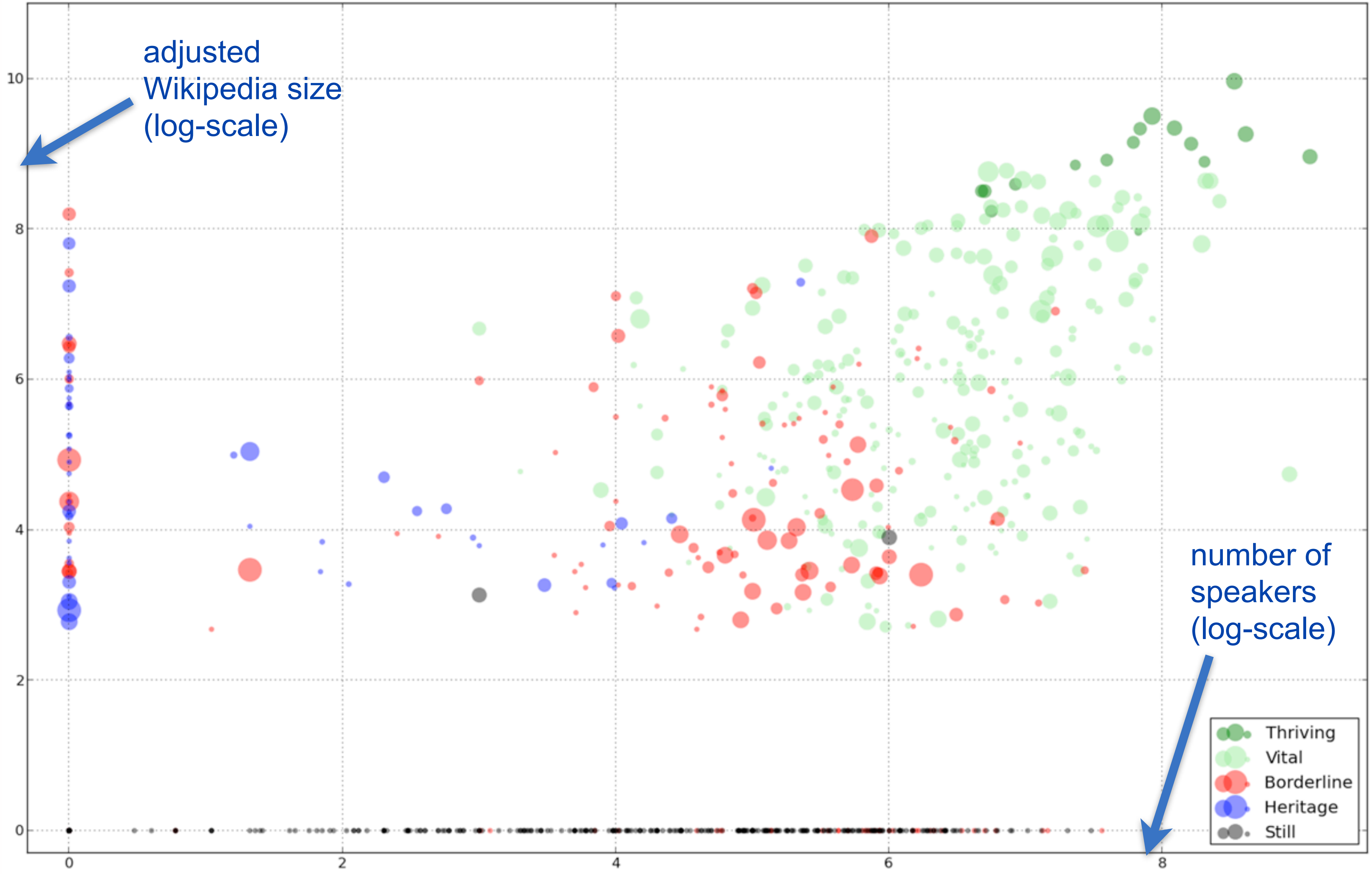
Languages are Dialects with a Treebank and a Dependency Parser

Cross-Lingual Parsing for Low-Resource Languages

Jörg Tiedemann
Department of Modern Languages
University of Helsinki

Digital Language Death (Kornai, 2013)



adjusted Wikipedia size (log-scale)

