Goal: Learn *inflection* ↔ *root* mappings for world’s languages with and without direct supervision

Definition: A **supervised** learning algorithm uses training data from generalized rules can be formed.

Inflection	**Root**	**Part of Speech**
	English	
swims	swim	3S Present Indicative
swimming	swim	Gerund
swam	swim	1S Past Indicative
swum	swim	Participle
	French	
abrège	abréger	1S Present Indicative
abrègent	abréger	3P Present Indicative
abrégerai	abréger	1S Future Indicative
conçu	concevoir	1P Future Anterior Indicative
crois	croire	1S Present Indicative
croyaient	croire	3P Imperfect Indicative
	Irish	
thóg	tóg	1S Past Indicative
thógadh	tóg	2P Imperfect Indicative
thógaidís	tóg	3P Imperfect Indicative
adhairim	adhair	1S Present Indicative
d’adhairfeá	adhair	2S Conditional
	Dutch	
bood	bieden	3S Past Indicative
gebiedt	bieden	2S Present Indicative
geboden	bieden	1S Past Perfect Indicative
verbood	verbieden	3S Past Indicative
verbiedt	verbieden	2S Present Indicative
aangeboden	aanbieden	1S Past Perfect Indicative

Inflection	**Root**	**Part of Speech**
	Turkish	
sandınız	sanmak	2P Past Def.Ind.Pos.Int.
sanaydınız	sanmak	2P Past Nar.Sub.Pos.Sta.
sanmayaydınız	sanmak	2P Past Nar.Sub.Neg.Sta.
sanmalıydınız	sanmak	2P Past Nar.Nec.Pos.Sta.
sanmalıymışsınız	sanmak	2P Past Dub.Nec.Pos.Sta.
sanmalıymışsınız	sanmak	2P Past Dub.Nec.Pos.Sta.
sanmamalıymışsınız	sanmak	2P Past Dub.Nec.Neg.Sta.
	Tagalog	
gugupitin	gupit	Indicative OF Cont.
pagugupitin	gupit	Causative A_2F Cont$_1$
paggugupitin	gupit	Causative A_2F Cont$_2$
papaggupitin	gupit	Causative A_2F Inf.
papaggugupitin	gupit	Causative A_2F Cont$_3$
	Swahili	
ninaagua	agua	1S Present Indicative
unaagua	agua	2S Present Indicative
niliagua	agua	1S Past Indicative
uliagua	agua	2S Past Indicative
nitaagua	agua	1S Future
utaagua	agua	2S Future
	Arabic	
katab	ktb	Active “write”
kattab	ktb	Active “cause to write”
ktutib	ktb	Passive “write”

Inflectional morphological phenomenon*

*from a computational linguistist's P.O.V. & using orthography instead of phonology

affixation	prefixation:	geuza	→	**mli**geuza	(*Swahili*)
	suffixation:	adhair	→	adhair**im**	(*Irish*)
	circumfixation:	mischen	→	**ge**misch**t**	(*German*)
	infixation:	palit	→	p**um**alit	(*Tagalog*)
point-of-affixation stem changes		pla**c**er	→	pla**ç**a	(*French*)
	elision:	clos**e**	→	closing	(*English*)
	gemination:	stir	→	stir**r**ed	(*English*)
	voicing:	zwer**f**t	→	zwer**v**en	(*Dutch*)
vowel harmony		abartmak	→	ab**a**rtm**a**s**a**nız	(*Turkish*)
		addetmek	→	add**e**tm**e**s**e**niz	(*Turkish*)
internal vowel shift		afbr**y**de	→	afbr**ø**d	(*Danish*)
		skr**i**ke	→	skr**ei**k	(*Norwegian*)
		sl**ee**p	→	sl**e**pt	(*English*)
agglutination and reduplication	agglutination:	ev	→	ev**de**	(*Turkish*)
	agglutination:		→	evde**ki**	
	agglutination:		→	evdeki**ler**	
	reduplication:	**gu**pit	→	**gugu**pit	(*Tagalog*)
	agglutination:		→	**i**gugupit	
	agglutination:		→	i**pa**gugupit	
	agglutination:		→	ip**in**agugupit	
	reduplication:	rumah	→	**rumah**rumah	(*Malay*)
	reduplication:	ibu	→	**ibu**ibu	
root and pattern		ktb	→	k**a**t**e**b	(*Arabic*)
		ktb	→	k**a**tt**a**b	
highly irregular forms		fi	→	erai	(*Romanian*)
		jānā	→	gayā	(*Hindi*)
		eiga	→	áttum	(*Icelandic*)

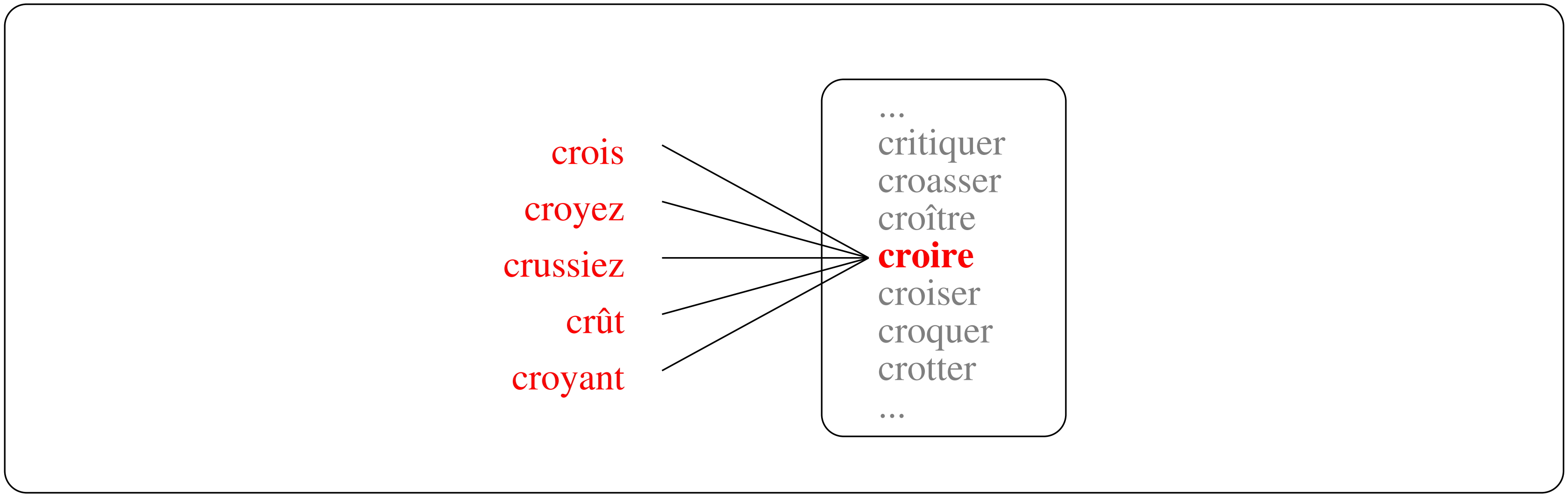
Task Definition

	Morphological Analysis	**Morphological Generation**
Input	inflection *burned*	root, part of speech *burn, VBD*
Output	root, (optional) part of speech *burn, VBD [Past Indicative]* *burn, VBN [Past Participle]*	inflection *burnt* *burned*

- Notice that both morphological analysis and morphological generation can often generate multiple correct answers.
- When the part-of-speech is omitted in morphological analysis, the task is often referred to as morphological stemming.

Major applications of computational morphology

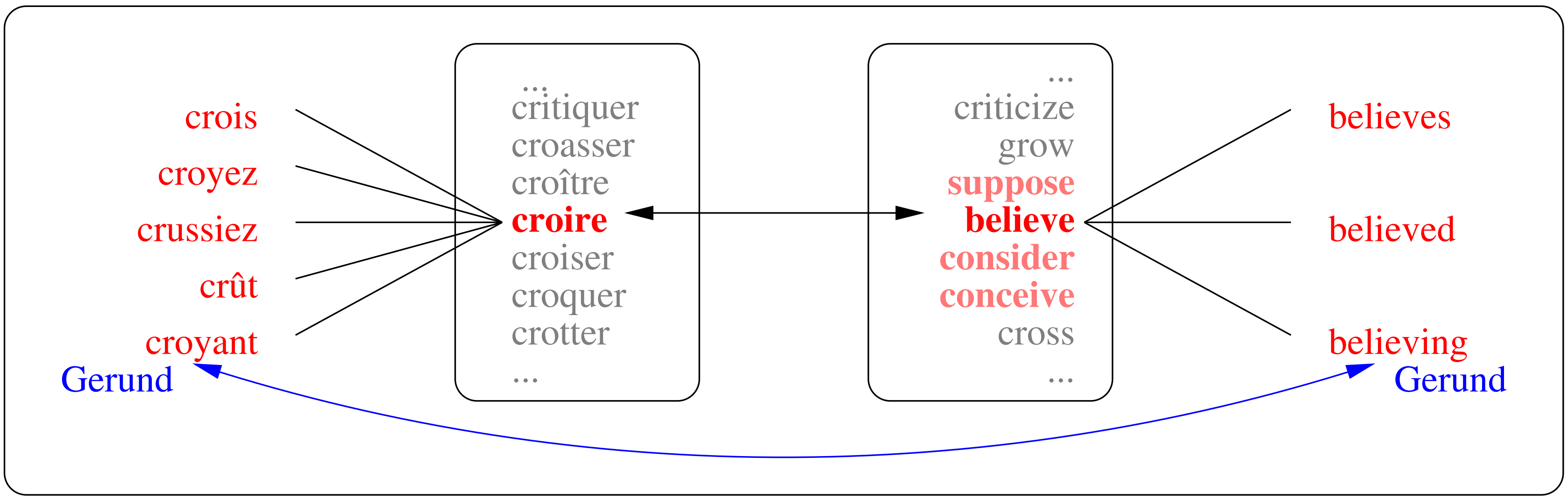
- Information retrieval
  - dimensionality reduction



- Machine translation
  - translation lexicon access, dimensionality reduction for contextual features, fine grained part of speech
- Other applications: parsing, word sense disambiguation, text generation, part of speech tagging

Major applications of computational morphology

- Information retrieval
  - dimensionality reduction
- Machine translation
  - translation lexicon access, dimensionality reduction for contextual features, fine grained part of speech



- Other applications: parsing, word sense disambiguation, text generation, part of speech tagging

Alignment Paradigm

- Prior approaches have focused (almost) exclusively on learning string transductions.
- This makes learning irregular morphology, and pairs such as the following, difficult:

sang	$\leftrightarrow$	sing
singed	$\leftrightarrow$	singe

- How can we learn that the past tense of *sing* isn't *singed*?
- Possible answer: a large amount of information for inflection-root mappings is available outside the string:

