Machine Translation: Learning without Word Alignments

Introduction to Natural Language Processing
Computer Science 585—Fall 2009
University of Massachusetts Amherst

David Smith
With slides from Charles Schafer

Specialized Translation Models: Named Entities

Translating Words in a Sentence

- Models will automatically learn entries in probabilistic translation dictionaries, for instance p(elle|she), from co-occurrences in aligned sentences of a parallel text.

- For some kinds of words/phrases, this is less effective. For example:
  - numbers
  - dates
  - named entities (NE)

The reason: these constitute a large open class of words that will not all occur even in the largest bitext. Plus, there are regularities in translation of numbers/dates/ NE.

Handling Named Entities

- For many language pairs, and particularly those which do not share an alphabet, transliteration of person and place names is the desired method of translation.

- General Method:
    1. Identify NE's via classifier
    2. Transliterate name
    3. Translate/reorder honorifics

- Also useful for alignment. Consider the case of Inuktitut-English alignment, where Inuktitut renderings of European names are highly nondeterministic.

Transliteration

Inuktitut rendering of English names changes the string significantly but not deterministically

Williams	**McLean**
ailiams	makalain
uialims	makkalain
uilialums	maklaain
uiliam	maklain
uiliammas	maklainn
uiliams	maklait
uilians	makli
uliams	maklii
viliams	makliik
	makliin
Campbell	maklin
kaampu	malain
kaampul	matliin
kaamvul	miklain
kamvul	mikliin
	miklin

Transliteration

Inuktitut rendering of English names changes the string significantly but not deterministically

Train a **probabilistic finite-state transducer** to model this ambiguous transformation

Williams	**McLean**
ailiams	makalain
uialims	makkalain
uilialums	maklaain
uiliam	maklain
uiliammas	maklainn
uiliams	maklait
uilians	makli
uliams	maklii
viliams	makliik
	makliin
Campbell	maklin
kaampu	malain
kaampul	matliin
kaamvul	miklain
kamvul	mikliin
	miklin

Transliteration

Inuktitut rendering of English names changes the string significantly but not deterministically

Williams	**McLean**
ailiams	makalain
uialims	makkalain
uilialums	maklaain
uiliam	maklain
uiliammas	maklainn
uiliams	maklait
uilians	makli
uliams	maklii
viliams	makliik
	makliin
Campbell	maklin
kaampu	malain
kaampul	matliin
kaamvul	miklain
kamvul	mikliin
	miklin

… Mr. Williams … … mista uialims …

Useful Types of Word Analysis

- Number/Date Handling

- Named Entity Tagging/Transliteration

- Morphological Analysis
  - Analyze a word to its root form (at least for word alignment)

    was -> is believing -> believe

    ruminerai -> ruminer ruminiez -> ruminer
  - As a dimensionality reduction technique
  - To allow lookup in existing dictionary

Learning Word Translation Dictionaries Using Minimal Resources

Learning Translation Lexicons for Low-Resource Languages

{Serbian Uzbek Romanian Bengali} → English

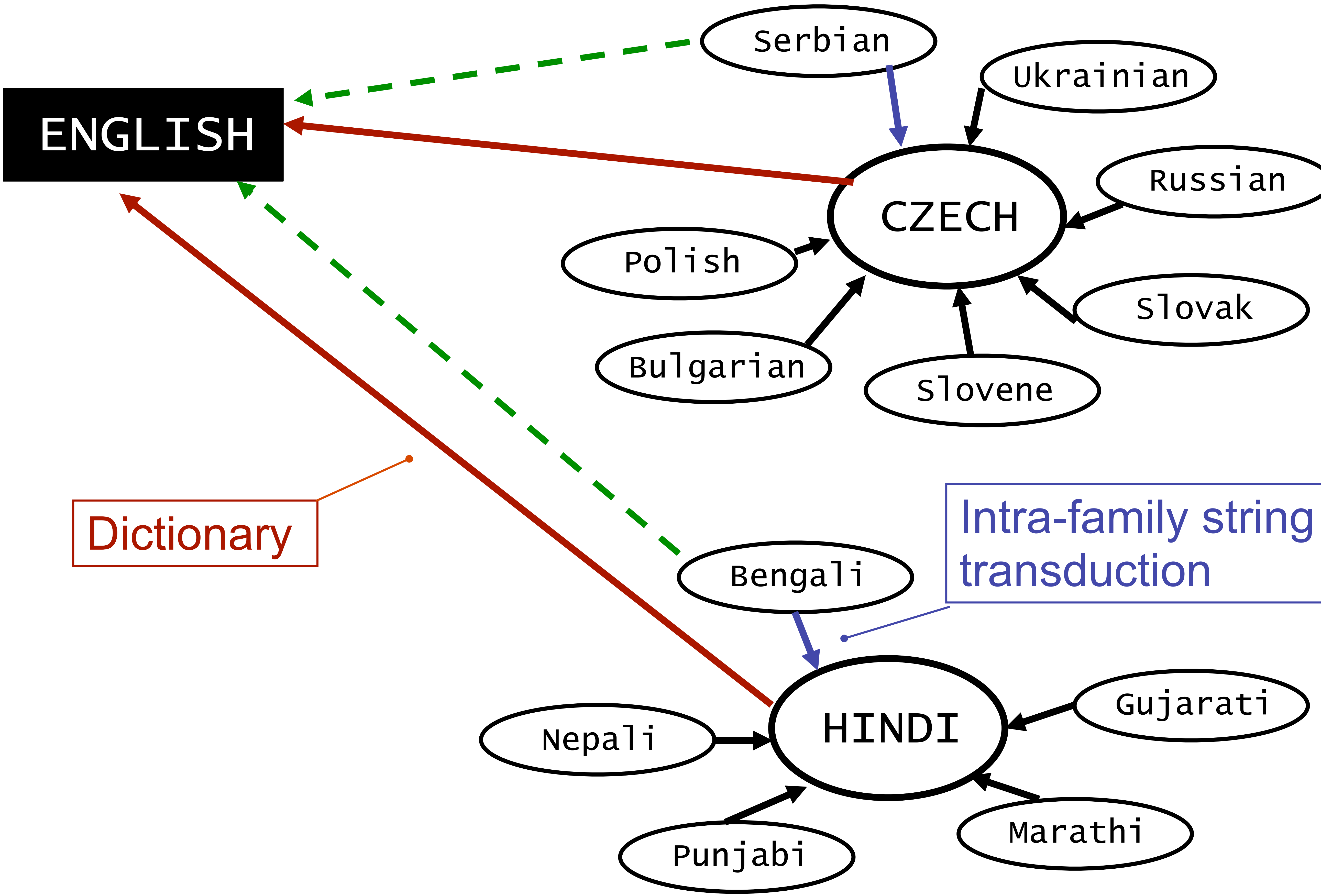
Problem: Scarce resources . . .

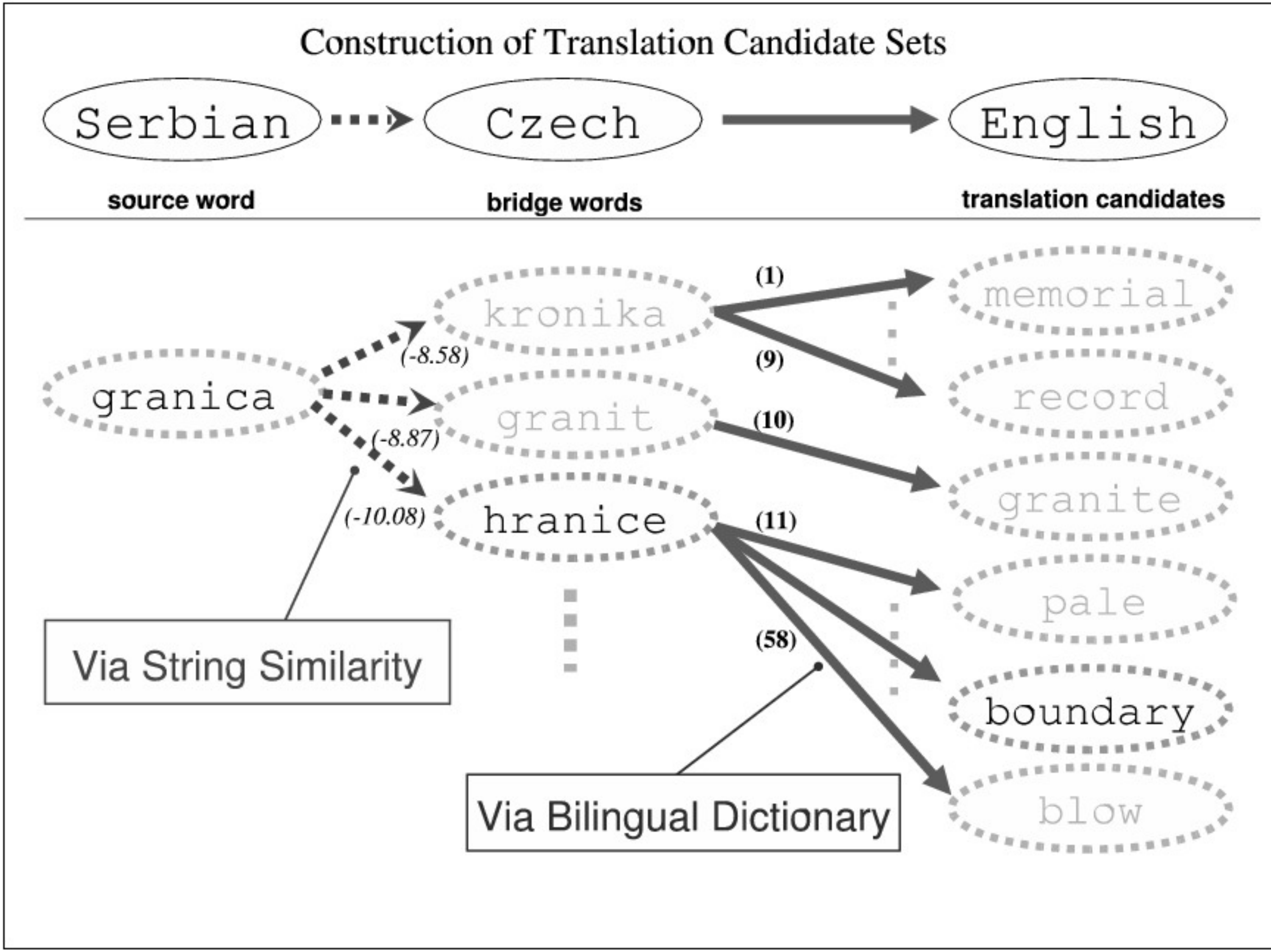
– Large parallel texts are very helpful, but often unavailable
– Often, no “seed” translation lexicon is available
– Neither are resources such as parsers, taggers, thesauri