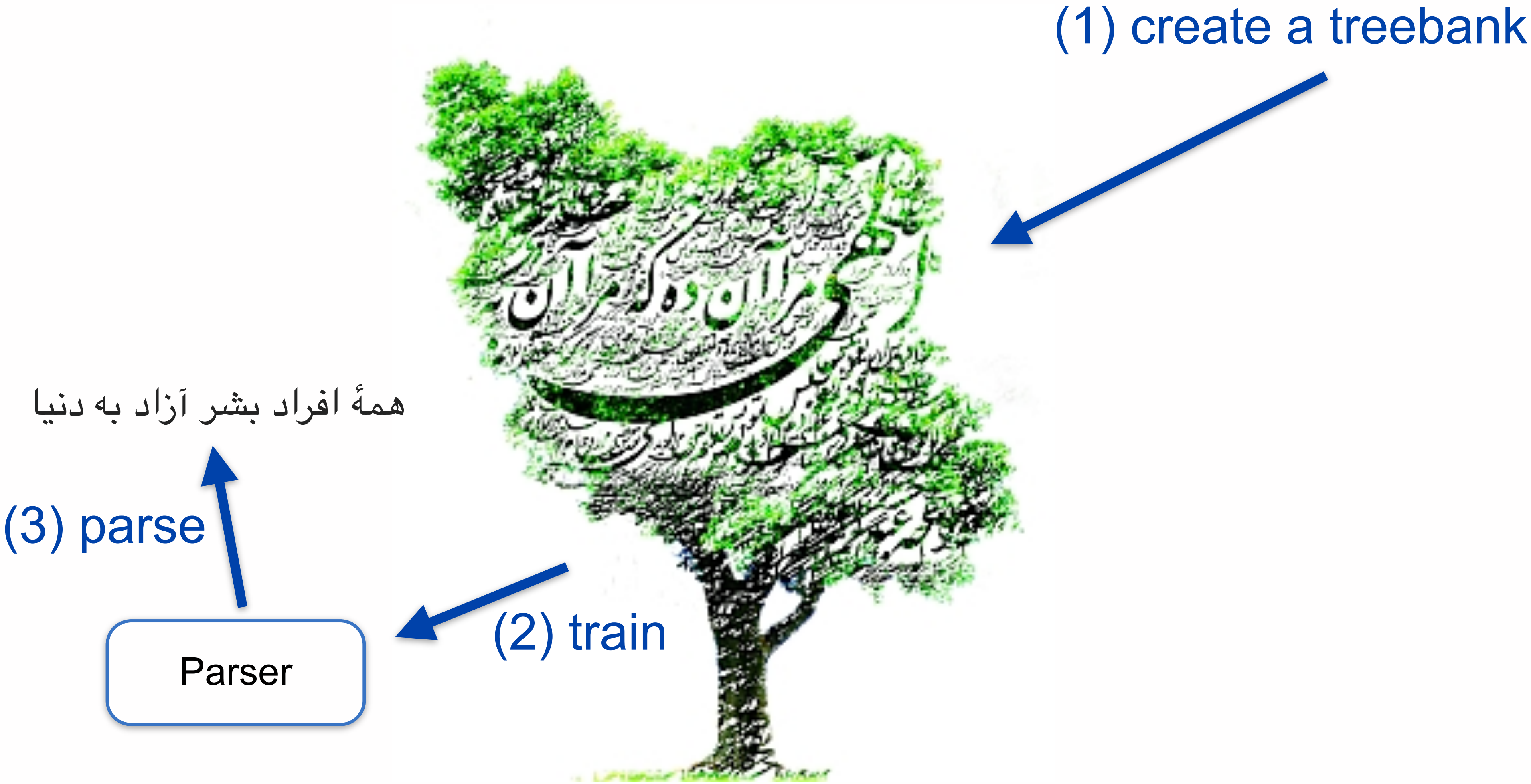
number of speakers (log-scale)

Adjusted wikipedia size plotted against number of speakers, log-log scales

How Important is Language Support?

© Jeff Stahler / Dist. by NEA, Inc.
dey sA I lack ryTN skilz*
* THEY SAY I LACK WRITING SKILLS
STAHLER
www.comics.com
5/8
© 2008 Jeff Stahler/ Dist. by NEA, Inc.

Statistical Parsing



(1) create a treebank

(2) train

Parser

(3) parse

همهٔ افراد بشر آزاد به دنیا

Languages without Treebanks

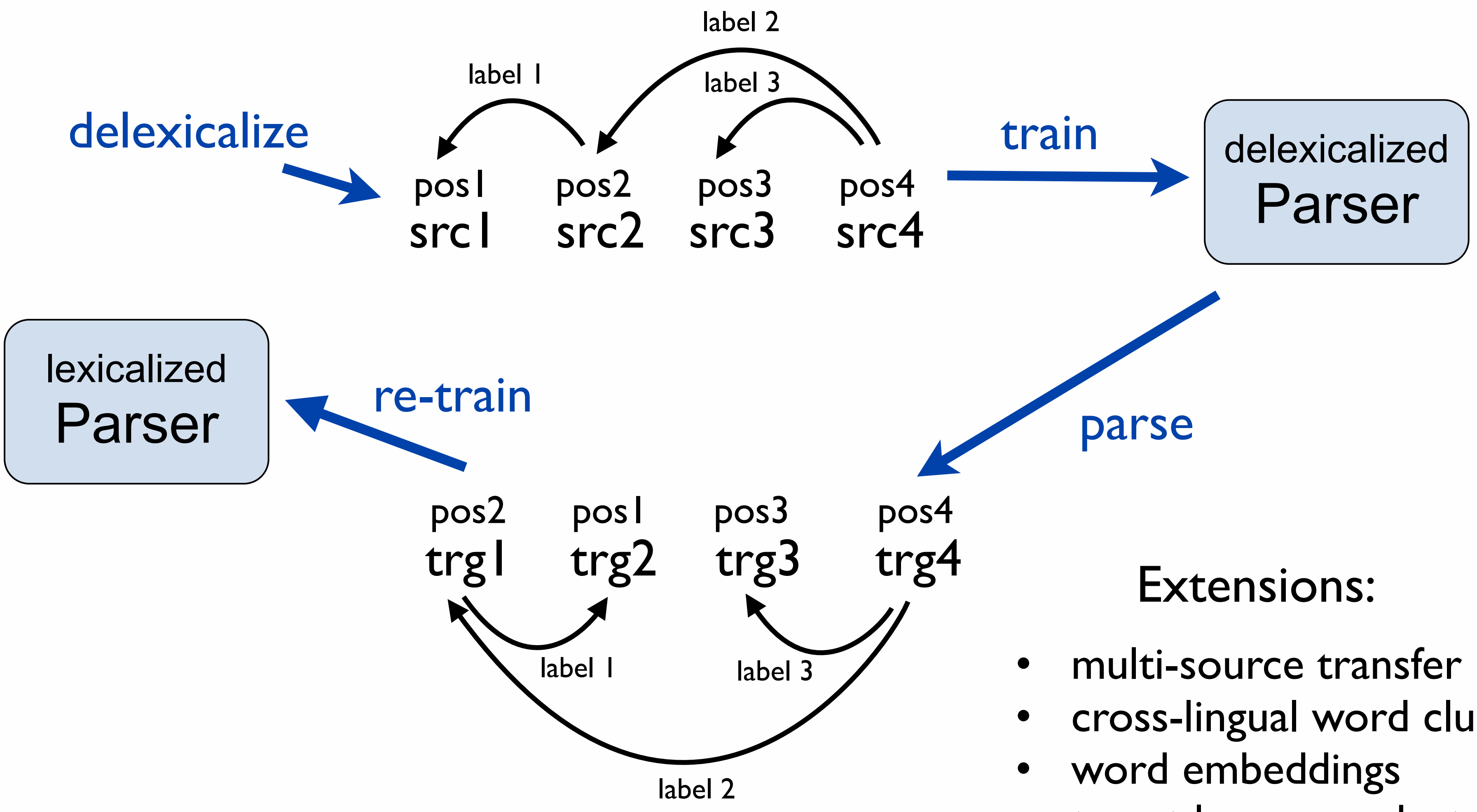
Can we make use of existing resources?