Context similarity

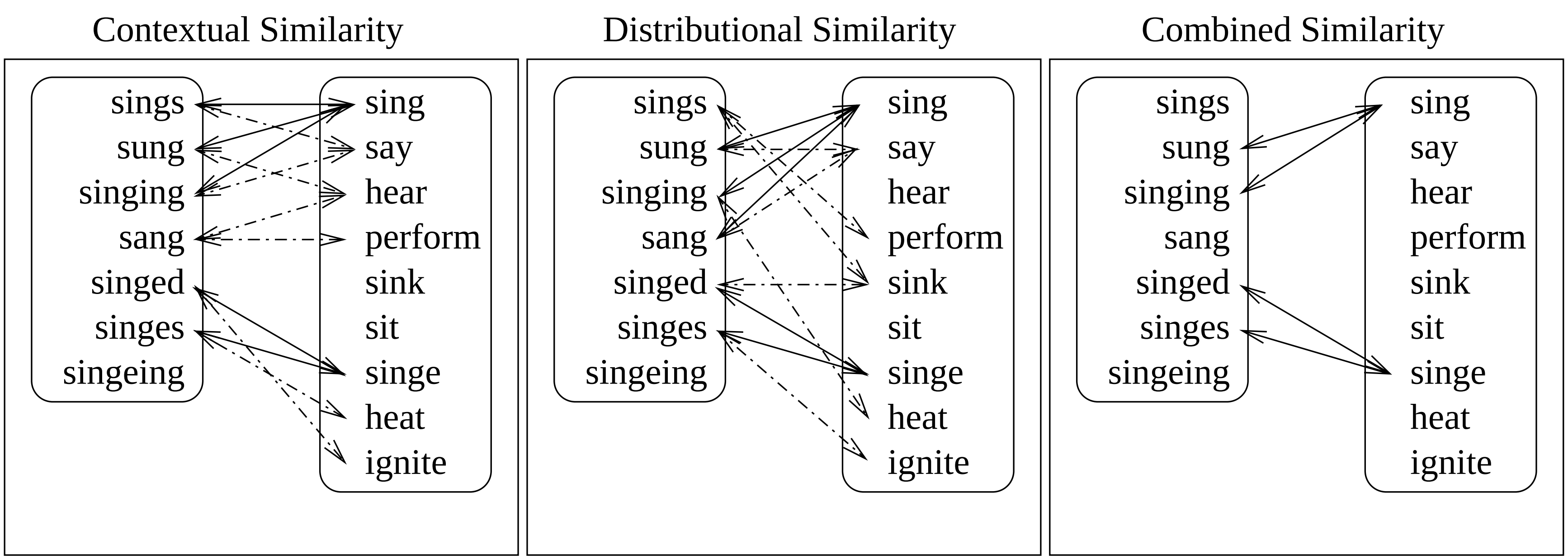
sing	songs
sang	songs
singe	hair
singes	hair

Distributional similarity

sing	1204
sang	1427
singe	2
singes	9

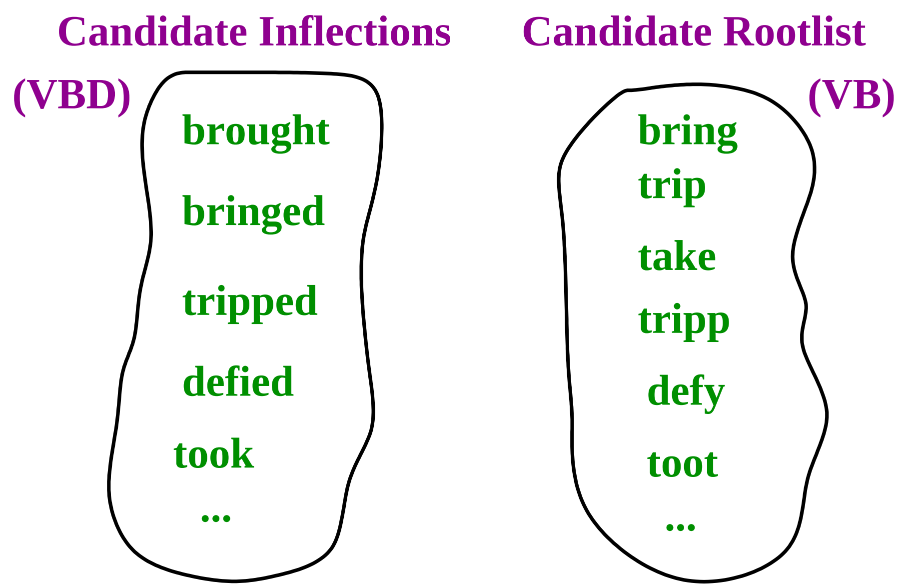
Alignment Paradigm

- Treat morphological analysis as an inflection-root mapping problem.
- Use multiple similarity measures and find a consensus answer.
  - Positionally weighted contextual similarity
  - Distributional similarity (frequency)
  - Weighted Levenshtein similarity (string-edit distance)
  - Bilingual projection



Resource Assumptions

- Noisy wordlists of inflection and root candidates for the language:



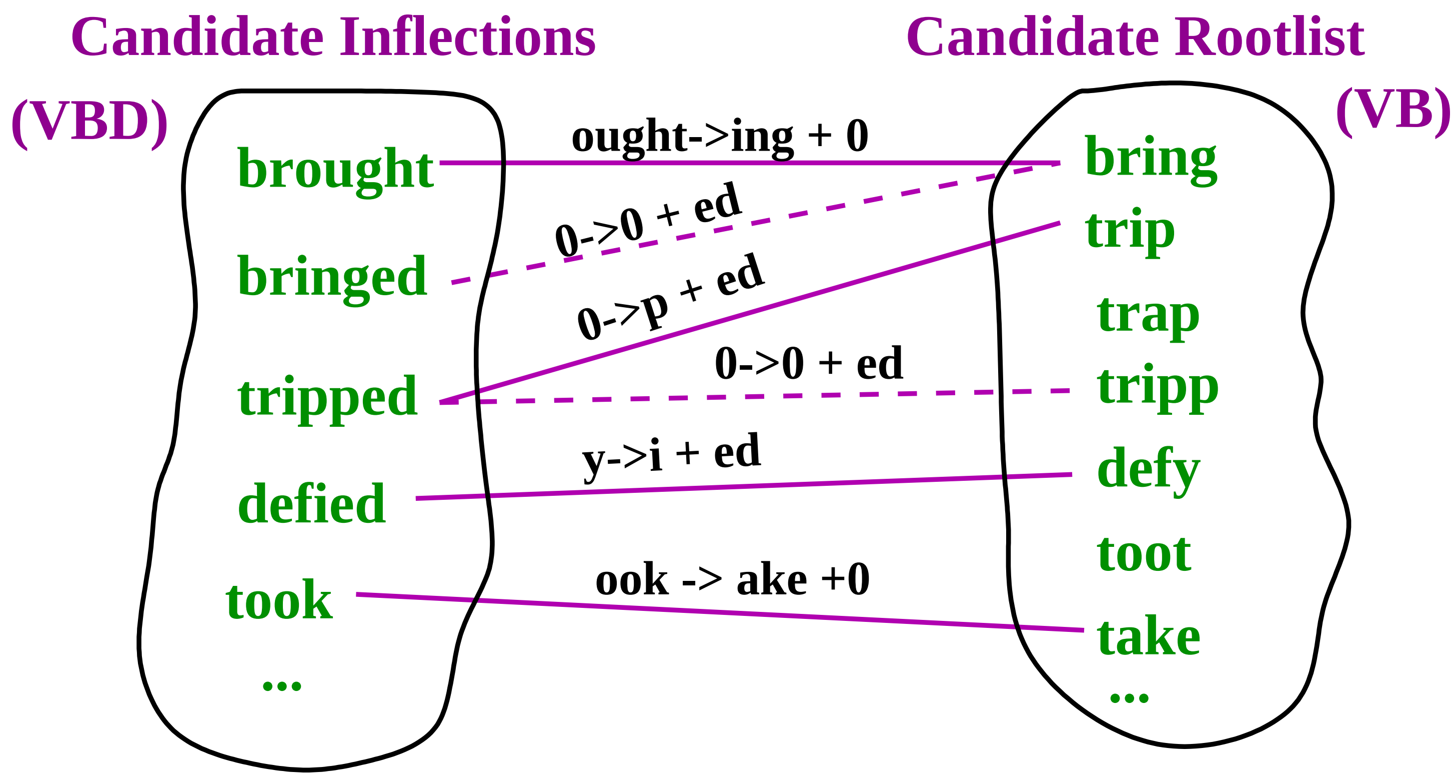
- Canonical suffixes of the language (optional):

Part of Speech	VB	VBD	VBZ	VBG	VBN
Canonical Suffixes	$+\epsilon$	+ed $+\epsilon$	+s	+ing	+en +ed $+\epsilon$