Solution: Use only monolingual corpora in source, target languages

– But use many information sources to propose and rank translation candidates

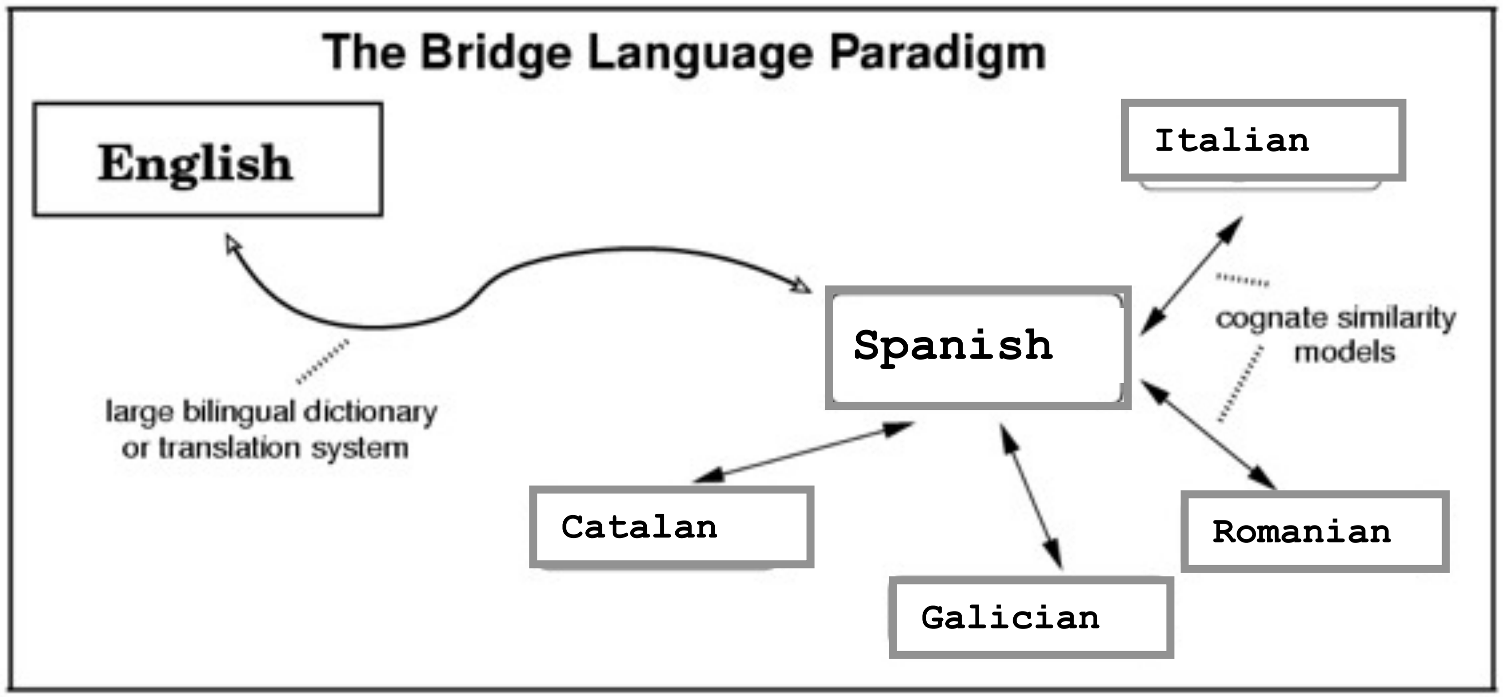
Bridge Languages





*** Constructing translation candidate sets**

Cognate Selection



some cognates

Spanish-Italian	homogenizar omogeneizzare
Polish-Serbian	befsztyk biftek
German-Dutch	gefestigt gevestigd

Spanish Word	Italian Word	Cognate?
electron	elettrone	
aventurero	avventuriero	
perífrasis	perifrasi	
divulgar	divulgare	
triada	triade	
agresivo	aggressivo	
insertar	inserto	
esprint	sprint	
trópico	tropico	
altimetro	altimetro	
alegato	lista	NO
variado	variato	
cepillar	piallare	
confusin	confusione	
fortificacion	fortificazione	
conjuncion	congiunzione	
encantador	incantatore	
heredero	erede	
vidrio	vetro	
vaciar	variare	NO
talisman	talismano	
sólido	solido	
criptografia	crittografia	
carencia	carenza	
cortesania	cortesia	NO
sadico	sadico	
concentracion	concentrazione	
venida	venuta	
agonizante	agonizzante	
extinguir	estinguere	

The Transliteration Problem

Arabic

Piedade	BEH YEH YEH DAL ALEF DAL YEH
Bolivia	BEH WAW LAM YEH FEH YEH ALEF
Luxembourg	LAM KAF SEEN MEEM BEH WAW REH GHAIN
Zanzibar	ZAIN NOON JEEM YEH BEH ALEF REH

Inuktitut

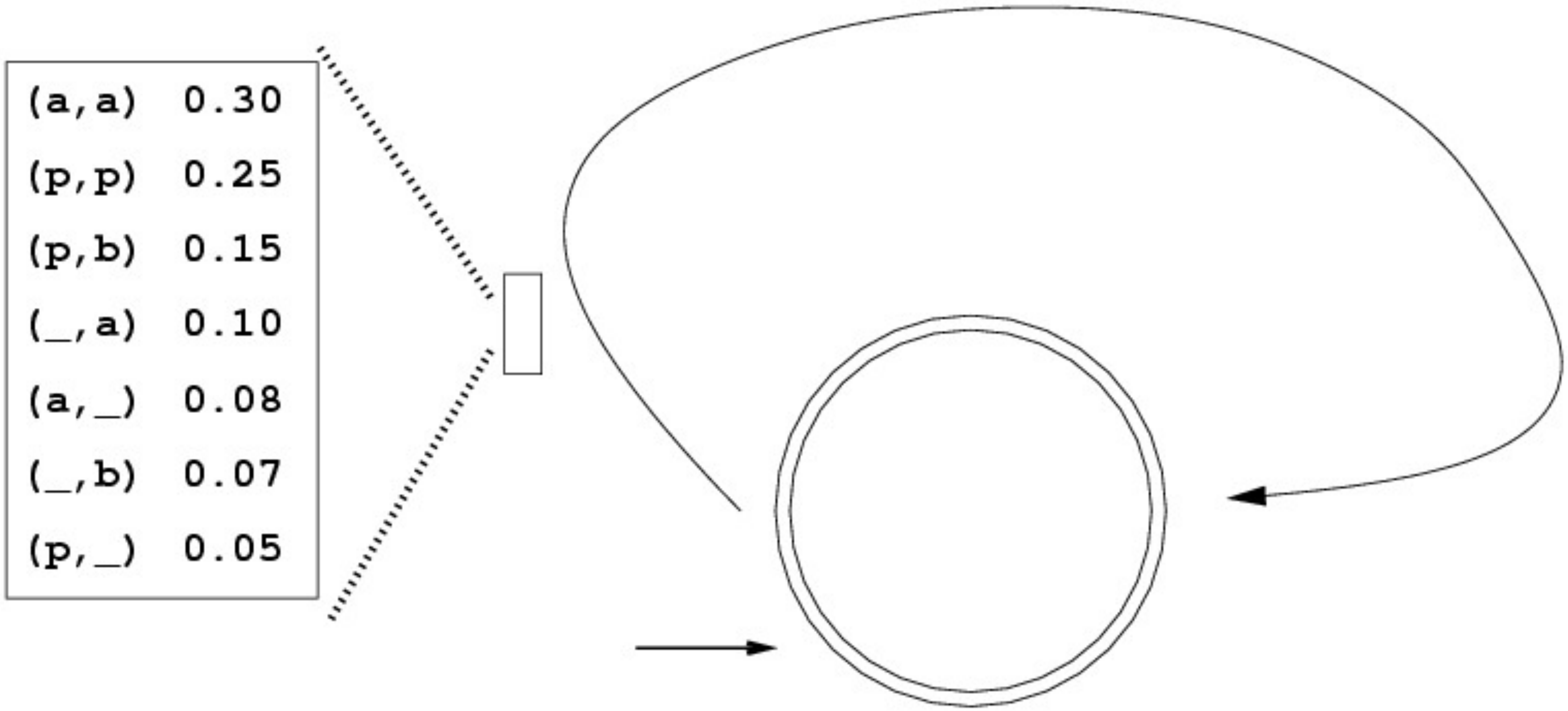
Williams: uialims uilialums uiliammas viliams