- in related languages
- in resource-rich languages

Cross-Lingual Methods

- **model transfer** (delexicalized models, target adaptation)
- **data transfer** (translations and annotation projection)

Cross-Lingual Methods I



Extensions:

- multi-source transfer
- cross-lingual word clusters
- word embeddings
- target language adaptation

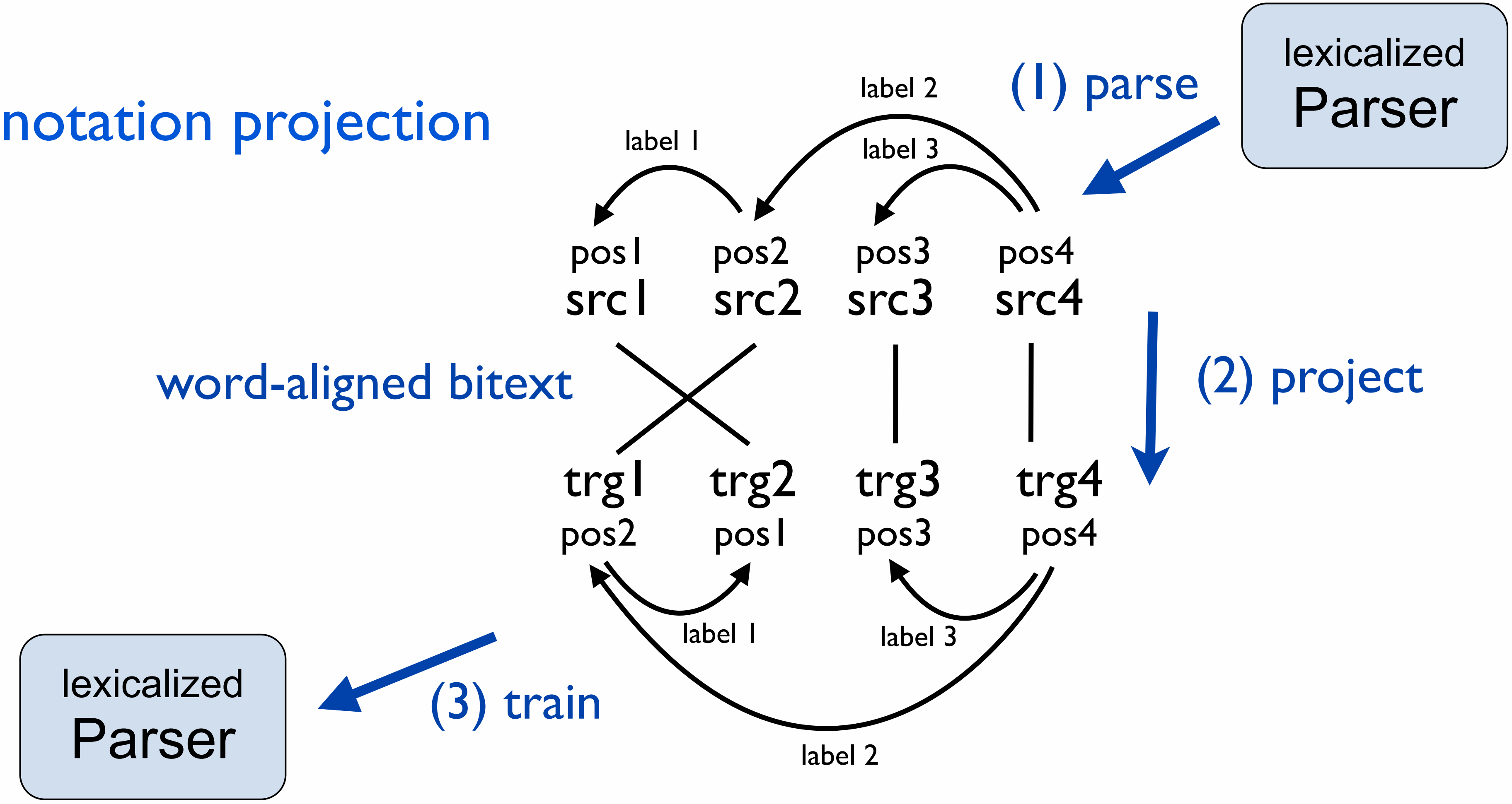
Delexicalized Parsing Across Languages

- `http://universaldependencies.org`
- *LAS = labeled attachment scores*

	← target (test) language →									
LAS	CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
CS		48.90	43.78	43.82	42.18	40.70	30.28	32.18	43.93	40.09
DE	47.27		47.80	53.63	33.45	51.60	37.63	39.41	53.63	46.14
EN	44.27	54.27		60.94	38.52	60.53	39.31	34.06	61.88	50.76
ES	48.40	52.59	50.10		32.80	65.40	43.84	34.46	69.54	46.79
FI	43.75	38.31	40.36	30.14		28.54	20.15	37.39	27.49	37.97
FR	43.63	53.04	52.55	66.42	31.44		41.82	34.53	69.62	44.98
GA	23.23	32.10	28.52	45.61	16.19	43.69		18.24	50.21	27.41
HU	31.83	38.42	29.77	31.17	36.68	30.94	17.59		30.42	25.86
IT	47.38	49.68	47.65	64.96	33.03	64.87	43.42	34.39		45.65
SV	41.20	50.48	47.16	51.93	36.46	51.07	37.76	40.48	55.65	