- Lots of plain text

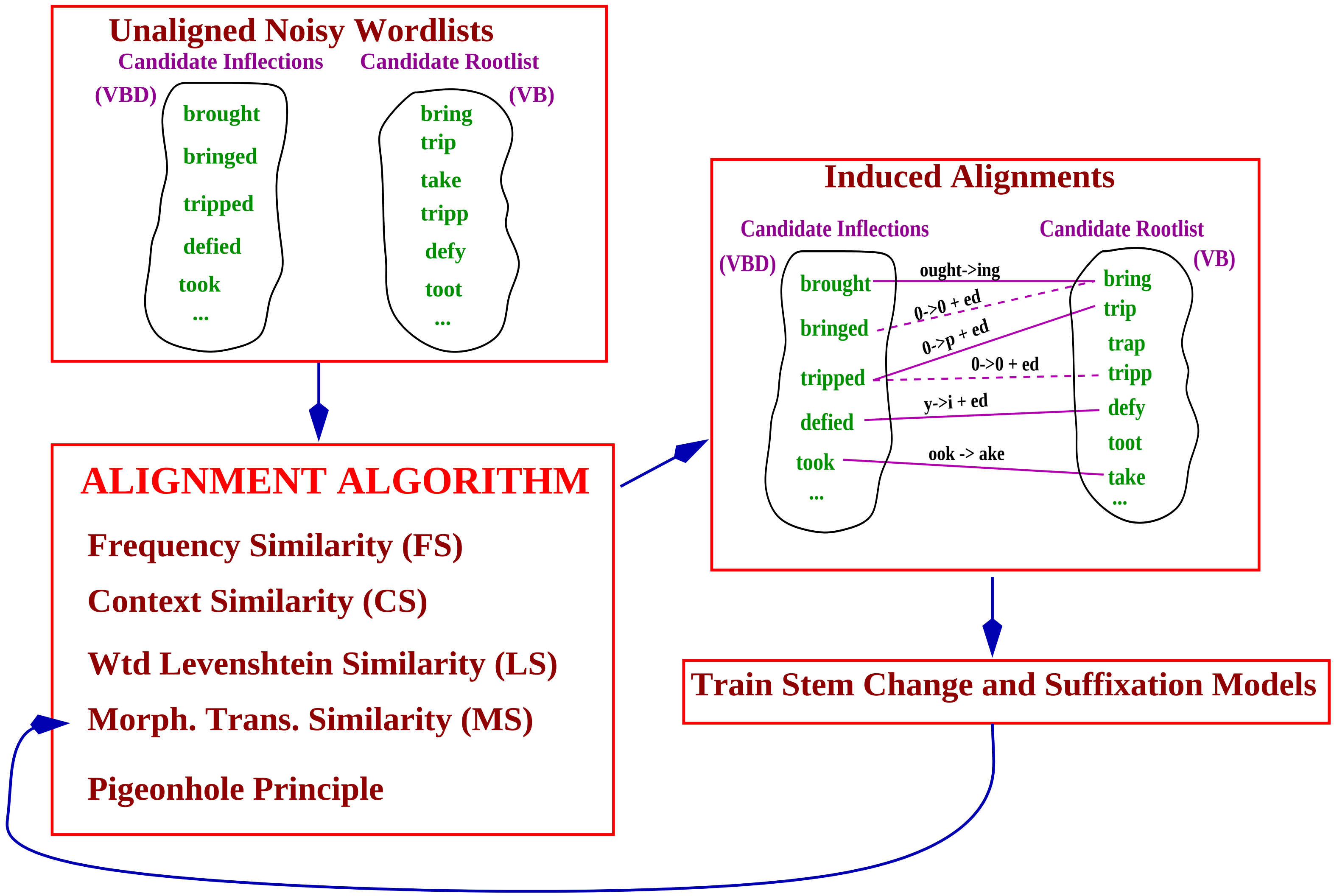
Approach

- Treat analysis as probabilistic alignment over large wordlists.



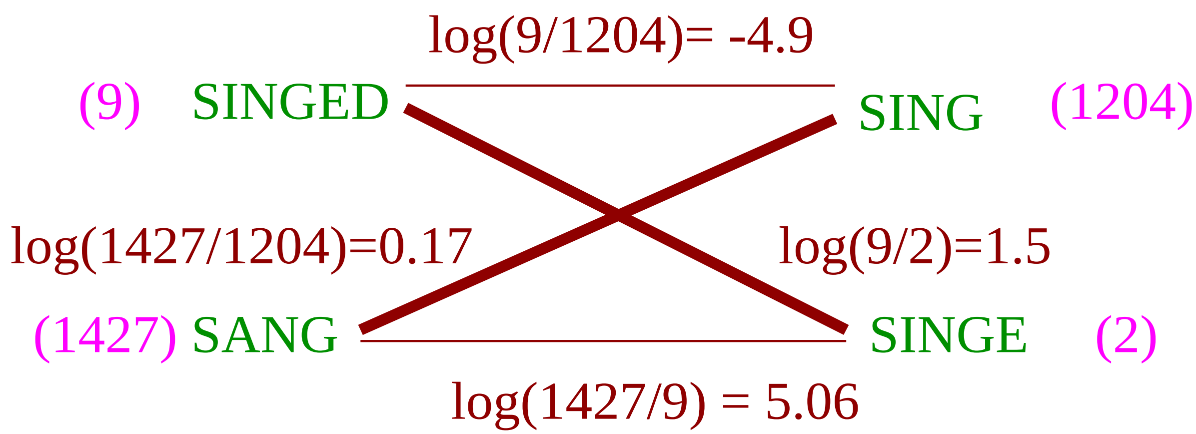
- Use these alignments to train string transduction models

Iterative Bootstrapping of Similarity Models



Distributional Similarity Models

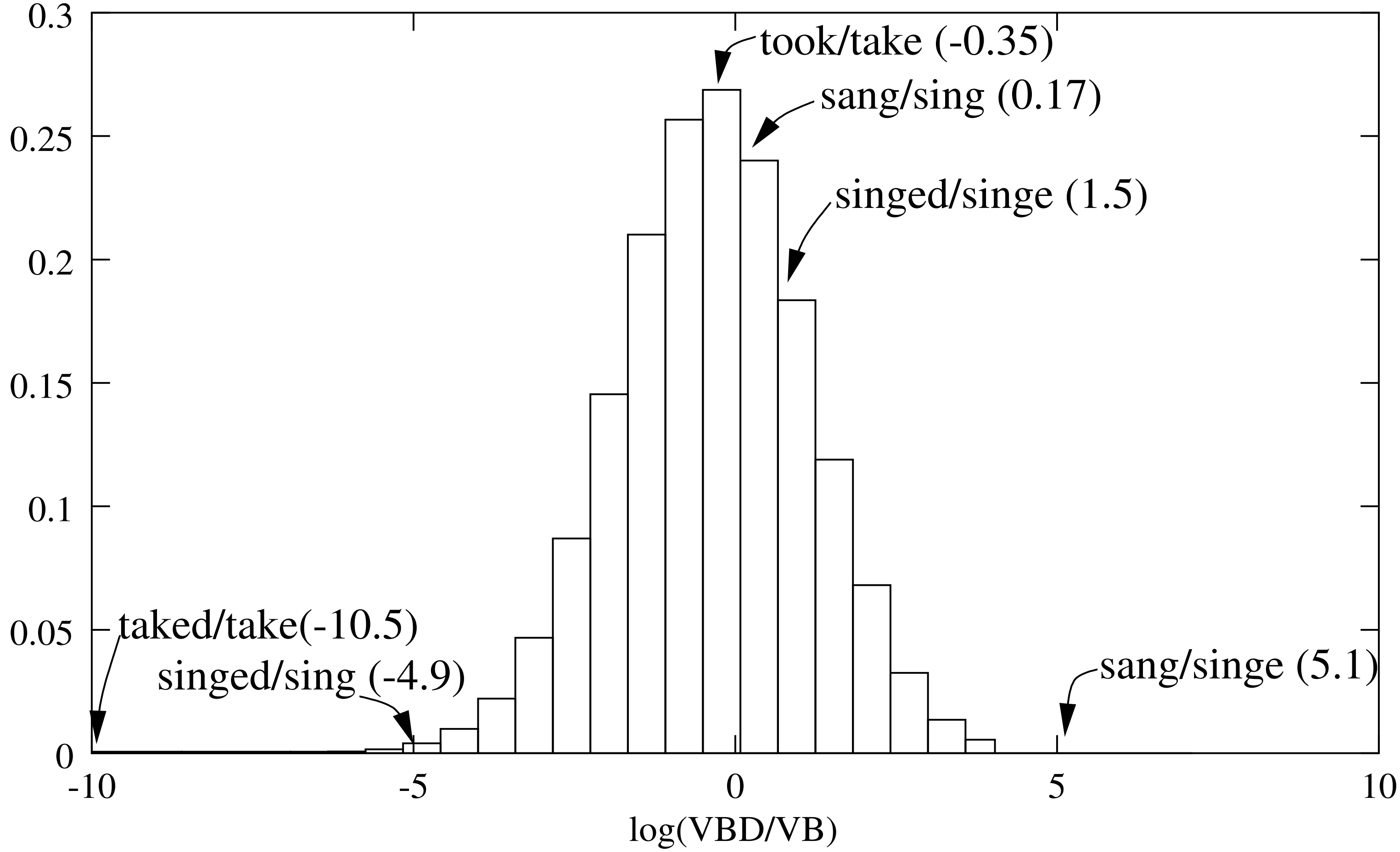
Favor inflection/root alignments with "good match" of frequencies



- How to quantify "good match"?
- How to penalize divergence?

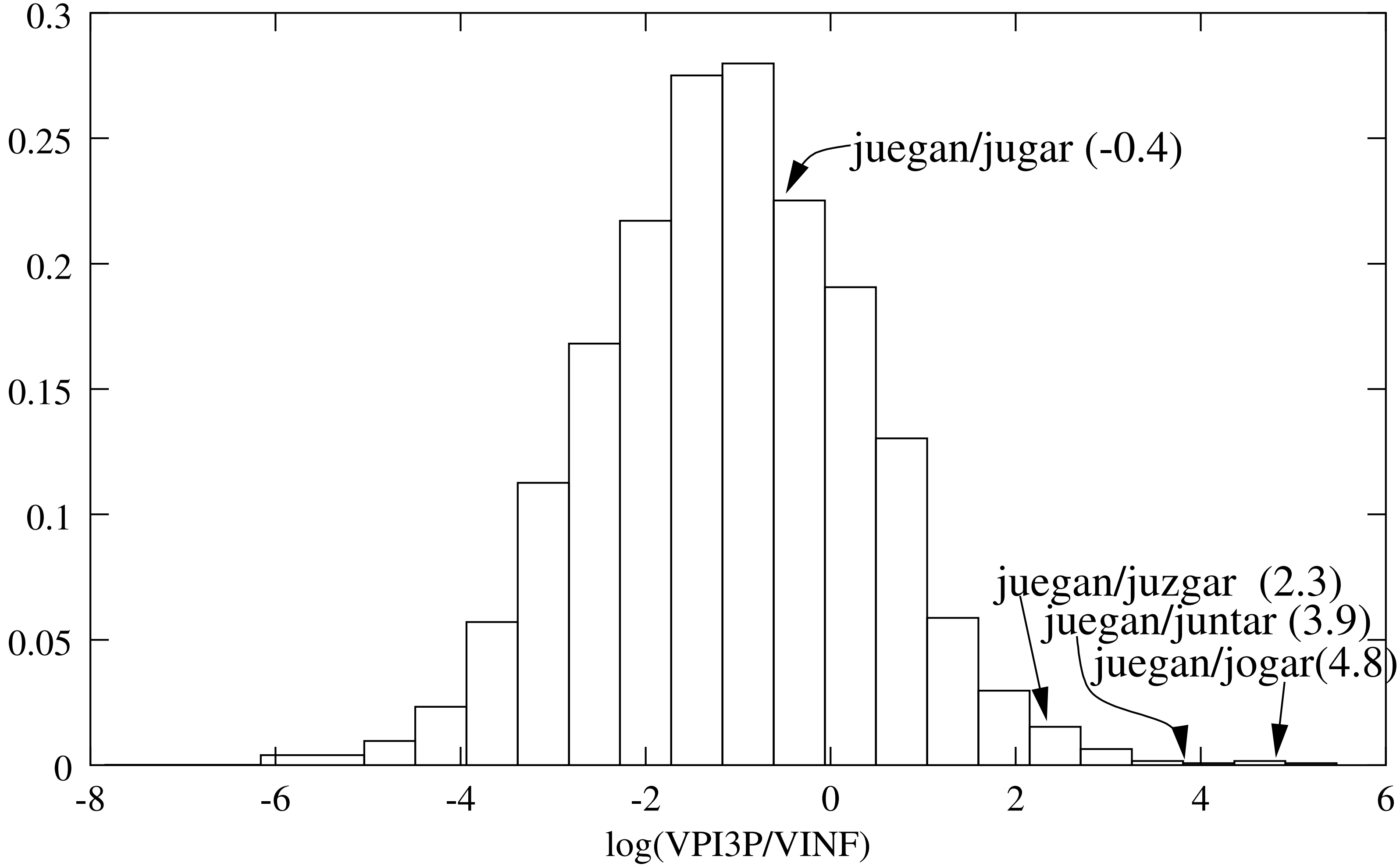
Distributional Similarity Models

Use empirical distributions of frequency ratios
to measure "goodness" of fit.