Campbell: kaampu kaampul kamvul kaamvul

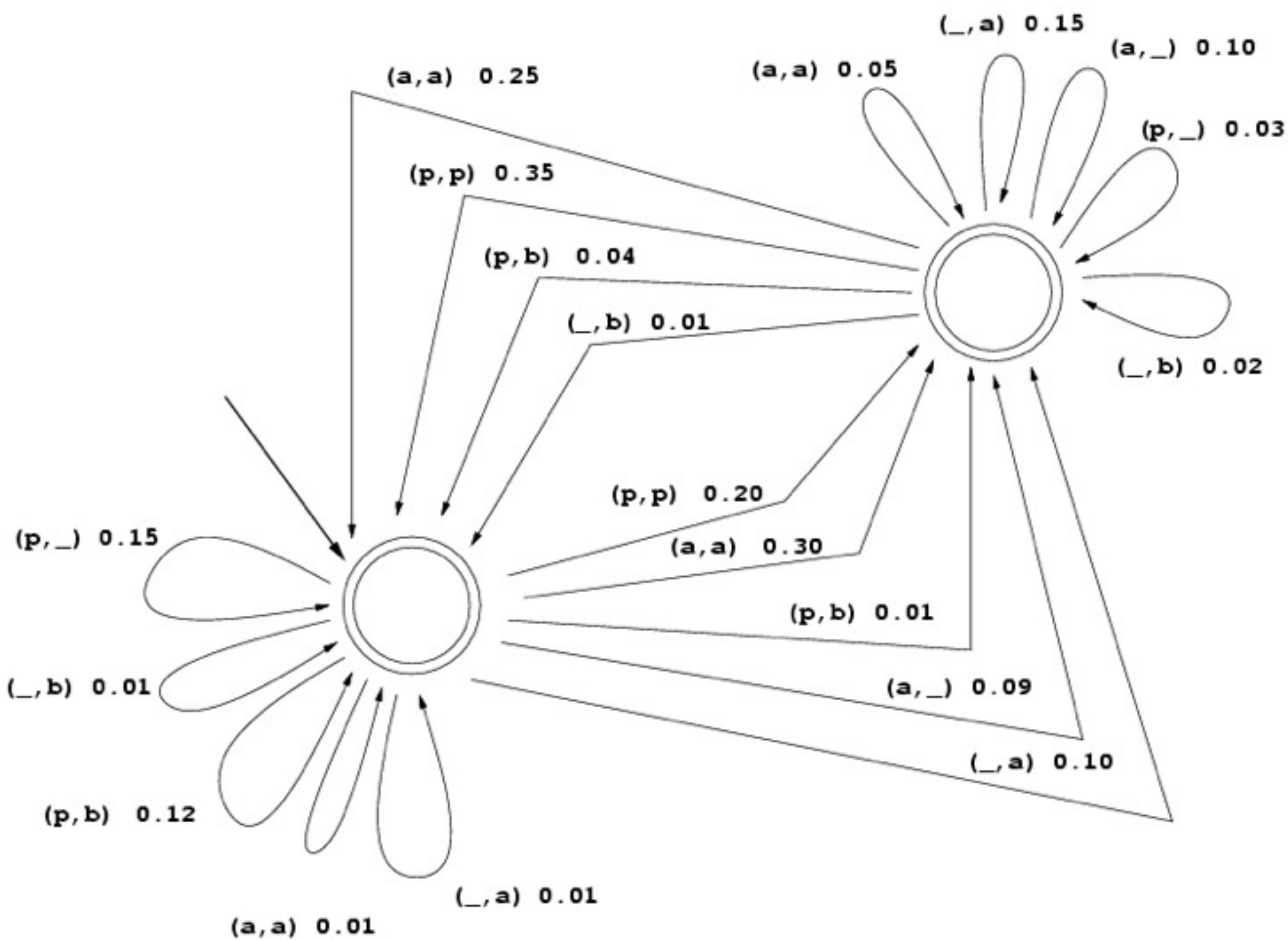
McLean: makalain maklainn makliin makkalain

Memoryless Transducer

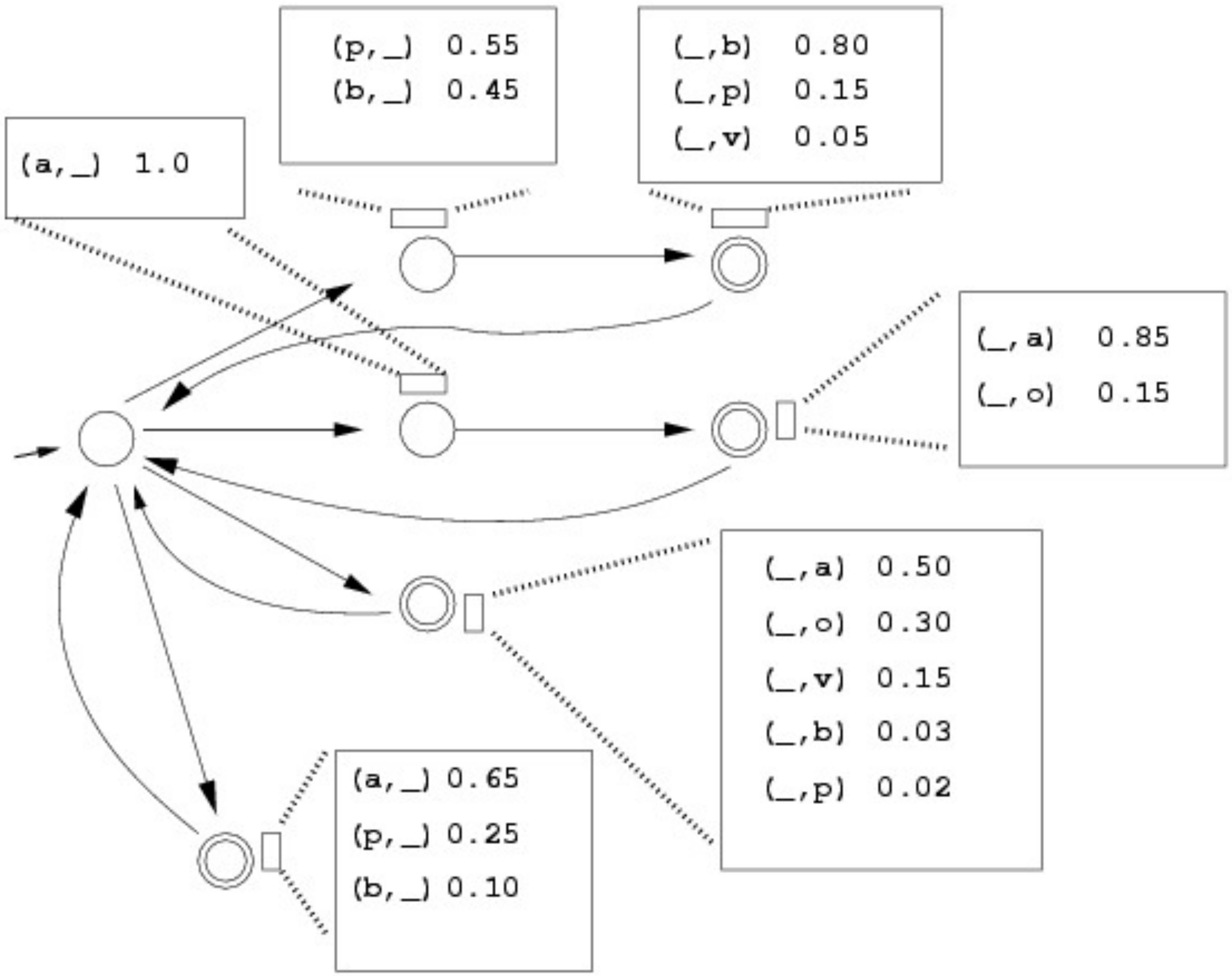
(Ristad & Yianilos 1997)



Two-State Transducer ("Weak Memory")



Unigram Interlingua Transducer



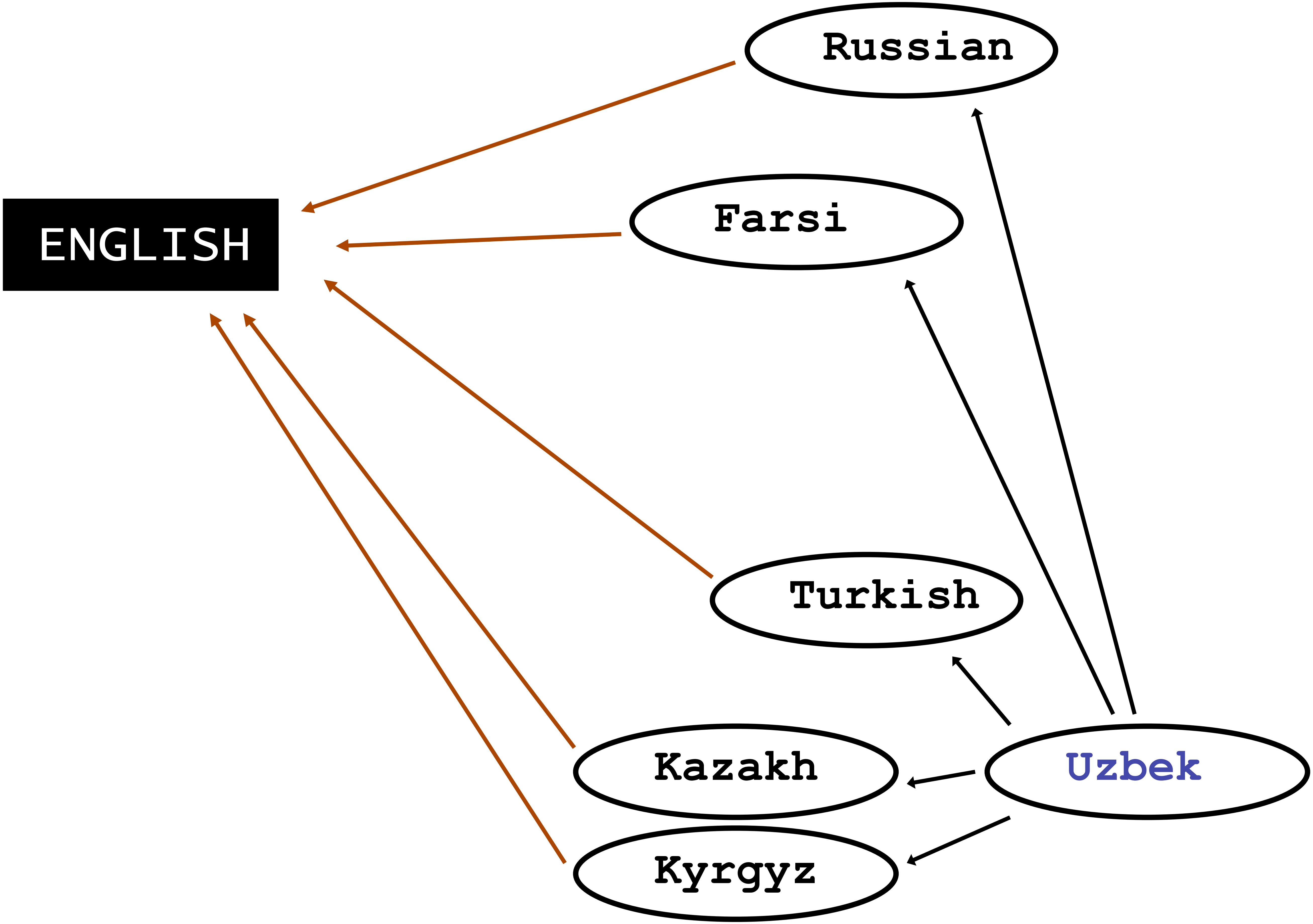
Examples: Possible Cognates Ranked by Various String Models