Cross-Lingual Methods II

Annotation projection



lexicalized Parser

(1) parse

label 2

label 1

label 3

pos1 pos2 pos3 pos4

src1 src2 src3 src4

word-aligned bitext

(2) project

trg1 trg2 trg3 trg4

pos2 pos1 pos3 pos4

label 1

label 3

label 2

(3) train

lexicalized Parser

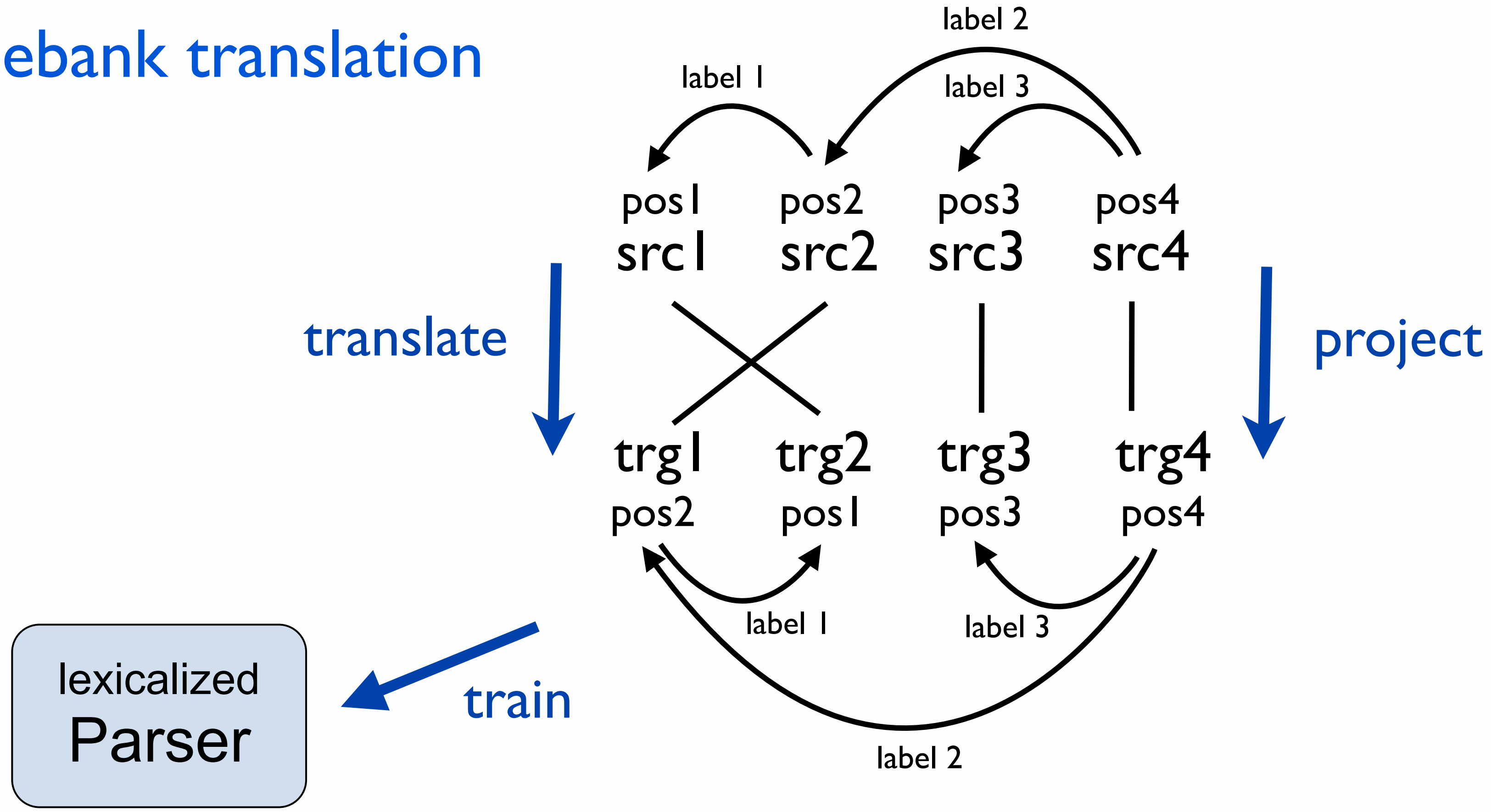
Annotation Projection Results

Example: Spanish as target language

	delexicalized		annotation projection	
PoS	**gold**	**predicted**	**gold**	**predicted**
cs	43,82	33,55	49,17	46,83
de	53,63	46,35	63,49	61,31
en	60,94	52,52	65,07	62,62
es	*75,47*	*69,03*	*84,05*	*80,16*
fi	30,14	26,03	42,37	40,96
fr	66,42	58,74	69,33	66,18
hu	31,17	28,67	48,97	47,36
it	64,96	57,98	65,76	63,31
sv	51,93	37,15	59,06	57,43