E[log(VBD/VB)] = -0.24

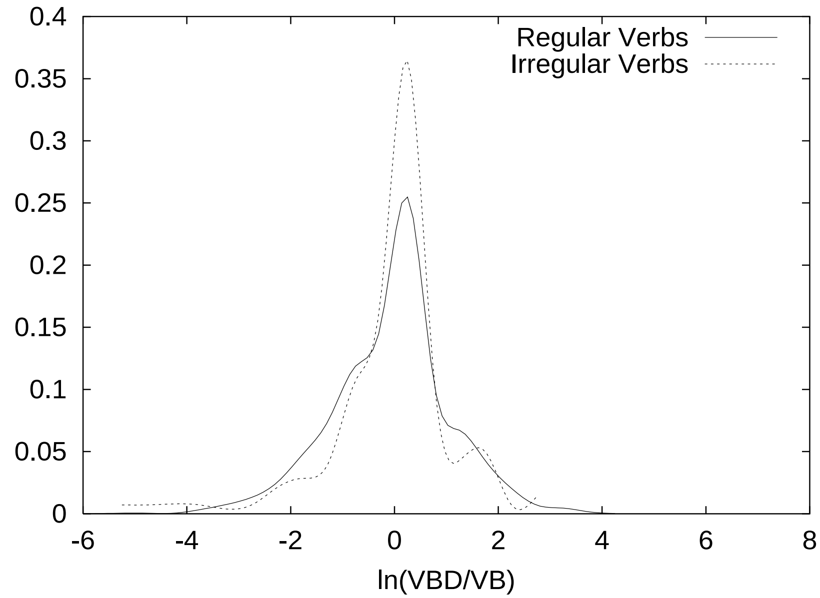
Distributional Similarity Models (Spanish)



E[log(VP13P/VINF)] = -1.21

Approximating Full Empirical Ratio Distributions

VerbType	$\frac{VBD}{VB}$	$\frac{VBG}{VB}$	Avg. Lemma Freq
Regular	.847	.746	861
Irregular	.842	.761	17406

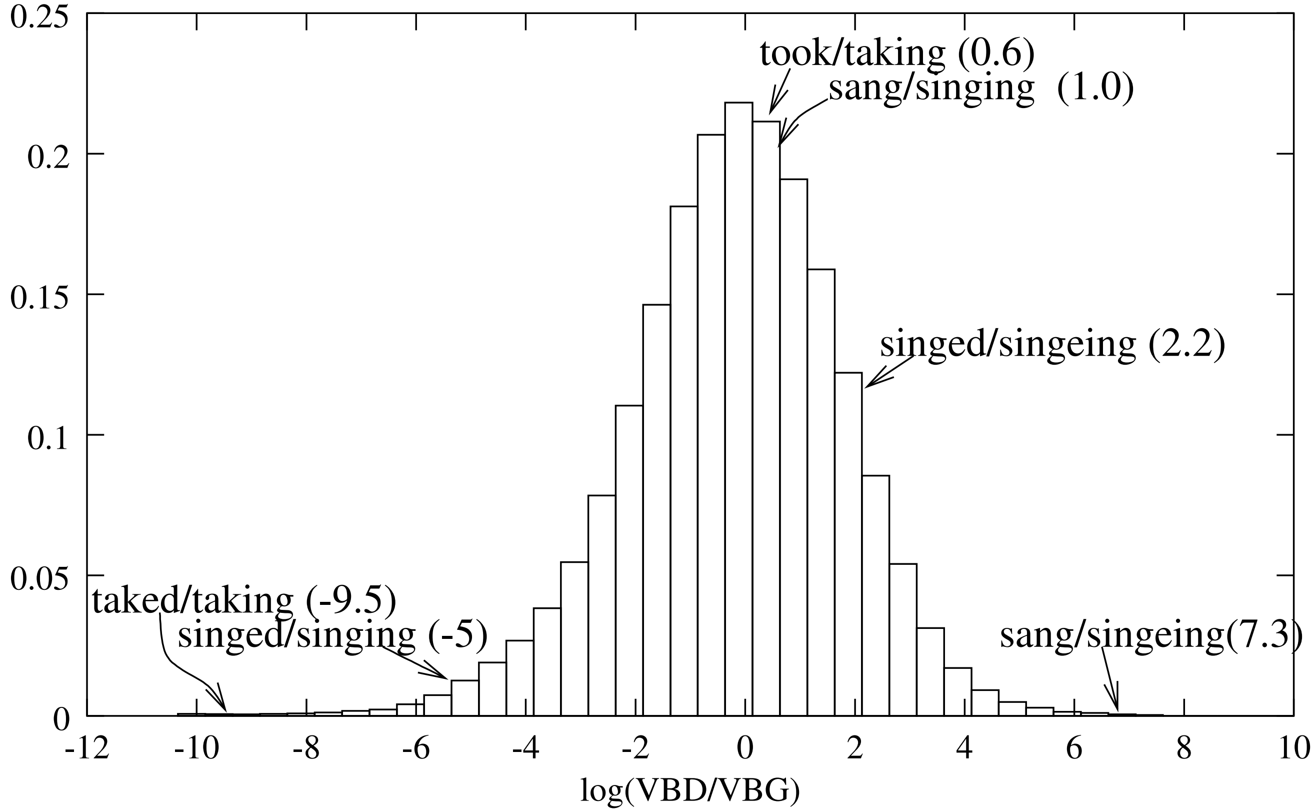


- Tense distribution of verb inflections not correlated with their regularity or stem changing properties
- Estimate initial empirical distributions from most confidently alignable pairs (often regular inflections) from other models

Multiple Ratio Estimators of Robustness

Comparing the distributional frequency directly with other inflections are also informative

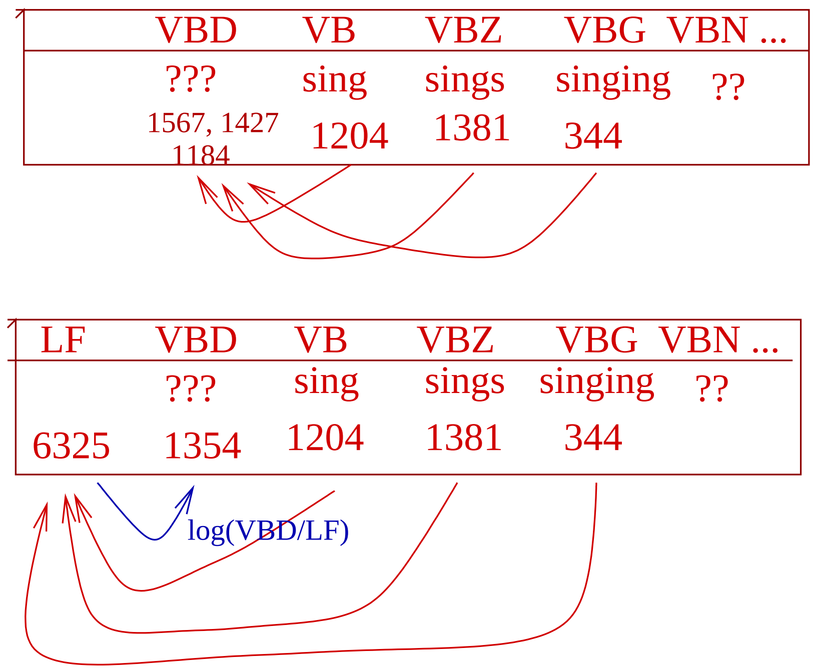
e.g. f(VBD)/f(VBG) or f(VBD)/f(VBZ)



- Any one individual inflection ratio can vary widely due to poor alignment or idiosyncratic tense usage

Reducing Frequency Model Dimensionality

- Use hidden variable of Estimated Lemma Frequency (LF) to capture and smooth multiple pairwise indicators



- Decreases pairwise model dimensionality ($T^2 \rightarrow T$)
- Improves robustness and coverage (all important for highly inflected languages)

