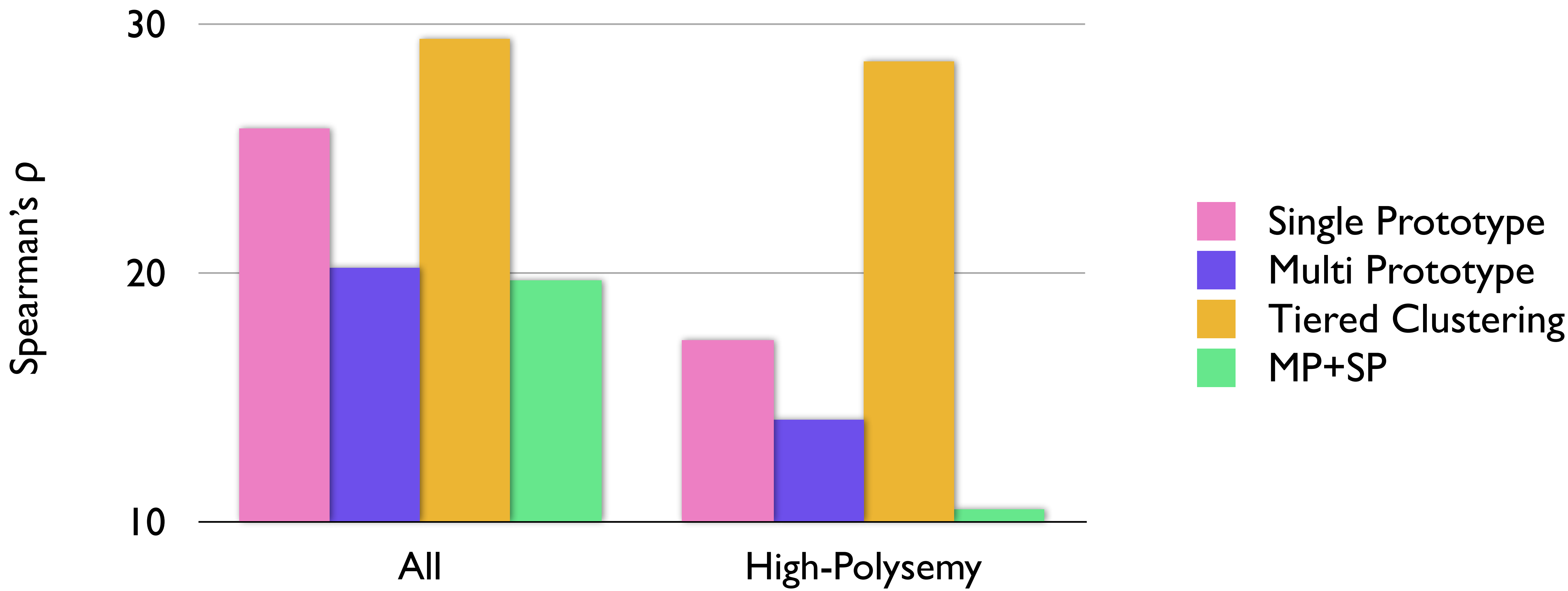
WN-Evocation Relatedness Correlation



- Much higher degree of unrelated words than WS-353, more uniform sample over word pairs
- High similarity set restricts to pairs in the top quartile by human rating

(Ma et. al 2009)

Padó Selectional Preference



- Background cluster captures commonalities between argument fillers, e.g. things that can eat or things that can be shot.

(Padó et. al 2007)