Cross-Lingual Methods III

Treebank translation



Treebank Translation Results

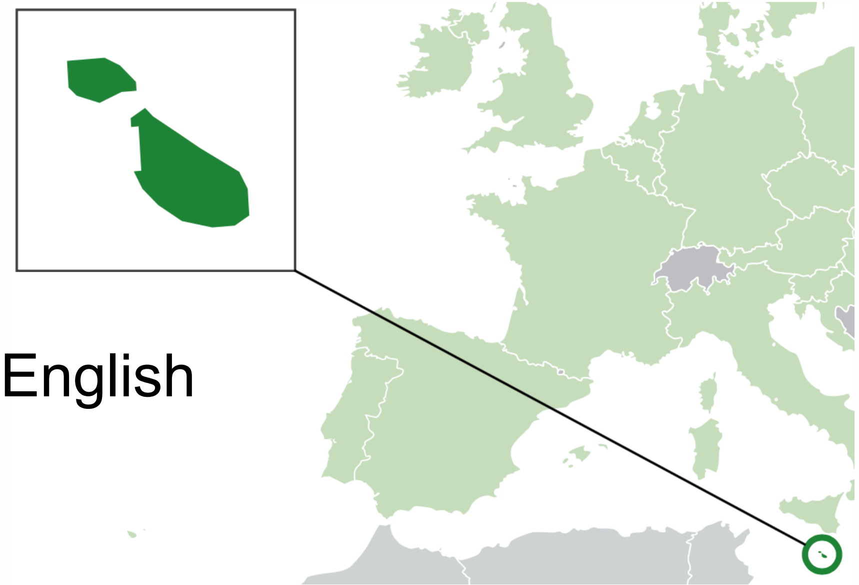
Example: Spanish as target language

	annotation projection		treebank translation	
PoS	**gold**	**predicted**	**gold**	**predicted**
cs	49,17	46,83	49,81	48,07
de	63,49	61,31	64,88	62,34
en	65,07	62,62	67,20	64,48
es	*84,05*	*80,16*	*84,05*	*80,16*
fi	42,37	40,96	36,11	34,45
fr	69,33	66,18	71,15	67,70
hu	48,97	47,36	43,16	41,07
it	65,76	63,31	68,74	66,10
sv	59,06	57,43	59,80	57,41

Does it all make sense?

What’s about real-world scenarios ...

Test-Case One: Maltese



Maltese

- ca 450,000 speakers
- official language of the EU
- influence from Arabic, Italian, English

Resources and tools

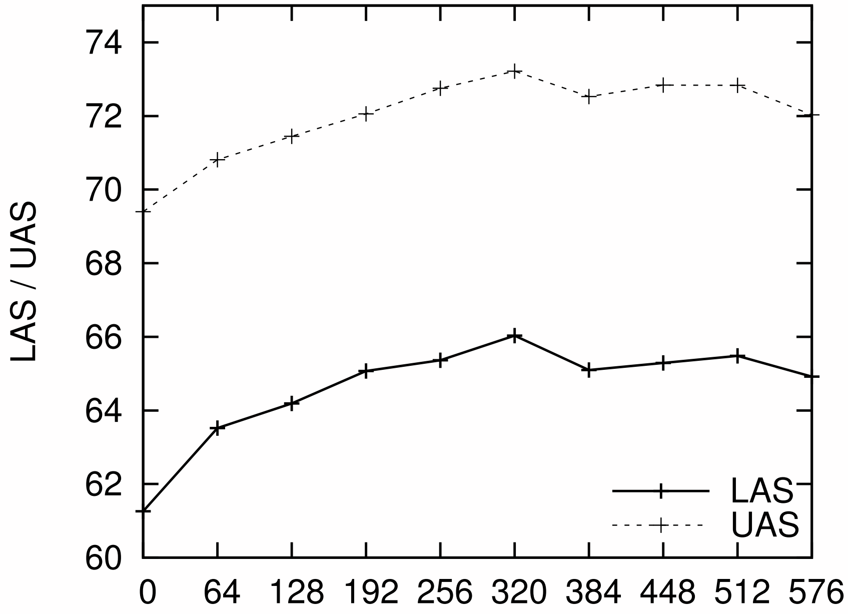
- lexical database with morphological information
- national corpus with automatic PoS annotation (Malti 3.0)
- PoS tagger (ca 97% accuracy)
- UD treebank in development (371 sentences)
- parallel data from the EU!

Cross-Lingual Dependency Parsing

Method	languages	LAS	UAS
Projection	all languages	62.51	71.54
Projection	en es fr it pt ro	62.52	71.28
Projection	bg cs en es it sl	62.77	**71.80**
Projection + inflectional info	bg cs en es it sl	**63.03**	71.54

predicted PoS

Adding projected data to 64 manually annotated trees:



Test-Case Two: Ingush

Nakh-Daghestanian language with ca 300,000 speakers

- no tagger
- no parser
- no parallel data

Russia
Caspian
Sea
Dagestan
Kabardino-
Balkaria
Ingushetia
Chechnya
North
Ossetia
Dagestan
South
Ossetia
Georgia
Tbilisi

Linguistic field work

- transcribed interviews
- interlinear annotation
- English glosses and translations

Ingush:	Cwaqqa hama dwajihwaajaacar, jihwaajarii?
Tokenized:	cwaqqa hama dwajihwaajaacar jihwaajarii
Interlinear glosses:	any thing DX-J.take away.PNW.NEG J.take away.PNW=Q
English:	Nothing had been taken away, right?