String Transduction Models Ranking Spanish Bridge Words for Romanian Source Word *inghiti*									
C1	**C2**	**C3**	**R&Y**	**2STEF**	**UIT**	**SN**	**AI**	**CDUI**	**JDCO**
S:ingrato	S:ingrato	S:ingrato	S:ingrato	S:ingrato	S:ingrato	S:ingrato	S:ingrato	S:ingrato	S:ingrato
S:ingerir	**S:ingerir**	S:engaste	S:grito	S:negrito	**S:ingerir**	S:ingente	S:negrito	S:infarto	S:engaste
S:engaste	S:engaste	**S:ingerir**	S:gaita	S:grito	S:grito	**S:ingerir**	S:negrita	S:engaste	S:anguila
S:ingreso	S:ingreso	S:inglete	S:grita	**S:ingerir**	S:grita	S:ingle	**S:ingerir**	S:ingreso	S:infarto
S:ingerido	S:ingerido	S:ingreso	S:negrito	S:negrita	S:inglete	S:angra	S:grito	S:introito	S:aguita
S:inglete	S:grito	S:ingerido	S:infarto	S:grita	S:gaita	S:ingerido	S:grita	S:negrito	S:ingreso
S:grito	S:inglete	S:infarto	S:negrita	S:gaita	S:negrito	S:ingenio	S:gaita	S:ingerido	S:intriga
S:infarto	S:infarto	S:grito	**S:ingerir**	S:ingerido	S:infarto	S:engan	S:ingenito	S:negrita	S:intuir
S:grita	S:negrito	S:introito	S:engaste	S:ingreso	S:introito	S:engatado	S:inglete	**S:ingerir**	S:indulto
S:introito	S:grita	S:engreir	S:haiti	S:haiti	S:engreir	S:invita	S:tahiti	S:inglete	S:inglete

String Transduction Models Ranking Turkish Bridge Words for Uzbek Source Word аввалги									
C1	**C2**	**C3**	**R&Y**	**2STEF**	**UIT**	**SN**	**AI**	**CDUI**	**JDCO**
T:evvelki	**T:evvelki**	**T:evvelki**	**T:evvelki**	T:vali	**T:evvelki**	T:edilgi	**T:evvelki**	**T:evvelki**	**T:evvelki**
T:evvelce	**T:evvelce**	**T:evvelce**	T:evveli	T:veli	**T:evvelce**	T:dalga	T:evveli	**T:evvelce**	**T:evvelce**
T:kalga	**T:evvelkí**	T:kalga	T:evvela	T:vals	T:edilgi	T:delgi	T:aval	T:evveli	**T:evvelkí**
T:evvelkí	T:kalga	T:salgi	**T:evvel**	T:delgi	T:algi	T:kalga	T:algi	T:evvela	T:ilkelci
T:vals	T:salgi	T:vals	T:algi	**T:evvelki**	T:salgi	T:evel	**T:evvel**	T:ilkelci	T:sivilce
T:salgi	T:vals	**T:evvelkí**	**T:evvelce**	T:kalga	T:vals	T:dalgl	T:evvela	T:eksilti	T:ilkelce
T:villa	T:villa	T:delgi	T:edilgi	T:dalga	T:delgi	**T:evvelki**	T:salgi	T:zavalli	T:akilci
T:silgi	T:silgi	T:villa	T:aval	T:villa	T:silgi	T:evlat	T:vali	**T:evvelkí**	T:eksilti
T:edilgi	T:ilkelci	T:evveli	T:evel	T:vale	T:kalga	T:dolgu	**T:evvelce**	**T:evvel**	T:asilce
T:volta	T:akilci	T:silgi	T:delgi	T:yilgi	T:dalga	T:veli	**T:evvelkí**	T:ilkelce	T:otelci

Romanian *inghiti* (ingest)
Uzbek *avvalgi* (previous/former)

* **Effectiveness of cognate models**



* Multi-family bridge languages

Similarity Measures

for re-ranking cognate/transliteration hypotheses

1. Probabilistic string transducers
2. Context similarity
3. Date distribution similarity
4. Similarities based on monolingual word properties

Similarity Measures

1. Probabilistic string transducers

2. Context similarity

3. Date distribution similarity

4. Similarities based on monolingual word properties

Compare Vectors

nezavisnost vector
Projection of context vector from Serbian to English term space

independence vector
Construction of context term vector

freedom vector
Construction of context term vector

	expression	religion	justice	majesty	sovereignty	declaration	country	ornamental
nezavisnost	0	0	2	1.5	1.5	1.5	4	1.5
independence	3	1	10	0	479	836	191	0
freedom	681	184	104	0	21	4	141	0

Compute cosine similarity between nezavisnost and “independence”

… and between nezavisnost and “freedom”

Similarity Measures

1. Probabilistic string transducers

2. Context similarity

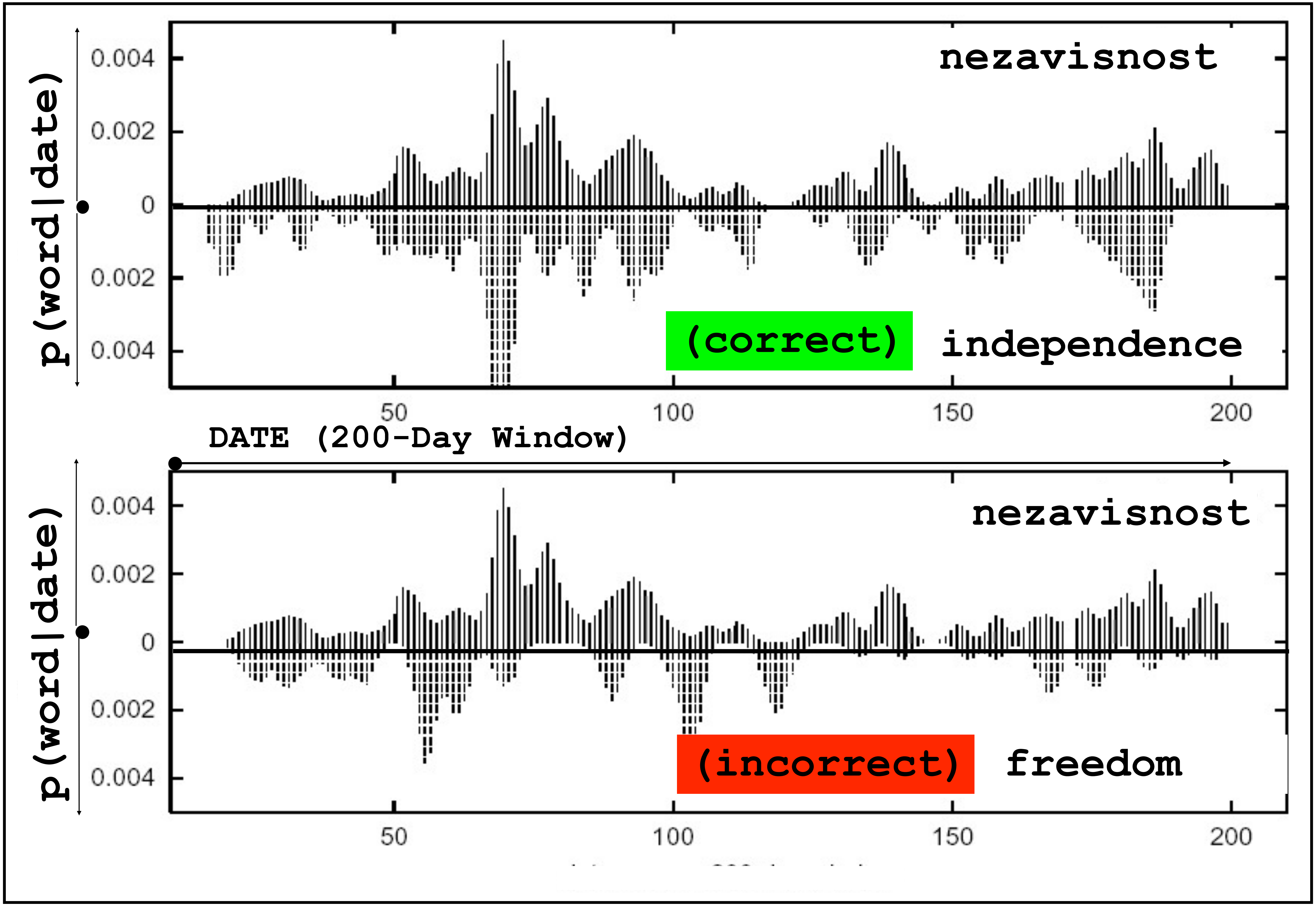
3. Date distribution similarity

4. Similarities based on monolingual word properties

Date Distribution Similarity

- Topical words associated with real-world events appear within news articles in bursts following the date of the event
- Synonymous topical words in different languages, then, display similar distributions across dates in news text: this can be measured
- We use cosine similarity on date term vectors, with term values `p(word|date)`, to quantify this notion of similarity