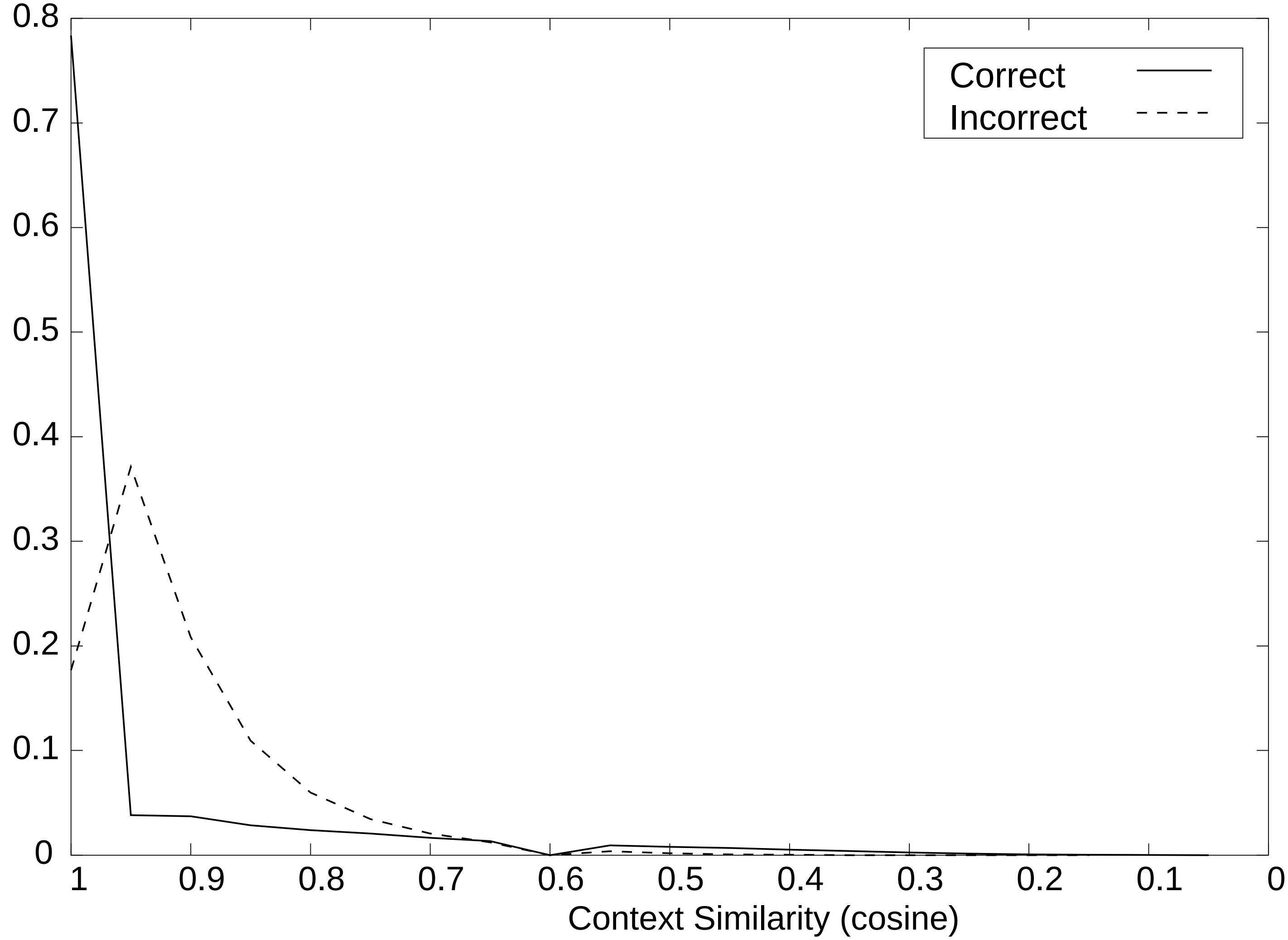
Context Similarity

Measure cosine similarity between aggregate, position-weighted context vectors

	hands	head	baby	violently	himself	deer	away
shook	128	103	21	17	-	-	1
shake	151	98	8	12	-	-	-
shoot	-	-	-	-	56	8	1
shoo	-	-	-	-	-	-	6

- Pool of basic regular expressions to locate potential salient positions.
- Regex choice and relative weighting optimized empirically on strongest alignments from other models.

Context similarity distributions for correctly and incorrectly aligned inflection-root pairs (French)



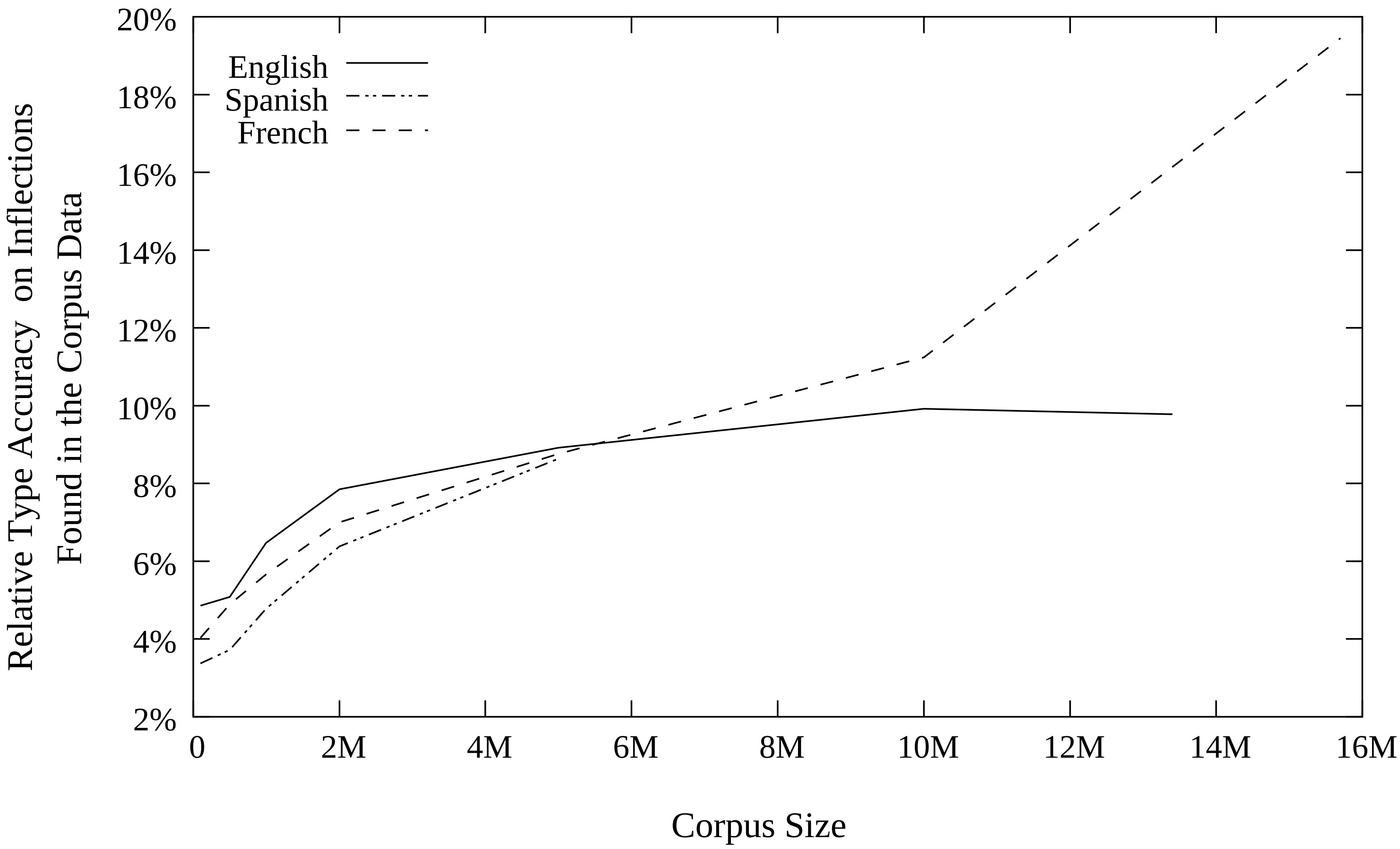
Context model sensitivity to window position

Language	Left 6...0	5...1	4...2	Center 3...3	2...4	1...5	Right 0...6
S-V-O							
Portuguese	12.05%	21.38%	27.55%	26.58%	29.01%	29.01%	**32.87%**
Estonian	31.87%	41.43%	44.01%	42.20%	44.28%	**44.70%**	32.21%
Free / S-V-O							
Russian	19.91%	41.08%	47.35%	40.91%	**47.90%**	47.39%	47.35%
Verb Second (V2)							
German	9.97%	11.34%	13.09%	**14.78%**	13.78%	13.34%	9.96%
S-O-V							
Turkish	**52.66%**	49.06%	48.15%	44.40%	47.03%	45.41%	25.28%
Basque	**25.88%**	21.35%	21.91%	19.75%	21.66%	19.81%	6.44%

- Bag of 6 words surrounding target word
  - 6...0: $word_1$ $word_2$ $word_3$ $word_4$ $word_5$ $word_6$ target
  - 3...3: $word_1$ $word_2$ $word_3$ target $word_4$ $word_5$ $word_6$
  - 0...6: target $word_1$ $word_2$ $word_3$ $word_4$ $word_5$ $word_6$

Language	Left 6...0	Center 3...3	Right 0...6
S-V-O			
Spanish	6.37%	20.31%	**29.38%**
Portuguese	12.05%	26.58%	**32.87%**
French	9.08%	38.40%	**45.60%**
Italian	3.69%	9.99%	**14.98%**
Romanian	10.42%	18.71%	**20.86%**
English	13.25%	21.98%	**25.67%**
Danish	7.21%	24.61%	**34.59%**
Swedish	2.09%	10.36%	**18.69%**
Icelandic	10.93%	23.43%	**29.98%**
Estonian	31.87%	**42.20%**	32.21%
Finnish	5.40%	12.09%	**12.15%**
Tagalog	10.10%	15.08%	**17.08%**
Swahili	8.63%	8.68%	**11.02%**
Free / S-V-O			
Czech	3.30%	**11.05%**	11.02%
Polish	8.16%	18.91%	**20.87%**
Russian	19.91%	40.91%	**47.35%**
Verb Second (V2)			
German	9.97%	**14.78%**	9.96%
Dutch	11.78%	**16.32%**	15.50%
S-O-V			
Turkish	**52.66%**	44.40%	25.28%
Basque	**25.88%**	19.75%	6.44%