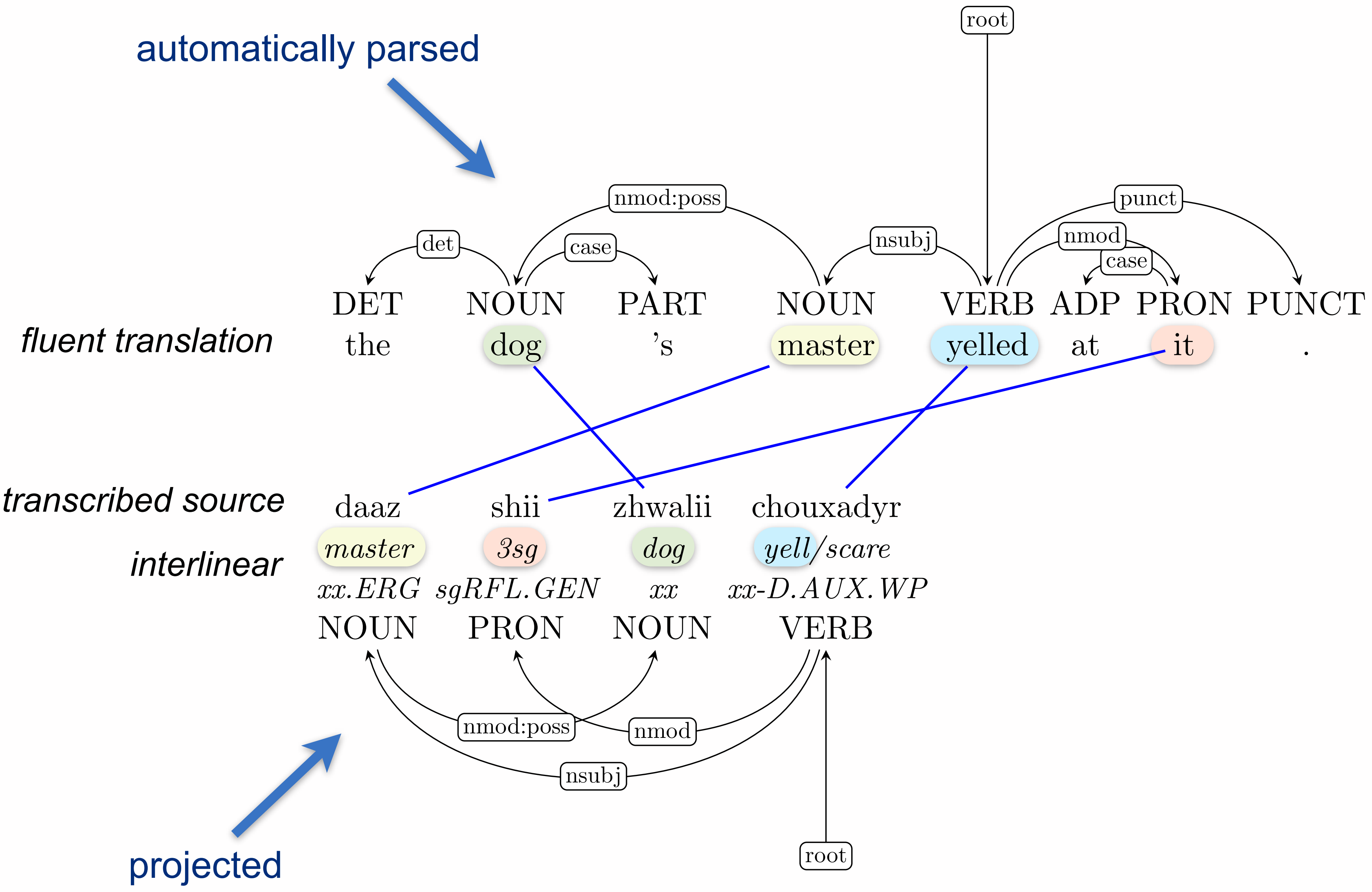
Step 1: Build an Interlinear Tagger

delexicalized

reference	predicted	including xx P	R	without xx P	R	token
xx.NW.D.NEG	xx.NW.D.NEG	100	100	100	100	xeattaadaac
DEM.PL.OBL	DEM.OBL	100	67	100	67	cy
xx.PL.DAT	xx.PL.DAT	100	100	100	100	bierazhta
D.PST=PTC	D.xx.PST=CUM	50	67	67	67	dar=q
DX-xx-J.xx.NW.J.NEG	DX-xx.AUX.NEG.PRS	25	20	25	25.00	dwachyjeannajaac
D.PST=PTC	D.xx.PST=CUM	50	67	67	67	dar=q
xx:NEG.PRS	xx.PRS.NEG	33	50	50	50	xaac
xx-J.xx.CVtemp	xx-D.xx.CVtemp	67	67	50	50	chyjiecha
J.xx.NEG.WP	J.AUX.NEG.WP	75	75	75	100	jaxandzar

(scores in %)	unambiguous	ambiguous (train)	ambiguous (test+train)	unknown
precision	95.06	83.64	49.19	72.13
recall	95.44	83.50	49.72	66.27
accuracy	90.38	70.74	4.24	34.39

Step 2: Gloss Alignment and Transfer


automatically parsed
root
nmod:poss
det
case
nsubj
punct
nmod
case
DET NOUN PART NOUN VERB ADP PRON PUNCT
fluent translation
the dog 's master yelled at it .
transcribed source
daaz shii zhwalii chouxadyr
interlinear
master *3sg* *dog* *yell/scare*
xx.ERG *sgRFL.GEN* *xx* *xx-D.AUX.WP*
NOUN PRON NOUN VERB
nmod:poss
nmod
nsubj
root

projected

And the Results are …

?

Conclusions

Cross-lingual parsing

- transfer / multilingual models are weak
- annotation projection is more robust
- treebank translation is possible

Tools for low-resource languages

- bootstrap data via annotation projection
- creative use of linguistic field work

Useful in applications and research?