Date Distribution Similarity - Example



Similarity Measures

1. Probabilistic string transducers

2. Context similarity

3. Date distribution similarity

4. Similarities based on monolingual word properties

Relative Frequency

$$rf(w_F) = \frac{f_{C_F}(w_F)}{|C_F|}$$

$$rf(w_E) = \frac{f_{C_E}(w_E)}{|C_E|}$$

Cross-Language Comparison:

$$\min\left(\frac{rf(w_F)}{rf(w_E)}, \frac{rf(w_E)}{rf(w_F)}\right)$$

[min-ratio method]

Precedent in Yarowsky & Wicentowski (2000); used relative frequency similarity for morphological analysis

Combining Similarities: Uzbek

Individual Bridge Language Results For Uzbek Using Combined Similarity Measures				
Rank	**Turkish**	**Russian**	**Farsi**	**Kyrgyz**
1	0.04	**0.12**	0.03	0.06
5	0.10	**0.23**	0.05	0.08
10	0.13	**0.26**	0.07	0.10
20	0.16	**0.28**	0.08	0.11
50	0.21	**0.30**	0.12	0.13
100	0.24	**0.31**	0.15	0.16
200	0.26	**0.32**	0.19	0.19

Multiple Bridge Language Results For Uzbek Using Combined Similarity Measures					
Rank	**Tur+Rus**	**Tur+Rus +Farsi**	**Tur+Rus +Eng**	**Tur+Rus +Farsi +Kaz+Kyr**	**Tur+Rus +Farsi +Kaz+Kyr+Eng**
1	0.12	0.13	0.13	**0.14**	**0.14**
5	0.26	0.27	0.26	0.28	**0.29**
10	0.30	0.31	0.31	**0.34**	**0.34**
20	0.35	0.37	0.35	**0.39**	**0.39**
50	0.39	0.41	0.39	0.42	**0.43**
100	0.41	0.43	0.41	**0.46**	0.45
200	0.43	0.45	0.42	**0.48**	0.46

Combining Similarities: Romanian, Serbian, & Bengali

Multiple Bridge Language Results For Romanian Using Combined Similarity Measures				
Rank	Spanish	Spanish +Russian	Spanish +English	Spanish +Russian +English
1	0.17	0.18	**0.19**	**0.19**
5	0.31	0.35	0.34	**0.37**
10	0.37	0.41	0.41	**0.43**
20	0.43	0.46	0.46	**0.48**
50	0.51	0.53	0.53	**0.55**
100	0.57	0.60	0.58	**0.61**
200	0.60	**0.62**	0.59	**0.62**

Multiple Bridge Language Results For Serbian Using Combined Similarity Measures						
Rank	Cz	Rus	Bulg	Cz +English	Cz+Slovak +Rus+Bulg	Cz+Slovak +Rus+Bulg +English
1	0.13	0.15	**0.19**	0.13	**0.19**	**0.19**
5	0.24	0.24	0.31	0.25	**0.38**	**0.38**
10	0.29	0.28	0.35	0.30	0.42	**0.43**
20	0.32	0.31	0.40	0.34	**0.48**	**0.48**
50	0.38	0.36	0.44	0.39	0.54	**0.55**
100	0.40	0.40	0.48	0.42	**0.59**	**0.59**
200	0.41	0.42	0.50	0.43	**0.60**	**0.60**

Bridge Language Results for Bengali Using Combined Similarity Measures		
Rank	Hindi	Hindi +English
1	0.03	**0.05**
5	0.11	**0.14**
10	0.13	**0.17**
20	0.16	**0.21**
50	0.19	**0.25**
100	0.22	**0.28**
200	0.23	**0.29**

Observations

* With no Uzbek-specific supervision, we can produce an Uzbek-English dictionary which is 14% exact-match correct

* Or, we can put a correct translation in the top-10 list 34% of the time (useful for end-to-end machine translation or cross-language information retrieval)

* Adding more bridge languages helps

Multiple Bridge Language Results For Uzbek Using Combined Similarity Measures

Rank	Tur+Rus	Tur+Rus +Farsi	Tur+Rus +Eng	Tur+Rus +Farsi +Kaz+Kyr	Tur+Rus +Farsi +Kaz+Kyr+Eng
1	0.12	0.13	0.13	**0.14**	**0.14**
5	0.26	0.27	0.26	0.28	**0.29**
10	0.30	0.31	0.31	**0.34**	**0.34**
20	0.35	0.37	0.35	**0.39**	**0.39**
50	0.39	0.41	0.39	0.42	**0.43**
100	0.41	0.43	0.41	**0.46**	0.45
200	0.43	0.45	0.42	**0.48**	0.46

Polylingual Topic Models

Automated Analysis of Text

- Previously: analyzing trends in text collections (Hall et al., '08)
- Monolingual models often work well: collections in English only
- Multilingual text collections are increasingly common
- Automated tools are most important for multilingual collections:
  - Don't know the language → cannot eyeball the data
  - Humans typically only know a few languages
  - New data will appear in other languages
- Simultaneously analyze document content in many languages

Multiple Languages

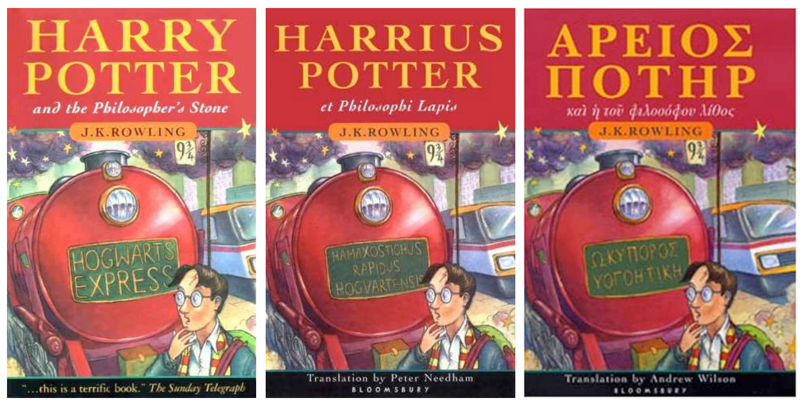
- Most statistical topic models are implicitly monolingual
- Why model multiple languages explicitly?

graph	problem	rendering	algebra	und	la
graphs	problems	graphics	algebras	von	des
edge	optimization	image	ring	die	le
vertices	algorithm	texture	rings	der	du
edges	programming	scene	modules	im	les

- Hodgepodge of English, German, French topics
- Imbalanced corpus: maybe only one or two French topics

Parallel vs. Comparable Corpora

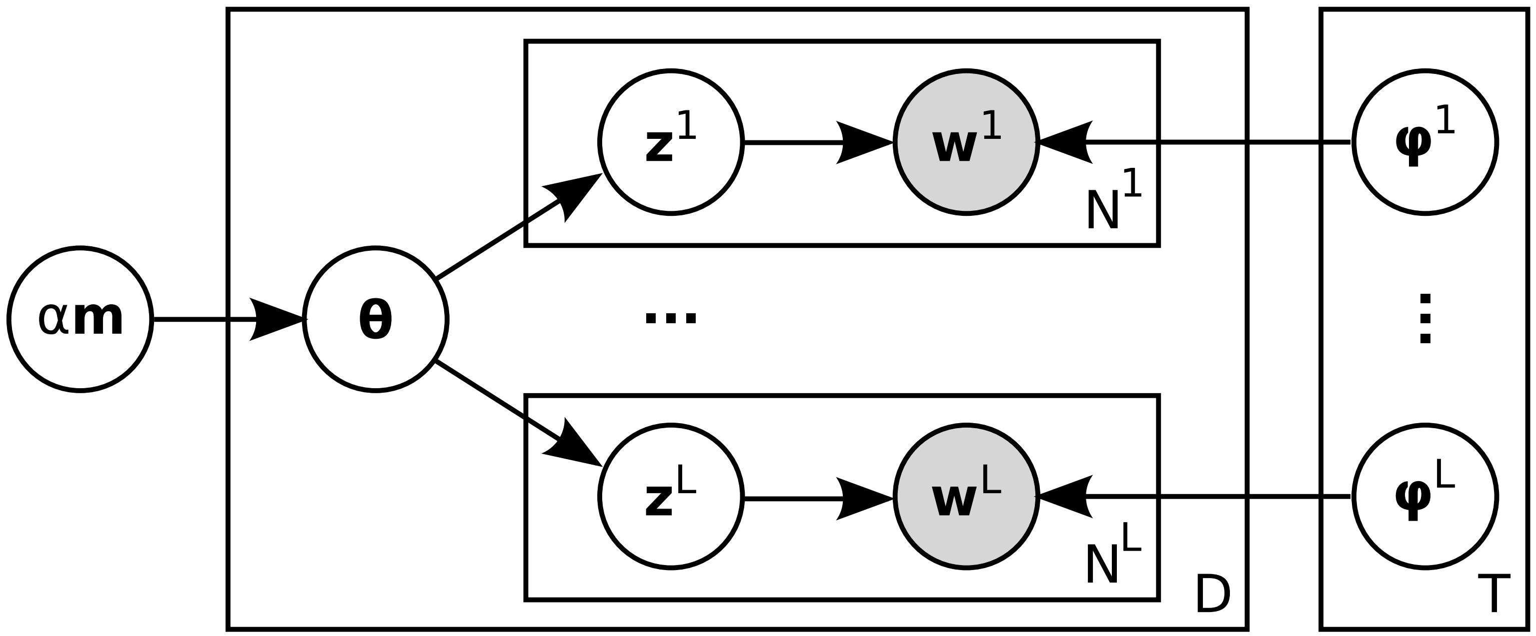
- A set of aligned documents is a "document tuple"



- Fully parallel corpora: documents are direct translations
- Corpora with a few parallel "glue" document tuples
- Comparable corpora: documents have similar semantic content

Polylingual Topic Model

- Generates a document tuple $\mathbf{w} = \mathbf{w}^1, \ldots, \mathbf{w}^L$ by drawing...