Context model sensitivity to corpus size



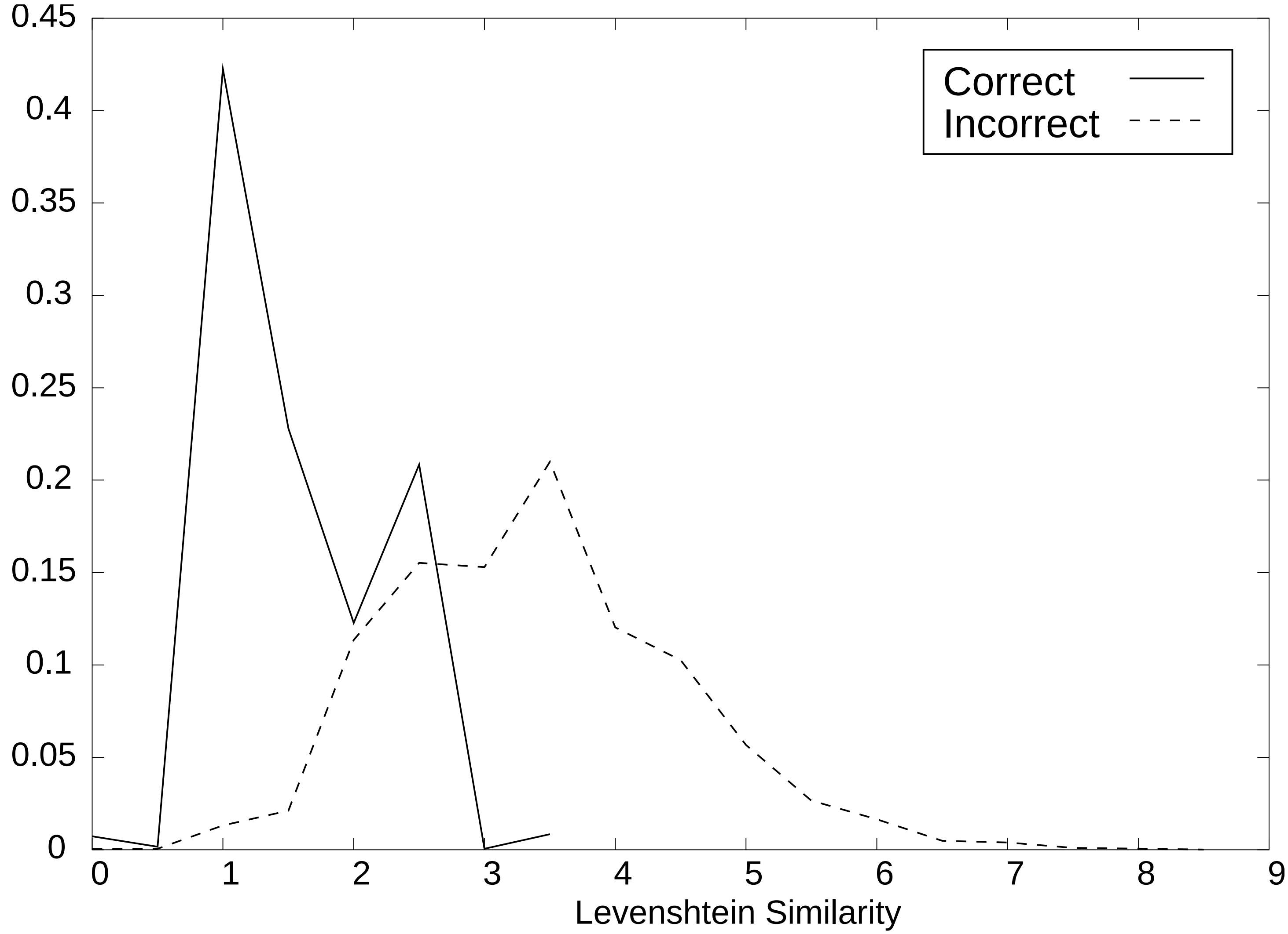
Weighted Levenshtein Similarity

- Measure of string edit distance
- Transition cost matrix weighted by relative rarity of letter change in current paired data

	a	o	ue	m	n	...
a	0	δ_1	δ_2	δ_4	δ_4	...
o	δ_1	0	δ_2	δ_4	δ_4	...
ue	δ_2	δ_2	0	δ_4	δ_4	...
m	δ_4	δ_4	δ_4	0	δ_3	...
n	δ_4	δ_4	δ_4	δ_3	0	...
...	...	...	...	...	...	...

- Initially set to 4 basic parameters for V-V, V^+-V^+, C-C, C-V^+ in current paired data
- Cost matrix re-estimated on subsequent alignments

Levenshtein similarity distributions for correctly and incorrectly aligned inflection-root pairs (English)



Language	Prefix Penalty			No	Suffix Penalty		
	1	0.5	0.25	Penalty	0.25	0.5	1.0
Spanish	**94.13%**	93.88%	93.36%	91.62%	63.38%	69.64%	76.08%
Portuguese	**96.26%**	96.01%	95.38%	93.74%	71.79%	76.24%	80.89%
Catalan	**92.01%**	91.37%	90.51%	88.02%	71.84%	74.96%	79.17%
Occitan	**92.64%**	92.40%	91.98%	88.63%	67.28%	71.21%	75.14%
French	**93.44%**	92.80%	91.78%	88.09%	69.17%	72.05%	76.45%
Italian	**95.26%**	94.86%	94.37%	91.79%	71.58%	75.70%	80.74%
Romanian	**91.14%**	90.48%	89.34%	82.67%	46.08%	51.93%	58.64%
Latin	**85.35%**	84.67%	82.89%	70.60%	21.66%	26.47%	34.02%
English	**93.36%**	92.38%	89.82%	84.80%	46.96%	54.28%	61.76%
Danish	**94.77%**	93.31%	92.69%	90.94%	70.44%	76.80%	81.00%
Norwegian	**94.47%**	93.81%	93.14%	91.40%	72.42%	78.51%	82.96%
Swedish	**90.24%**	88.93%	87.15%	82.74%	42.88%	51.00%	59.83%
Icelandic	**91.33%**	90.95%	90.30%	88.65%	62.38%	67.55%	74.11%
Hindi	**96.88%**	96.48%	**96.88%**	96.48%	87.50%	87.50%	88.67%
Sanskrit	**78.34%**	77.39%	75.75%	67.43%	29.68%	34.41%	41.58%
Estonian	**81.86%**	81.20%	80.70%	78.52%	62.59%	65.71%	69.17%
Tamil	**89.61%**	87.60%	87.27%	83.42%	62.31%	69.01%	74.54%
Finnish	**74.86%**	73.57%	71.88%	62.88%	27.35%	32.43%	39.52%
Turkish	**95.03%**	94.69%	94.09%	89.63%	48.85%	55.83%	64.61%
Uzbek	**84.67%**	84.43%	83.89%	81.02%	51.19%	55.21%	60.40%
Basque	**80.74%**	79.85%	78.69%	73.91%	38.45%	44.21%	49.47%
Czech	76.49%	76.40%	76.42%	**78.26%**	67.13%	70.13%	72.79%
Polish	**93.22%**	93.02%	92.78%	91.71%	68.54%	73.57%	78.83%
Russian	**84.73%**	83.41%	82.12%	80.87%	66.33%	69.59%	72.96%
German	**91.58%**	91.53%	91.39%	91.44%	82.02%	84.04%	86.34%
Dutch	**80.49%**	80.11%	80.22%	78.08%	69.56%	71.59%	73.79%
Irish	**92.89%**	**92.89%**	92.21%	87.75%	48.71%	54.16%	62.18%
Welsh	**85.99%**	84.81%	83.52%	76.50%	35.41%	42.08%	50.69%
Tagalog	20.27%	23.80%	28.35%	61.03%	72.24%	**72.32%**	71.40%
Swahili	29.83%	34.94%	41.69%	68.38%	**79.95%**	79.35%	78.31%
Klingon	24.84%	27.34%	30.46%	98.74%	**99.55%**	99.17%	99.08%

Using Alignment Models to Bootstrap String Transduction Models

- Combine alignment models to get training data
- Models each have different dynamic ranges
  - Levenshtein Similarity: [0,inf) $\Rightarrow$ lower score is better
  - Context Similarity: [0,1] $\Rightarrow$ higher score is better
  - Frequency Similarity (-inf,inf)
- Bidirectional averaged relative rankings

$$
\begin{aligned}
&sim(root_i, inf_j) = \\
&\quad \lambda_F(\text{rank}_{fs}(root_i|inf_j)+\text{rank}_{fs}(inf_j|root_i))+ \\
&\quad \lambda_C(\text{rank}_{cs}(root_i|inf_j)+\text{rank}_{cs}(inf_j|root_i))+ \\
&\quad \lambda_L(\text{rank}_{ls}(root_i|inf_j)+\text{rank}_{ls}(inf_j|root_i))+ \\
&\quad \lambda_M(\text{rank}_{ms}(root_i|inf_j)+\text{rank}_{ms}(inf_j|root_i))
\end{aligned}
$$

- Favors mutual affinity

Choosing λ's for Model Combination

$$
\begin{aligned}
&sim(root_i, inf_j) = \\
&\quad \lambda_F(\text{rank}_{fs}(root_i|inf_j) + \text{rank}_{fs}(inf_j|root_i)) + \\
&\quad \lambda_C(\text{rank}_{cs}(root_i|inf_j) + \text{rank}_{cs}(inf_j|root_i)) + \\
&\quad \lambda_L(\text{rank}_{ls}(root_i|inf_j) + \text{rank}_{ls}(inf_j|root_i)) + \\
&\quad \lambda_M(\text{rank}_{ms}(root_i|inf_j) + \text{rank}_{ms}(inf_j|root_i))
\end{aligned}
$$

- Initially, Levenshtein model will produce the cleanest training data for bootstrapping $\Rightarrow$ set $\lambda_L >> \lambda_C$ and λ_F
- The output of the string transduction model can then be used to refine the alignment models
- As context and frequency models are refined, λ_C and λ_F are increased relative to λ_L

Comparison of Models (English)

Combination of Similarity Models	# of Iterations	All Words (3888)	Highly Irregular (128)	Regular (1877)	Semi-Regular (1883)
FS (*Frequency Sim*)	(Iter 1)	9.8	18.6	8.8	10.1
LS (*Levenshtein Sim*)	(Iter 1)	31.3	19.6	20.0	34.4
CS (*Context Sim*)	(Iter 1)	28.0	32.8	30.0	25.8
CS+FS	(Iter 1)	32.5	64.8	32.0	30.7
CS+FS+LS	(Iter 1)	71.6	76.5	71.1	71.9
CS+FS+LS+MS	(Iter 1)	96.5	74.0	97.3	97.4
CS+FS+LS+MS	(Convg)	**99.2**	**80.4**	**99.9**	**99.7**
Mooney&Califf		82.5	5.0	100.0	84.0

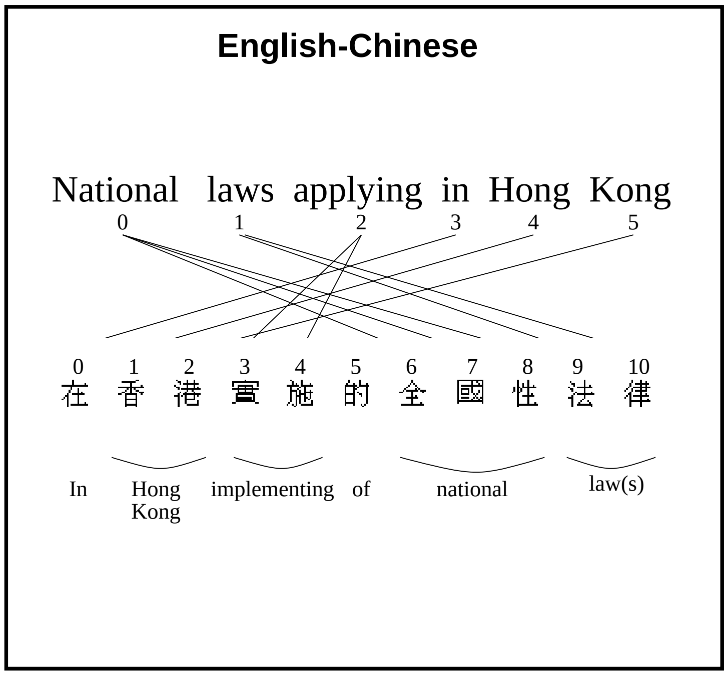
- Frequency and Context similarity models assume that words must start with the same initial letter at first iteration. This is not an unreasonable assumption for suffixal languages.