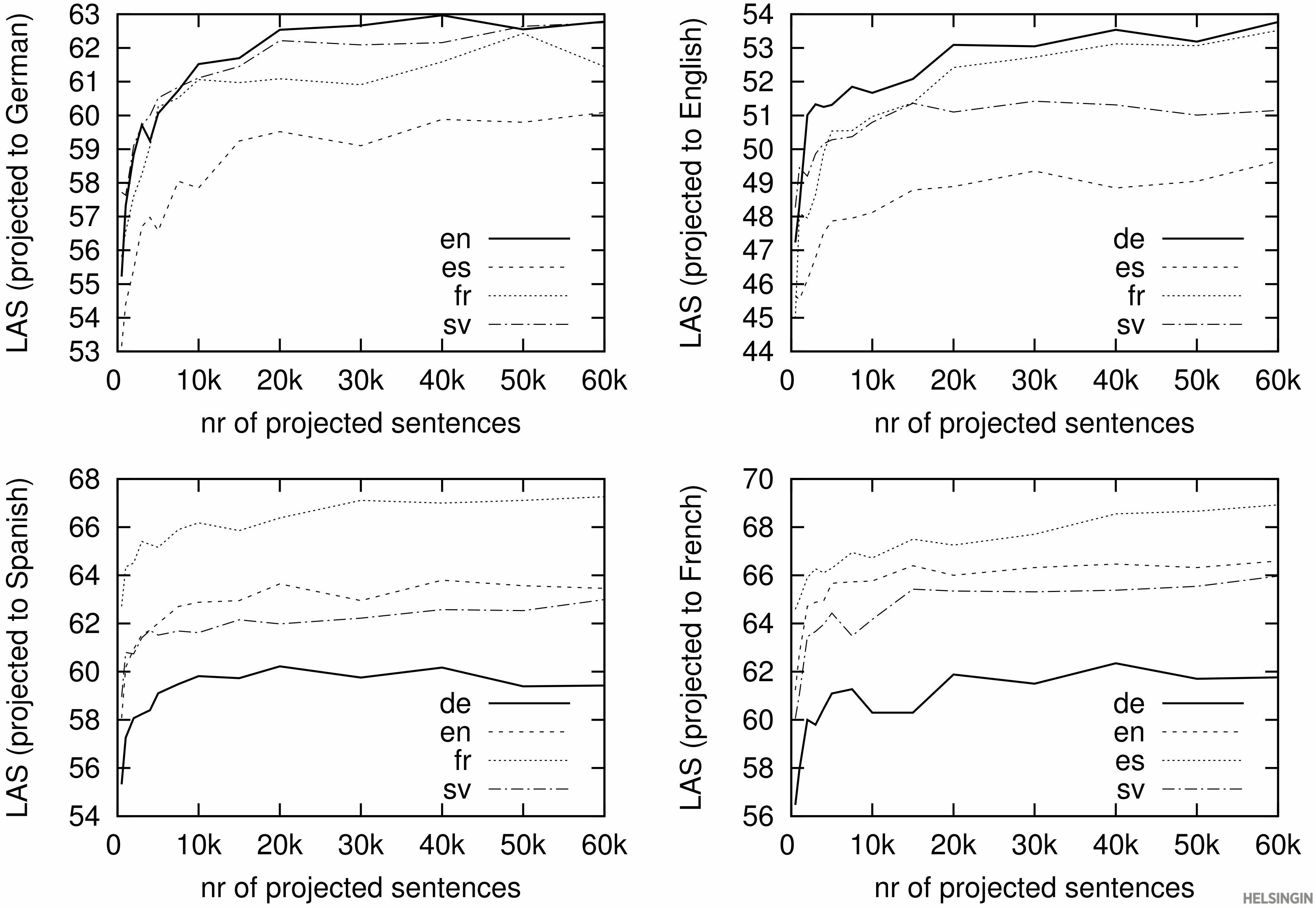
Questions?

Multi-Source System Combinations

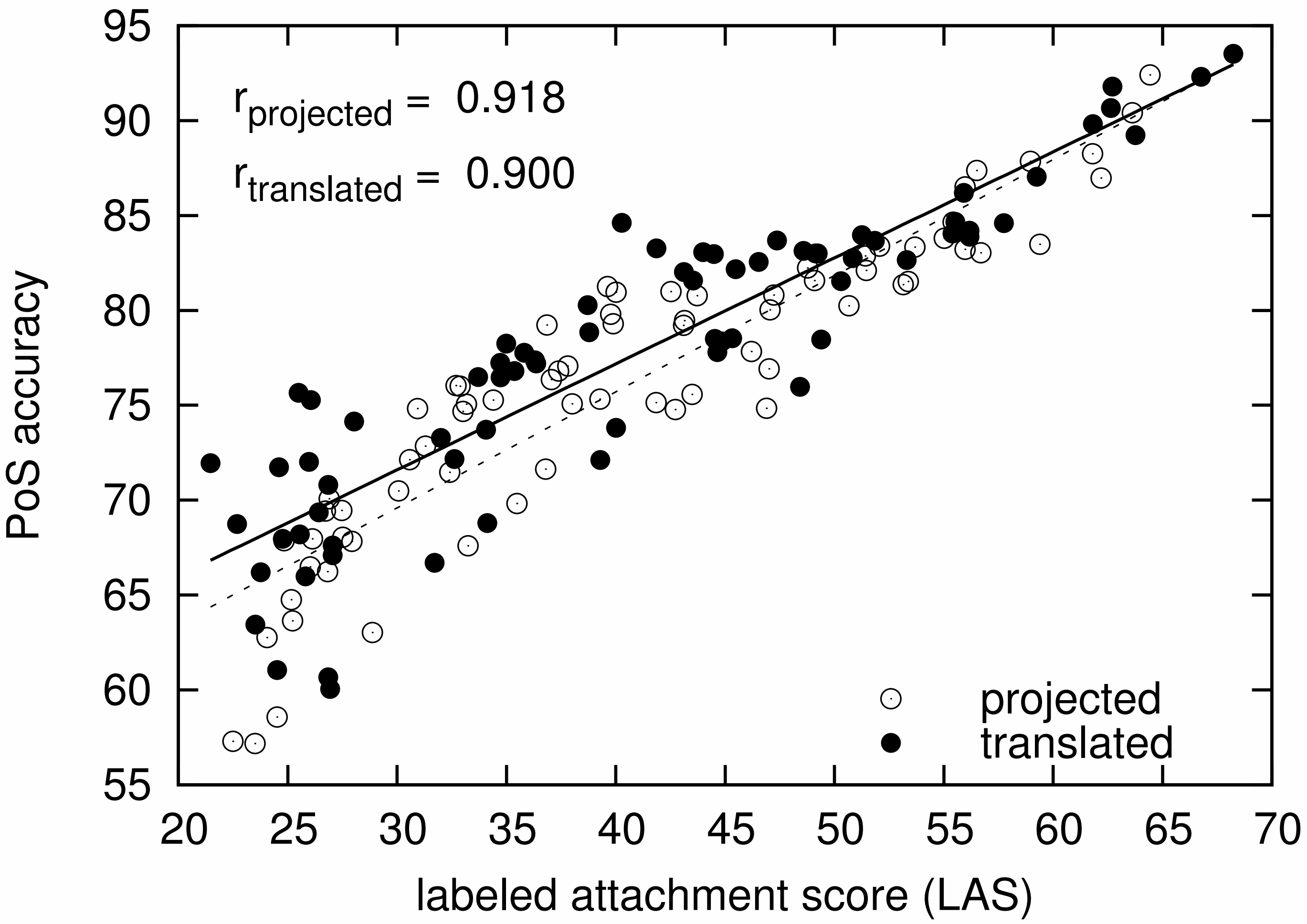
	DE	EN	ES	FR	SV
monolingual baseline with gold PoS	78.38	91.46	82.30	82.30	84.52
delexicalized monolingual with gold PoS	70.84	82.44	71.45	73.71	74.55
best delexicalized cross-lingual with gold PoS	52.53	48.24	62.66	62.39	59.42
best cross-lingual model with gold PoS	**67.60**	**61.56**	**69.36**	**72.78**	**73.40**
monolingual PoS tagger accuracy	95.24	97.56	95.37	95.08	95.86
combined projected PoS tagger accuracy	88.47	88.24	88.06	89.83	88.07
monolingual baseline with predicted PoS	73.03	88.38	76.59	76.79	77.83
delexicalized monolingual with predicted PoS	64.25	72.81	60.49	64.06	65.77
best delexicalized cross-lingual with predicted PoS	48.36	43.87	52.94	52.47	49.84
combined cross-lingual with predicted PoS	**63.14**	**55.16**	**64.99**	**67.91**	**67.93**
combined cross-lingual with projected PoS model	**57.84**	**51.66**	**61.40**	**63.86**	**61.58**