- For real-world data, only the word tokens are observed

Key Characteristics

- A topic is a *set* of distributions over words, e.g., $\phi_t = \phi_t^1, \ldots, \phi_t^L$
- Works on aligned document tuples, rather than documents
- Each tuple can consist of only a subset of languages
- Tuple-specific distributions over topics
  - Ensure cross-language consistency: e.g., topic 13 in French is semantically similar to topic 13 in English
- Simple, Gibbs sampling inference algorithm
  - No more complicated than latent Dirichlet allocation

EuroParl: Example Topics ($T = 400$)

DA	centralbank europæiske ecb s lån centralbanks
DE	zentralbank ezb bank europäischen investitionsbank darlehen
EL	τράπεζα τράπεζας κεντρική εκτ κεντρικής τράπεζες
EN	**bank central ecb banks european monetary**
ES	banco central europeo bce bancos centrales
FI	keskuspankin ekp n euroopan keskuspankki eip
FR	banque centrale bce européenne banques monétaire
IT	banca centrale bce europea banche prestiti
NL	bank centrale ecb europese banken leningen
PT	banco central europeu bce bancos empréstimos
SV	centralbanken europeiska ecb centralbankens s lån

EuroParl: Example Topics ($T = 400$)

DA	mål nå målsætninger målet målsætning opnå
DE	ziel ziele erreichen zielen erreicht zielsetzungen
EL	στόχους στόχο στόχος στόχων στόχοι επίτευξη
EN	**objective objectives achieve aim ambitious set**
ES	objetivo objetivos alcanzar conseguir lograr estos
FI	tavoite tavoitteet tavoitteena tavoitteiden tavoitteita tavoitteen
FR	objectif objectifs atteindre but cet ambitieux
IT	obiettivo obiettivi raggiungere degli scopo quello
NL	doelstellingen doel doelstelling bereiken bereikt doelen
PT	objectivo objectivos alcançar atingir ambicioso conseguir
SV	mål målet uppnå målen målsättningar målsättning

EuroParl: Example Topics ($T = 400$)

DA	andre anden side ene andet øvrige
DE	anderen andere einen wie andererseits anderer
EL	άλλες άλλα άλλη άλλων άλλους όπως
EN	**other one hand others another there**
ES	otros otras otro otra parte demás
FI	muiden toisaalta muita muut muihin muun
FR	autres autre part côté ailleurs même
IT	altri altre altro altra dall parte
NL	andere anderzijds anderen ander als kant
PT	outros outras outro lado outra noutros
SV	andra sidan å annat ena annan

EuroParl: Analysis of Trained Model

- Is the model genuinely learning mixtures of topics?

  Mr President, yesterday, at the end of the vote on the budget, there was a moment when the three institutions concerned were represented by women and the President concluded by saying that the Millennium was ending on a high note.

  Monsieur le Président, hier, la fin du vote sur le budget fut un moment particulier dans la mesure où les trois institutions impliquées étaient représenté es par des personnes de sexe féminin. La Présidente a conclu en disant que le millénaire s'achevait sur une note positive.

EuroParl: Analysis of Trained Model

- Is the model genuinely learning mixtures of topics?

  Mr President, yesterday, at the end of the vote on the budget, there was a moment when the three institutions concerned were represented by women and the President concluded by saying that the Millennium was ending on a high note.

  Monsieur le Président, hier, la fin du vote sur le budget fut un moment particulier dans la mesure où les trois institutions impliquées étaient représenté es par des personnes de sexe féminin. La Présidente a conclu en disant que le millénaire s'achevait sur une note positive.

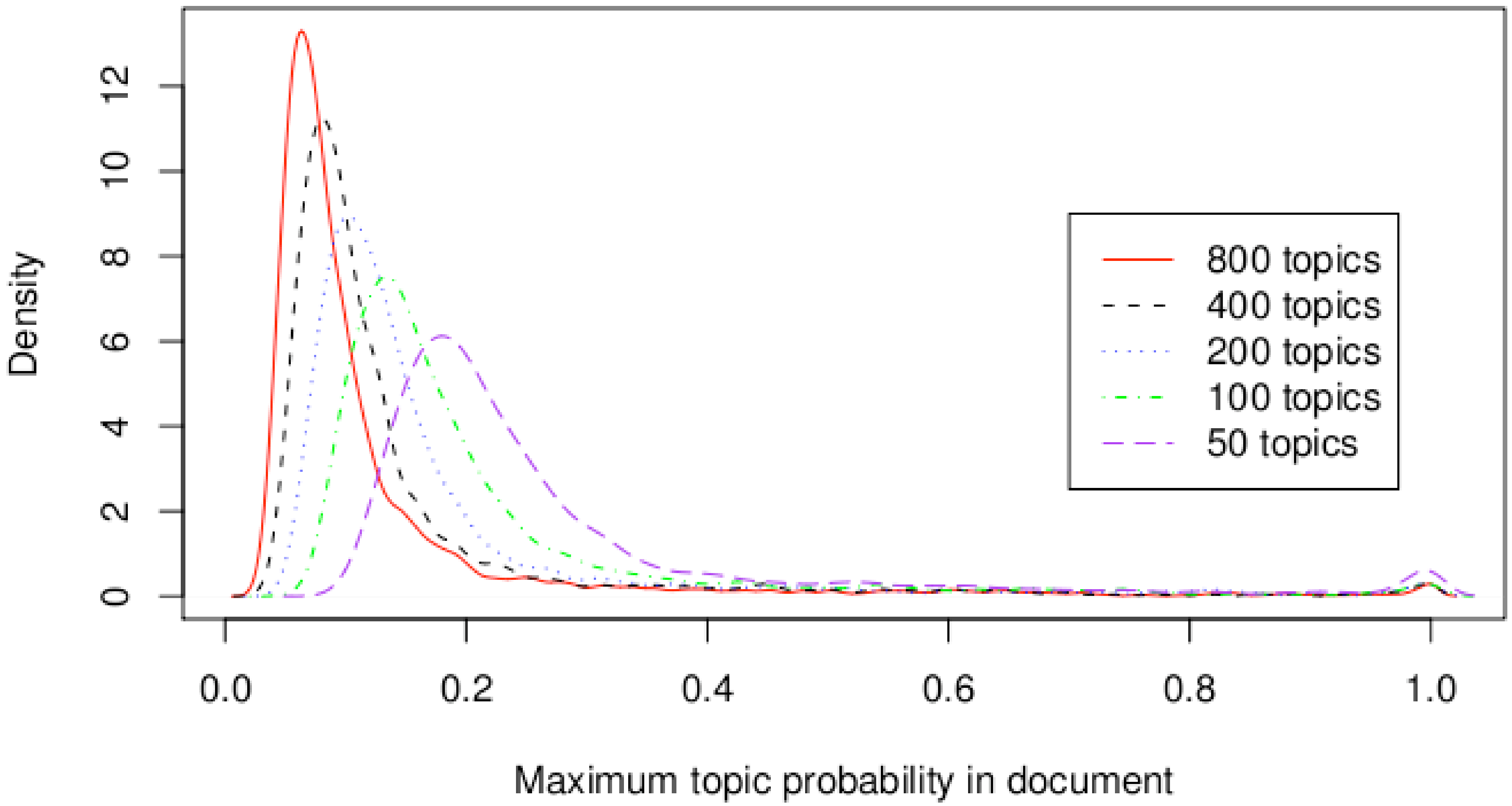
EuroParl: Analysis of Trained Model

- Is the model genuinely learning mixtures of topics?

  Mr President, yesterday, at the end of the vote on the budget, there was a moment when the three institutions concerned were represented by women and the President concluded by saying that the Millennium was ending on a high note.

  Monsieur le Président, hier, la fin du vote sur le budget fut un moment particulier dans la mesure où les trois institutions impliquées étaient représenté es par des personnes de sexe féminin. La Présidente a conclu en disant que le millénaire s'achevait sur une note positive.

EuroParl: Analysis of Trained Model



Parallel Corpora: "Glue" Tuples

- How many aligned documents are needed to get aligned topics?