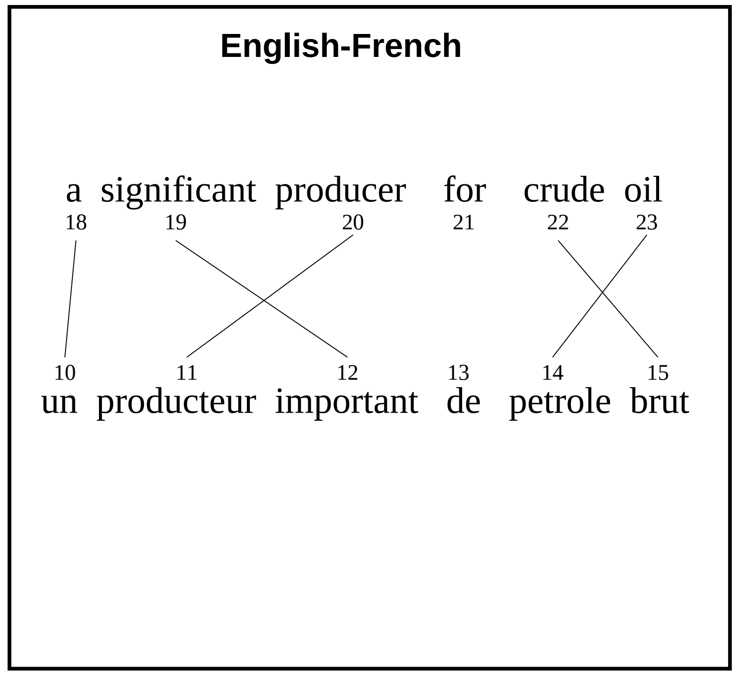
Performance on Irregular Verb Sample

Word	True Root	CS+FS+LS+MS (Convg)	Score	CS+FS+LS+MS (Itr 1)	CS+FS+LS (Itr 1)	CS+FS (Itr 1)	LS only (Itr 1)
got	get	go	1.30	go	go	go	gut
took	take	**take**	1.50	**take**	**take**	**take**	toot
became	become	**become**	2.35	**become**	**become**	**become**	**become**
clung	cling	**cling**	2.55	**cling**	**cling**	**cling**	**cling**
swore	swear	**swear**	2.80	**swear**	**swear**	**swear**	store
came	come	**come**	3.55	**come**	**come**	**come**	**come**
flung	fling	**fling**	4.60	**fling**	**fling**	**fling**	**fling**
strove	strive	**strive**	5.85	**strive**	**strive**	straddle	**strive**
swept	sweep	**sweep**	6.20	**sweep**	**sweep**	**sweep**	swap
woke	wake	**wake**	6.95	**wake**	**wake**	wind	**wake**
bore	bear	**bear**	7.75	**bear**	bar	**bear**	bare
lent	lend	**lend**	9.25	**lend**	**lend**	**lend**	**lend**
struck	strike	**strike**	11.60	**strike**	**strike**	**strike**	strut
bit	bite	**bite**	13.60	**bite**	**bite**	betray	bet
dove	dive	**dive**	17.25	**dive**	**dive**	dash	**dive**
caught	catch	**catch**	18.35	**catch**	cut	**catch**	cough
dealt	deal	**deal**	21.45	**deal**	**deal**	disagree	**deal**

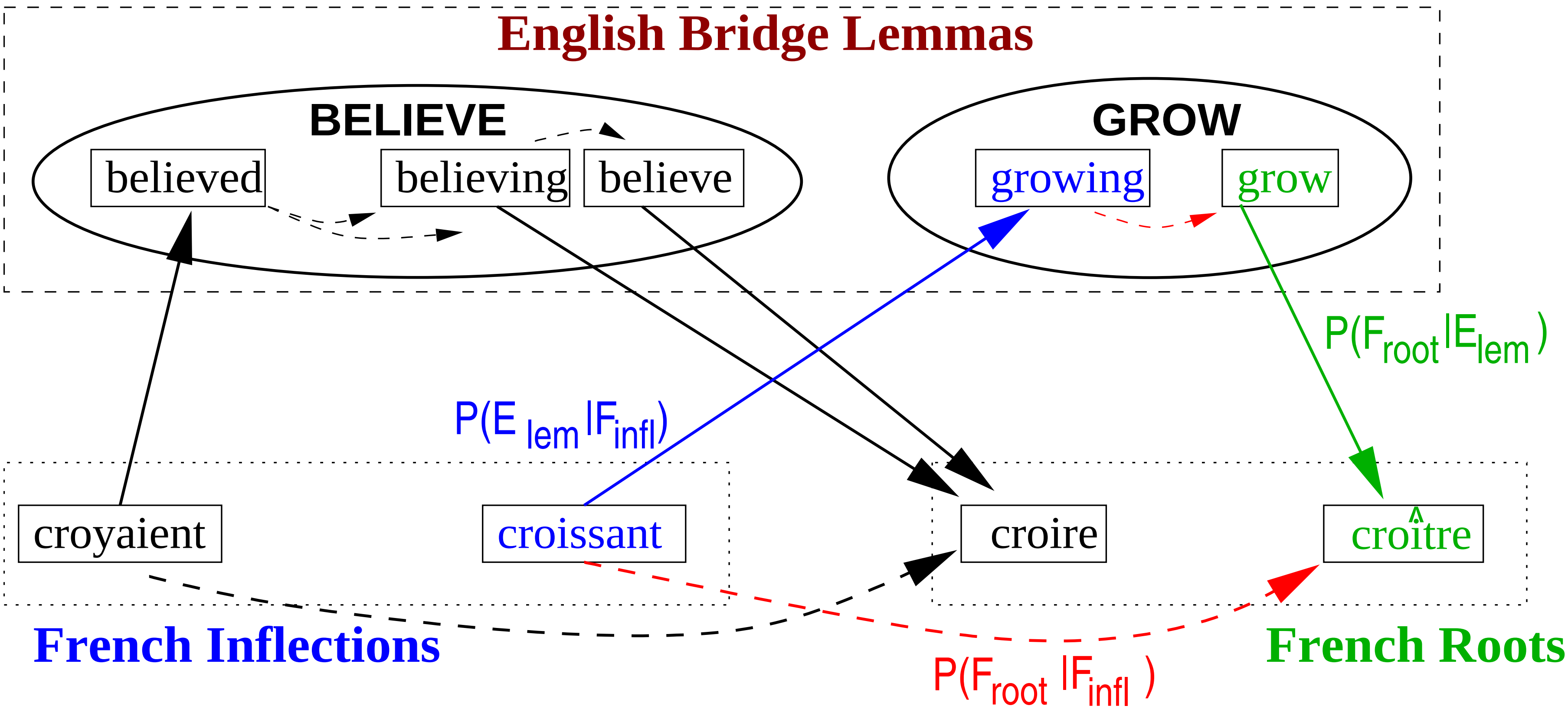
Multilingual Projection

- Historically, large NLP investment in tools for English and a handful of other languages (French, Japanese)
- Use existing morphological analyzers to bootstrap training data for morphological analyzer for a second language
- Word align bilingual corpus (EGYPT, Y. Al-Onaizan et al. 1999):

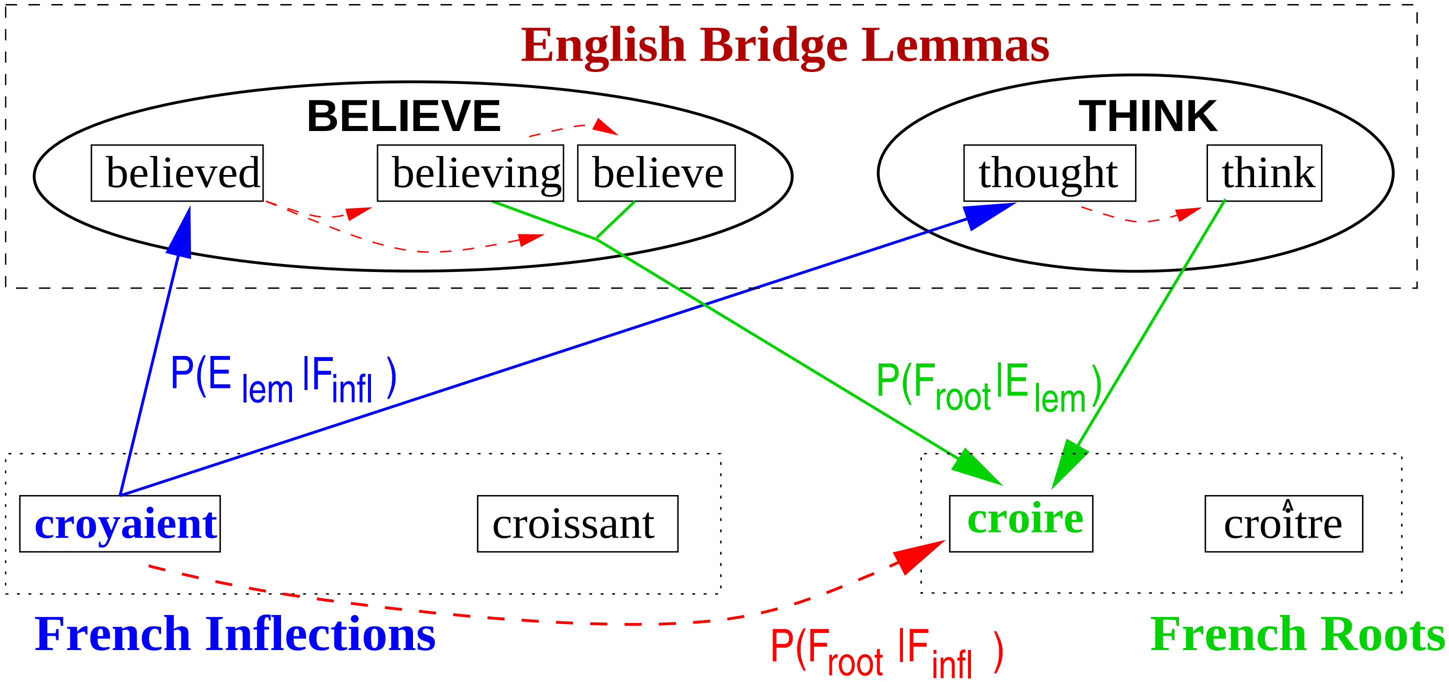




Morphological Analysis via Translingual Bridges: Leverage investment in existing analyzers