(labeled attachment scores)

How Much Data Do We Need?

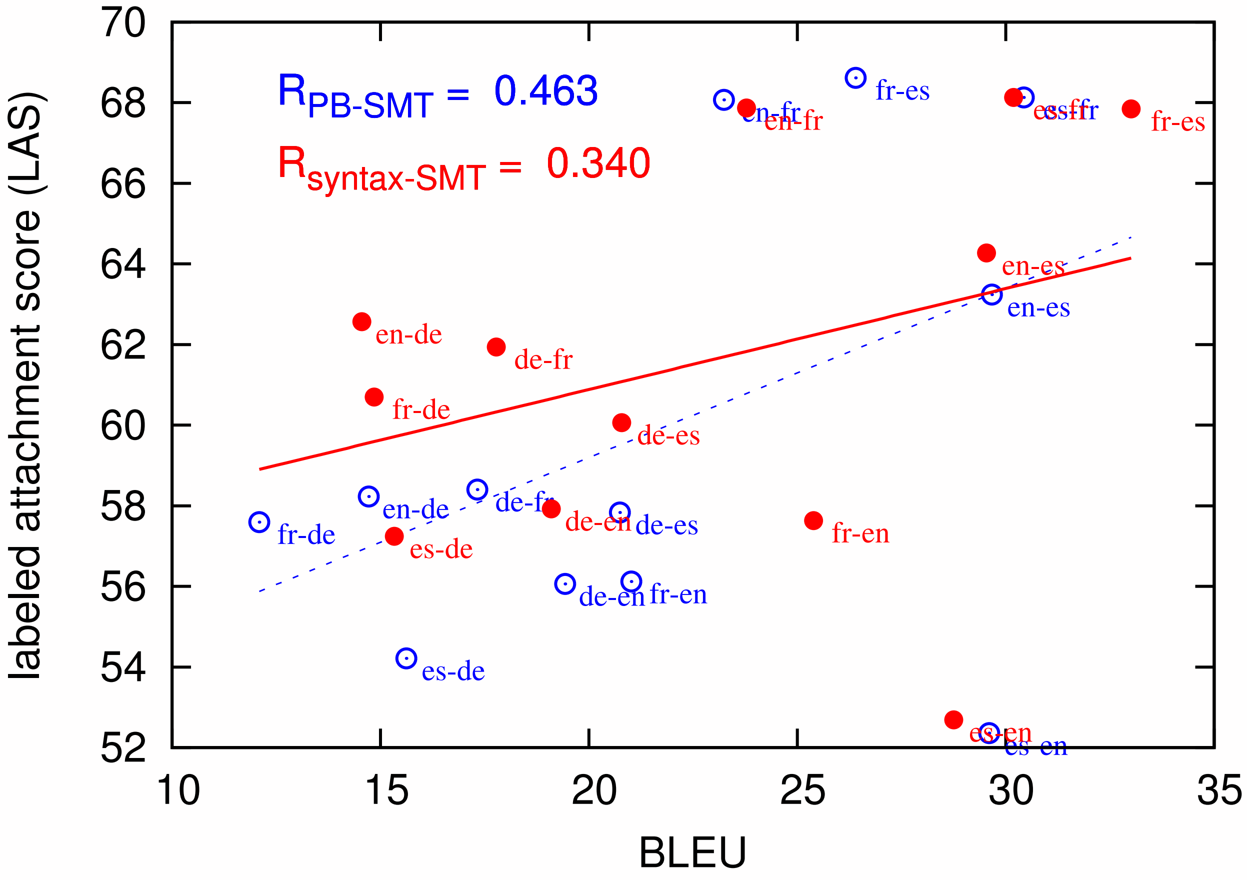


The Impact of PoS Tagging Performance



Translation Quality vs. Parsing Quality

Treebank translation approach:


$R_{PB\text{-}SMT} = 0.463$
$R_{syntax\text{-}SMT} = 0.340$
labeled attachment score (LAS)
BLEU
fr-es
en-fr
en-fr
es-fr
es-fr
fr-es
en-es
en-es
en-de
de-fr
fr-de
de-es
en-de
de-fr
de-en
de-es
fr-de
es-de
fr-en
de-en
fr-en
es-de
es-en
es-en