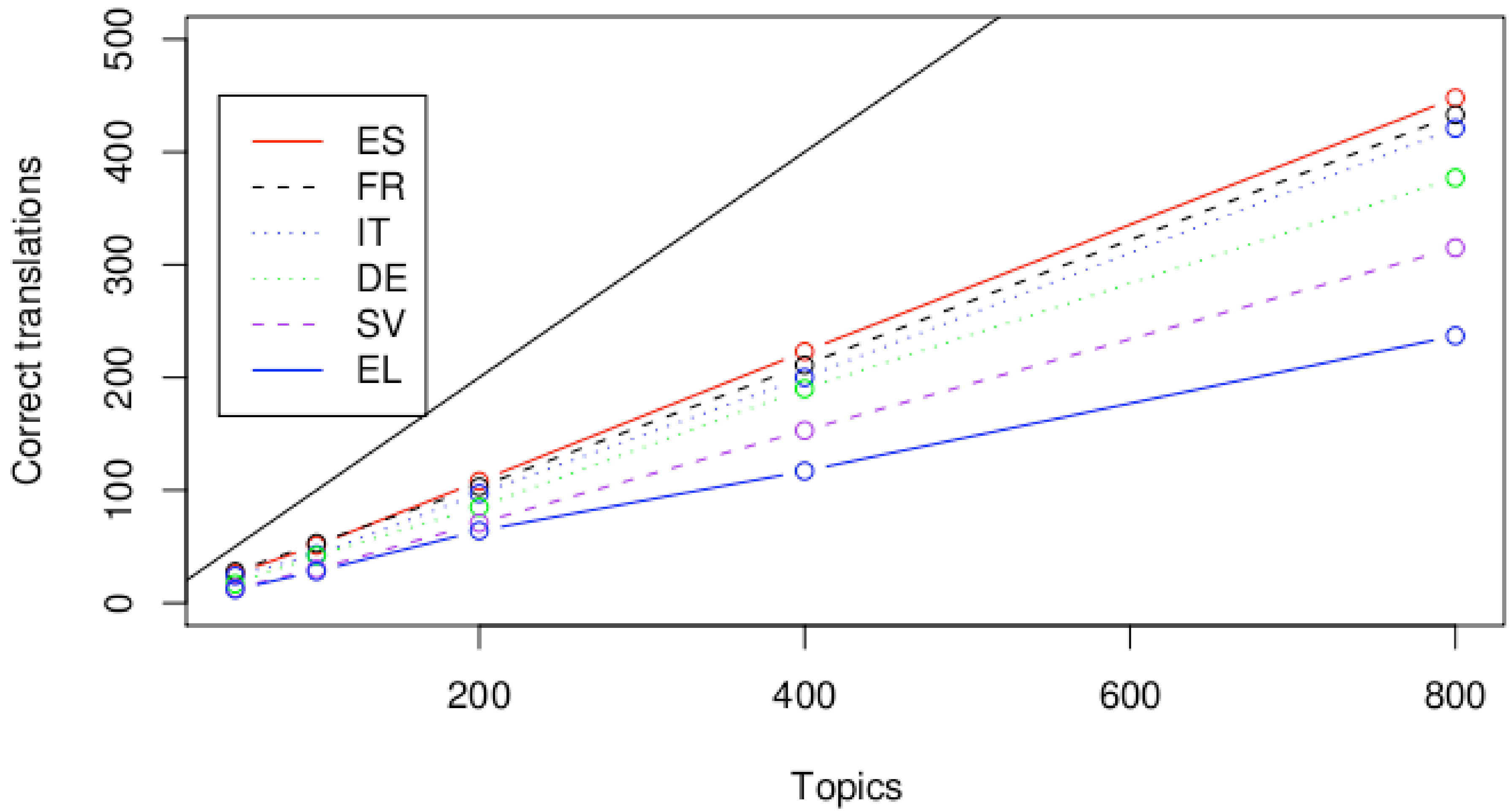
	topics with 1% "glue" tuples
DE	rußland russland russischen tschetschenien sicherheit
EN	china rights human country s burma
IT	ho presidente mi perché relazione votato

	topics with 25% "glue" tuples
DE	rußland russland russischen tschetschenien ukraine
EN	russia russian chechnya cooperation region belarus
IT	russia unione cooperazione cecenia regione russa

Generating Bilingual Lexica

- Bilingual lexicon: word pairs (e.g., English word, translation)
- High probability words in different languages for a topic are likely to include translations – can use these to generate lexica
- Form candidate translations: Cartesian product of most probable K words in English and in each translation language
- No morphological variants: e.g., rules/vorschriften, rule/vorschrift
- Count # of lexicon pairs that are in the candidate set
- Advantages: unsupervised; all words, not just nouns

Generating Bilingual Lexica ($K = 1$)

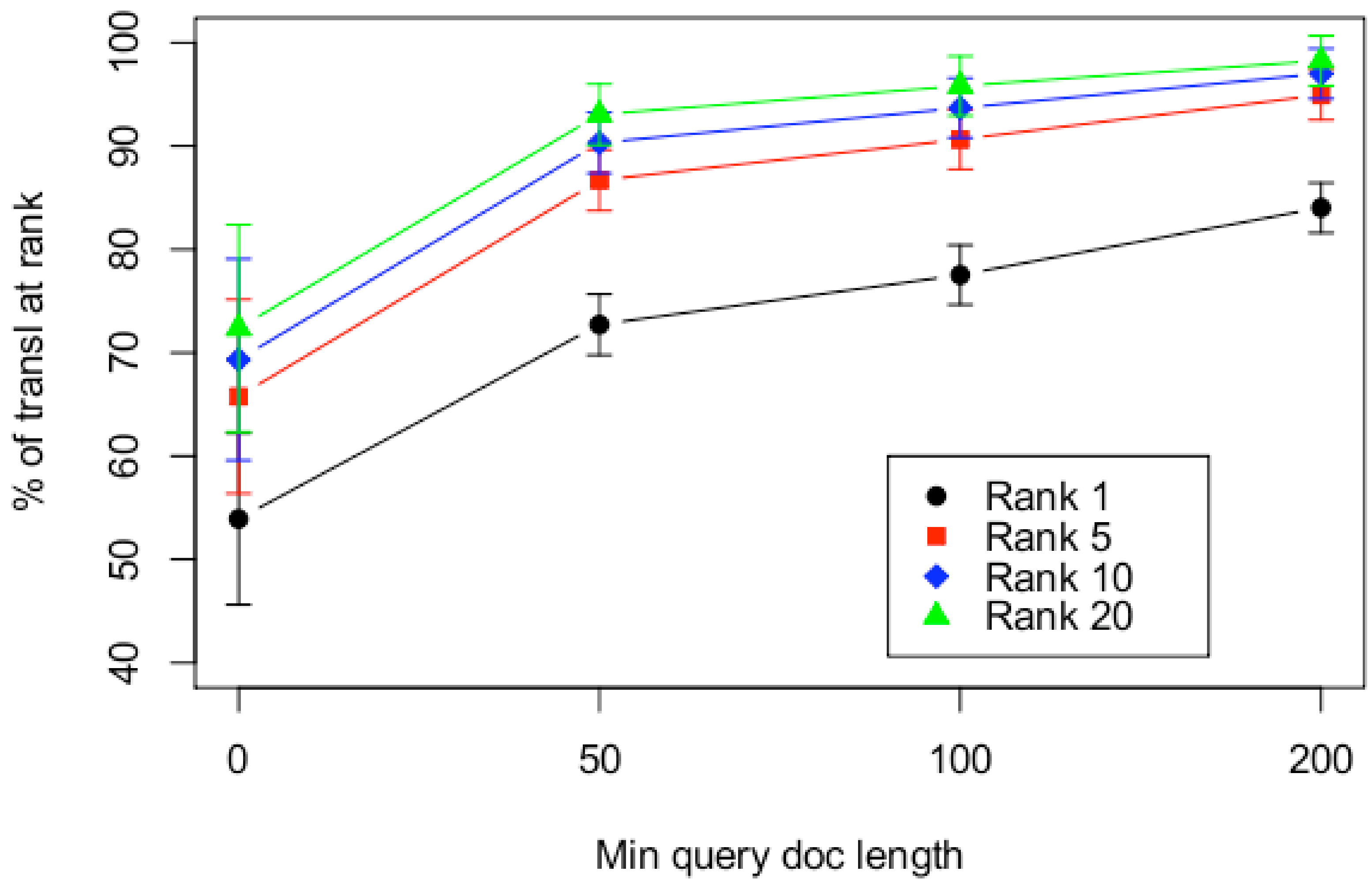


Finding Translations

- Train model on aligned document tuples
- Output: set of polylingual topics, e.g., $\phi_t = \phi_t^1, \ldots, \phi_t^L$

- Map each test document to the low-dimensional space defined by the polylingual topics $\rightarrow$ document-topic distributions
- For each query/target language pair:
  - Compute similarities for all query/target document pairs
  - For each query document, rank target documents by similarity
- Jensen-Shannon divergence, cosine distance

Finding Translations (Jensen-Shannon)



Comparable Corpora

- Directly parallel translations are rare, expensive to produce
- Comparable corpora more common: e.g., Wikipedia, web pages
  - Our data set: all Wikipedia articles in English, Farsi, Finnish, French, German, Greek, Hebrew, Italian, Polish, Russian, Turkish, Welsh
- Documents are topically similar but not direct translations
- More interesting questions, more real-world applications:
  - Do comparable document tuples support alignment of topics?
  - Do different languages have different perspectives?
  - Which topics do particular languages focus on?

Wikipedia: Example Topics ($T = 400$)

CY	sadwrn blaned gallair at lloeren mytholeg
DE	space nasa sojus flug mission
EL	διαστημικό sts nasa αγγλ small
EN	**space mission launch satellite nasa spacecraft**
FA	فضایی ماموریت ناسا مدار فضانورد ماهواره
FI	sojuz nasa apollo ensimmäinen space lento
FR	spatiale mission orbite mars satellite spatial
HE	החלל הארץ חלל כדור א תוכנית
IT	spaziale missione programma space sojuz stazione
PL	misja kosmicznej stacji misji space nasa
RU	космический союз космического спутник станции
TR	uzay soyuz ay uzaya salyut sovyetler

Wikipedia: Example Topics ($T = 400$)

CY	sbaen madrid el la josé sbaeneg
DE	de spanischer spanischen spanien madrid la
EL	ισπανίας ισπανία de ισπανός ντε μαδρίτη
EN	**de spanish spain la madrid y**
FA	ترین de اسپانیا اسپانیایی کوبا مادرید
FI	espanja de espanjan madrid la real
FR	espagnol espagne madrid espagnole juan y
HE	ספרד ספרדית דה מדריד הספרדית קובה
IT	de spagna spagnolo spagnola madrid el
PL	de hiszpański hiszpanii la juan y
RU	де мадрид испании испания испанский de
TR	ispanya ispanyol madrid la küba real

Wikipedia: Example Topics ($T = 400$)