General Alignment Model via Multiple Bridge Lemmas



$$P_{mp}(F_{root}|F_{infl}) = \sum_i P_a(F_{root}|E_{lem_i})\, P_a(E_{lem_i}|F_{infl})$$

$$\begin{aligned} P_{mp}(\text{croire}|\text{croyaient}) = \\ & P(\text{croire}|\textsc{believe})\, P(\textsc{believe}|\text{croyaient}) + \\ & P(\text{croire}|\textsc{think})\, P(\textsc{think}|\text{croyaient}) + \ldots \end{aligned}$$

Examples of Induced Morphological Analyses

Induced Morphological Analyses for CZECH

Inflection	Root	POST Analysis	TopBridge
bral	brát	át→a +l	marry
brala	brát	át→a +la	accept
brali	brát	át→a +li	marry
byl	být	ýt→y +l	be
byli	být	ýt→y +li	be
bylo	být	ýt→y +lo	be
chovala	chovat	t→ ϵ +la	behave
chová	chovat	at→ ϵ +á	behave
chováme	chovat	at→ ϵ +áme	behave
chodila	chodit	t→ ϵ +la	walk
chodí	chodit	it→ ϵ +í	walk
choďte	chodit	dit→ďt +e	swim
chránila	chránit	t→ ϵ +la	protect
chrání	chránit	it→ ϵ +í	protect
couval	couvat	t→ ϵ +l	back
chcete	chtít	tít→c +ete	want
chceš	chtít	tít→c +eš	want
chci	chtít	tít→c +i	want
chtějí	chtít	ít→ěj +í	want
chtěli	chtít	ít→ě +li	want
chtělo	chtít	ít→ě +lo	want

Induced Morphological Analyses for FRENCH

Inflection	Root	POST Analysis	TopBridge
abrège	abréger	éger→èg +e	shorten
abrègent	abréger	éger→èg +ent	shorten
abrégerai	abréger	er→ ϵ +erai	curtail
achète	acheter	eter→èt +e	buy
achètent	acheter	eter→èt +ent	buy
achètera	acheter	eter→èt +era	buy
adviendrait	advenir	enir→iendr +ait	happen
advient	advenir	enir→ien +t	happen
aliène	aliéner	éner→èn +e	alienate
aliènent	aliéner	éner→èn +ent	alienate
conçu	concevoir	cevoir→ç +u	conceive
crois	croire	re→ ϵ +s	believe
croyaient	croire	ire→y +aient	believe

Induced Morphological Analyses for SPANISH

Inflection	Root	POST Analysis	TopBridge
aborreció	aborrecer	er→ ϵ +ió	hate
aborrecía	aborrecer	er→ ϵ +ía	hate
aborrezco	aborrecer	cer→zc +o	hate
abrace	abrazar	zar→c +e	embrace
abrazado	abrazar	ar→ ϵ +ado	embrace
adquiere	adquirir	rir→er +e	get
anden	andar	ar→ ϵ +en	walk
anduvo	andar	ar→uv +o	walk
buscáis	buscar	ar→ ϵ +áis	seek
busque	buscar	car→qu +e	seek
busqué	buscar	car→qu +é	seek

Use projected pairs as training for supervised models

Model	Precision Typ	Precision Tok	Coverage Typ	Coverage Tok
FRENCH Verbal Morphology Induction				
French Hansards (12M words):				
MProj only	.992	.999	.779	.994
MProj+POST	.998	.999	.988	.999
MProj+POST+BKM	.994	.999	1.00	1.00
French Hansards (1.2M words):				
MProj only	.985	.998	.327	.976
MProj+POST	.995	.999	.958	.998
MProj+POST+BKM	.979	.998	1.00	1.00
French Hansards (120K words):				
MProj only	.962	.931	.095	.901
MProj+POST	.984	.993	.916	.994
MProj+POST+BKM	.932	.989	1.00	1.00
French Bible (300K words) via English Bible:				
MProj only	1.00	1.00	.052	.747
MProj+POST	.991	.998	.918	.992
MProj+POST+BKM	.954	.994	1.00	1.00

Use projected pairs as training for supervised models (continued)

CZECH Verbal Morphology Induction

Czech Reader's Digest (500K words):

MProj only	.915	.993	.152	.805
MProj+POST	.916	.917	.893	.975
MProj+POST+BKM	.878	.913	1.00	1.00

SPANISH Verbal Morphology Induction

Spanish Bible (300K words) via English Bible:

MProj only	.973	.935	.264	.351
MProj+POST	.988	.998	.971	.967
MProj+POST+BKM	.966	.985	1.00	1.00

Spanish Bible (300K words) via French Bible:

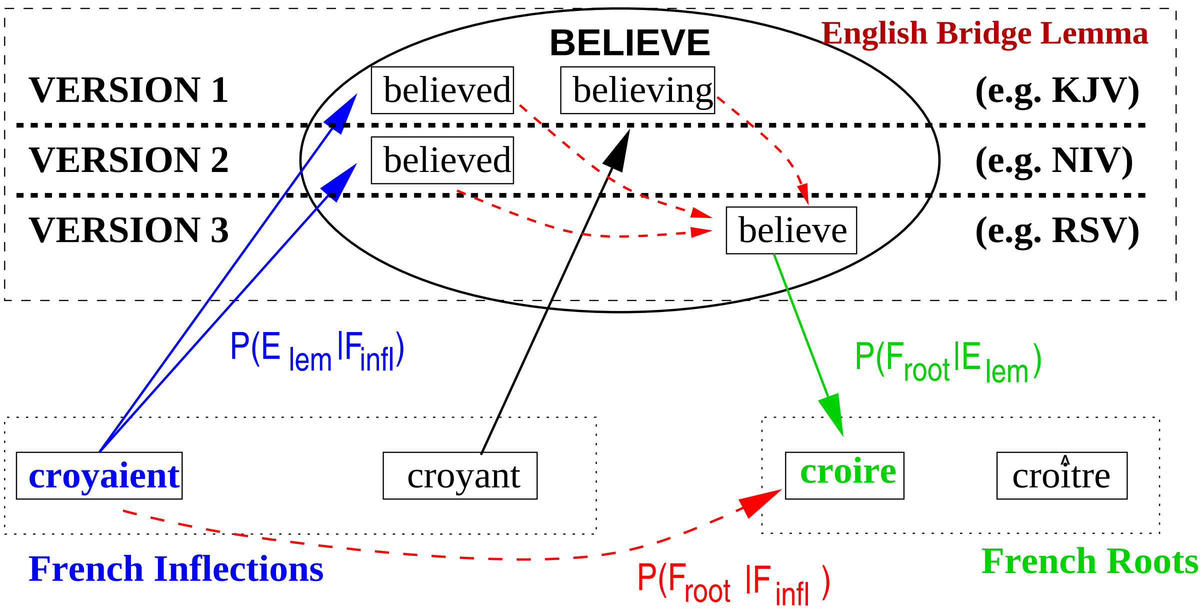
MProj only	.980	.935	.722	.765
MProj+POST	.983	.974	.986	.993
MProj+POST+BKM	.974	.968	1.00	1.00

Lemmatization Induction via Multiple Bible Versions

- Bible is easily alignable, available in <u>many</u> languages

.954 accuracy (by type)	single French Bible
.994 accuracy (by token)	on full *modern French* test set

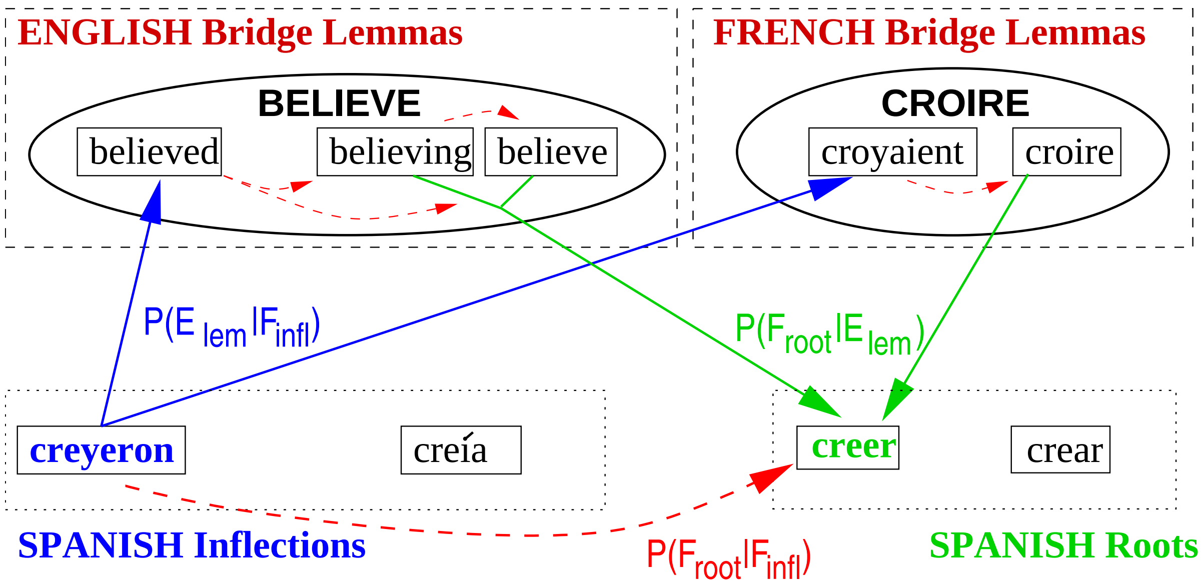
- Augment bridge potential using multiple English versions



⇒ .954 → .964 with *no* additional French resources

Lemmatization Induction via Multiple Languages

- Using bitext in multiple languages adds bridging pathways
- Use newly lemmatized Bible in French to improve Spanish analysis



Spanish via 1 English Bible	.966
Spanish via 1 French Bible	.974
Spanish via English + French Bibles	.981

(accuracy by type)

Use projected pairs as training for supervised models (continued)

Model	Precision Typ	Precision Tok	Coverage Typ	Coverage Tok
FRENCH Verbal Morphology Induction				
French Bible (300K words) via 3 English Bibles:				
MProj only	.928	.975	.100	.820
MProj+POST	.981	.991	.931	.990
MProj+POST+BKM	.964	.991	1.00	1.00
SPANISH Verbal Morphology Induction				
Spanish Bible (300K words) via 3 English Bibles:				
MProj only	.964	.948	.468	.551
MProj+POST	.990	.998	.978	.987
MProj+POST	.976	.987	1.00	1.00

Performance of Lemmatization Induction by Corpus Size