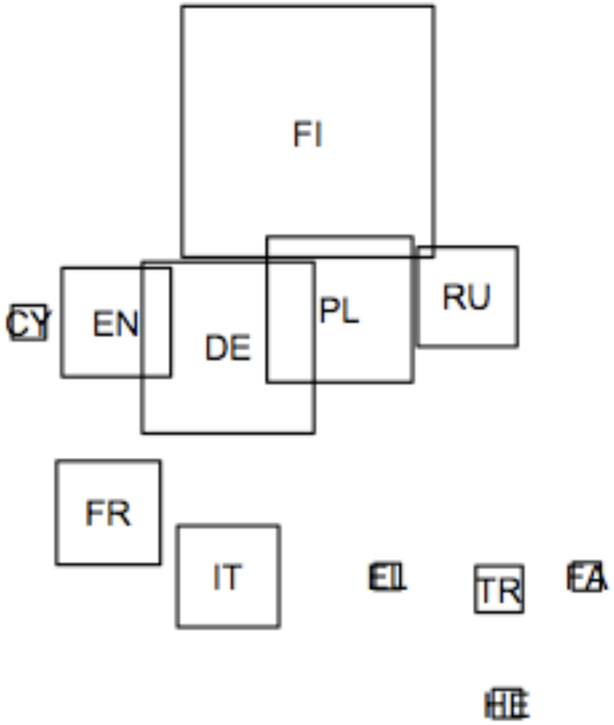
CY	bardd gerddi iaith beirdd fardd gymraeg
DE	dichter schriftsteller literatur gedichte gedicht werk
EL	ποιητής ποίηση ποιητή έργο ποιητές ποιήματα
EN	**poet poetry literature literary poems poem**
FA	شاعر شعر ادبیات فارسی ادبی آثار
FI	runoilija kirjailija kirjallisuuden kirjoitti runo julkaisi
FR	poète écrivain littérature poésie littéraire ses
HE	משורר ספרות שירה סופר שירים המשורר
IT	poeta letteratura poesia opere versi poema
PL	poeta literatury poezji pisarz in jego
RU	поэт его писатель литературы поэзии драматург
TR	şair edebiyat şiir yazar edebiyatı adlı

Topic Divergence Between Languages

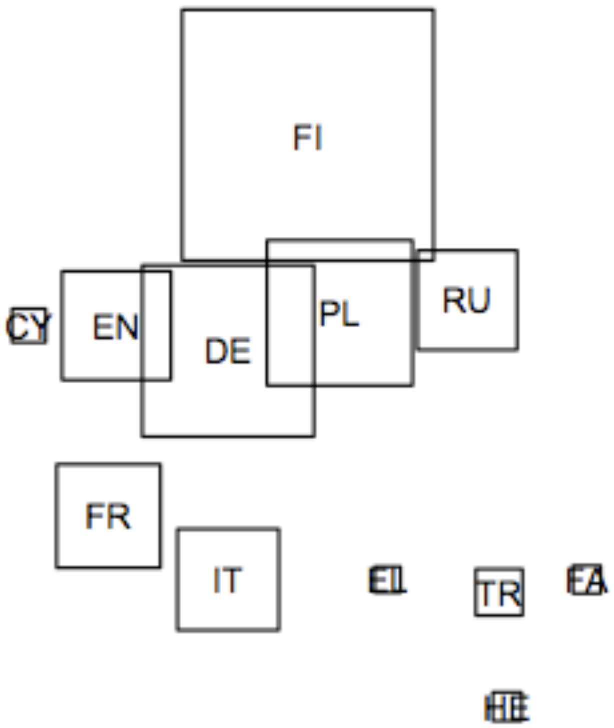
- Estimate document-specific distributions over topics
- Compute Jensen-Shannon divergence between documents in a tuple
- Average document-document divergences for each language pair:
  - "Disagreement" score for each language pair

- Almost all language pairs have divergences consistent with EuroParl data, even languages that have historically been in conflict

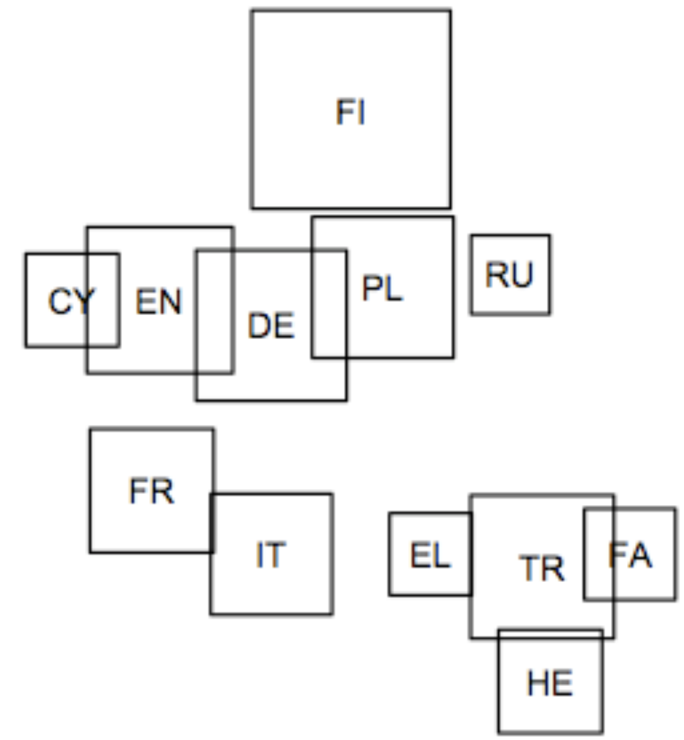
- Although individual articles may have high between-language divergence, Wikipedia is on average consistent between languages



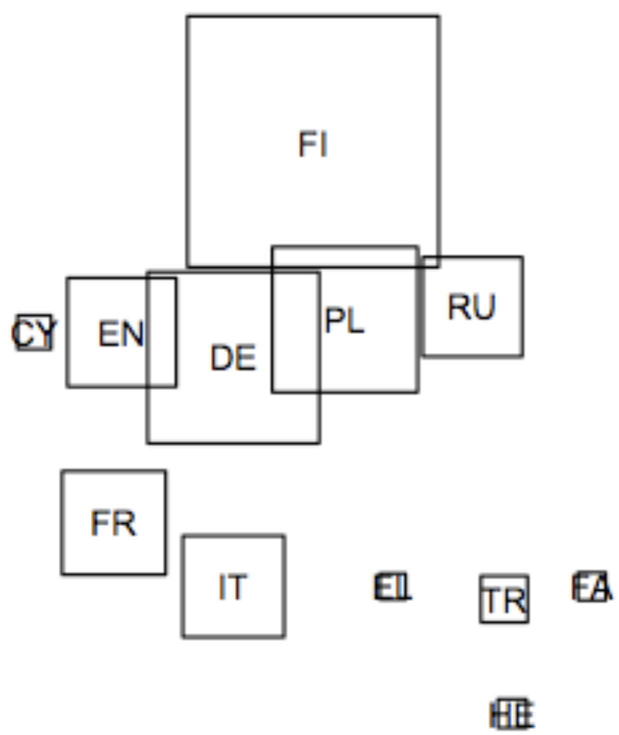
world ski km won



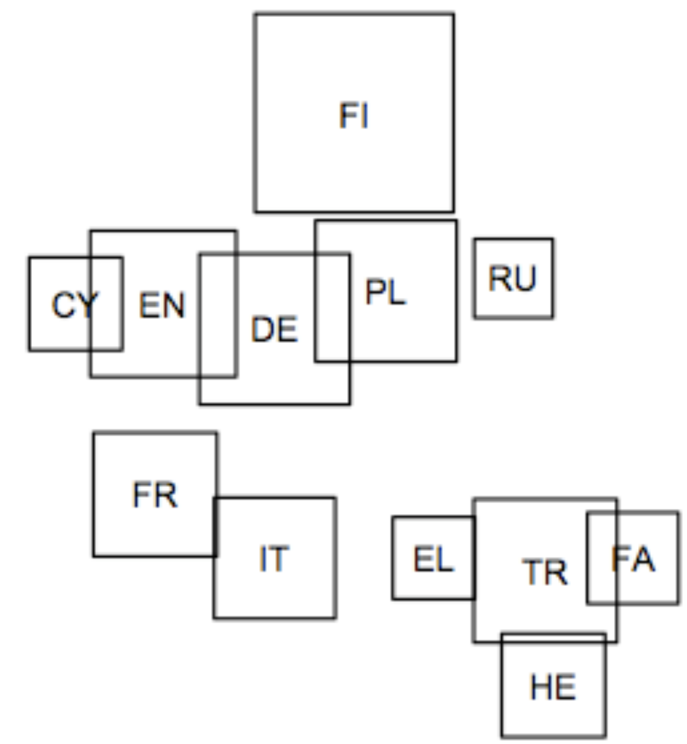
world ski km won



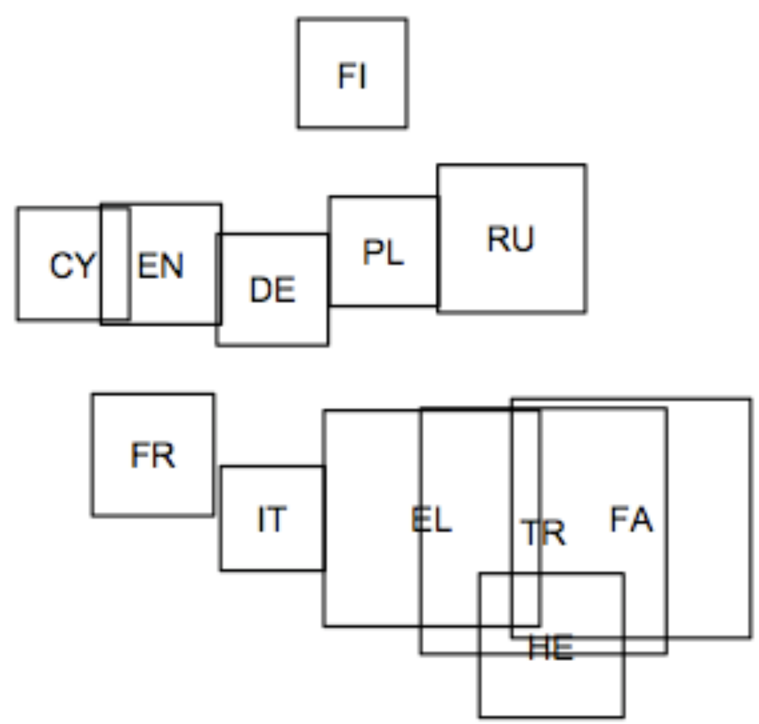
actor role television actress



world ski km won



actor role television actress



ottoman empire khan byzantine