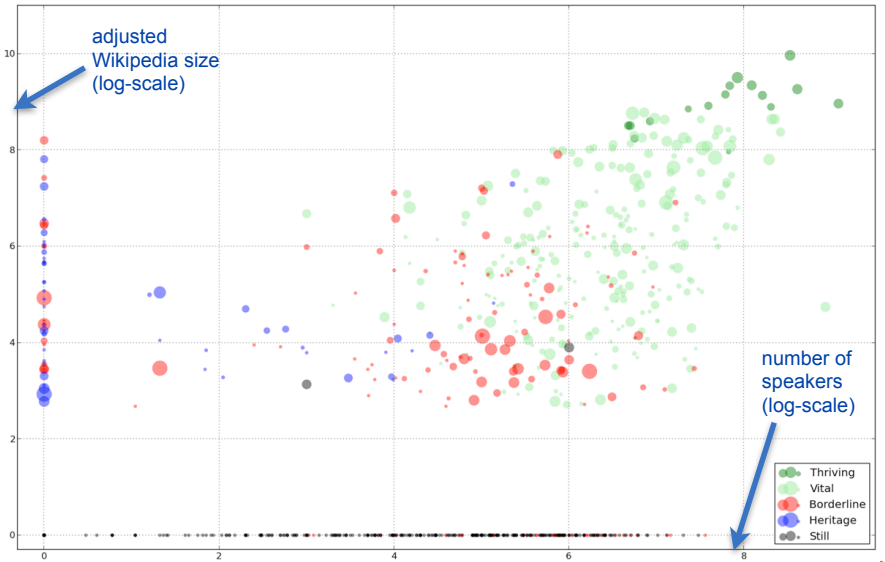
Languages are Dialects with a Treebank and a Dependency Parser

Cross-Lingual Parsing for Low-Resource Languages

Jörg Tiedemann
Department of Modern Languages
University of Helsinki

Digital Language Death (Kornai, 2013)



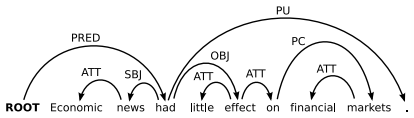
Adjusted wikipedia size plotted against number of speakers, log-log scales

How Important is Language Support?

Multilingual Dependency Parsing

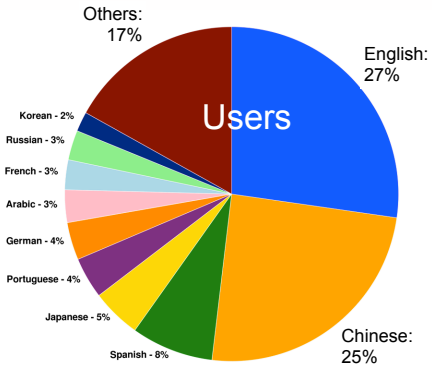
Parsing technology in many applications

- machine translation
- information extraction
- ...

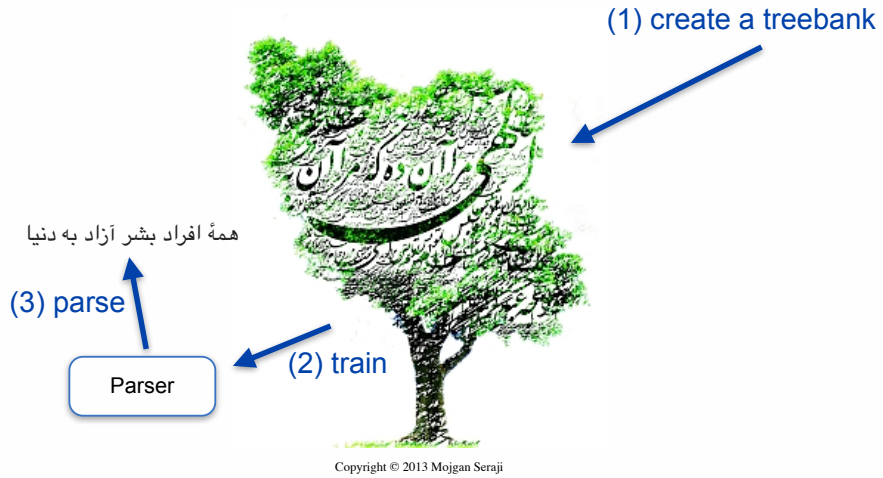


The World is not English only

- many languages on the Web
- most are under-resourced



Statistical Parsing



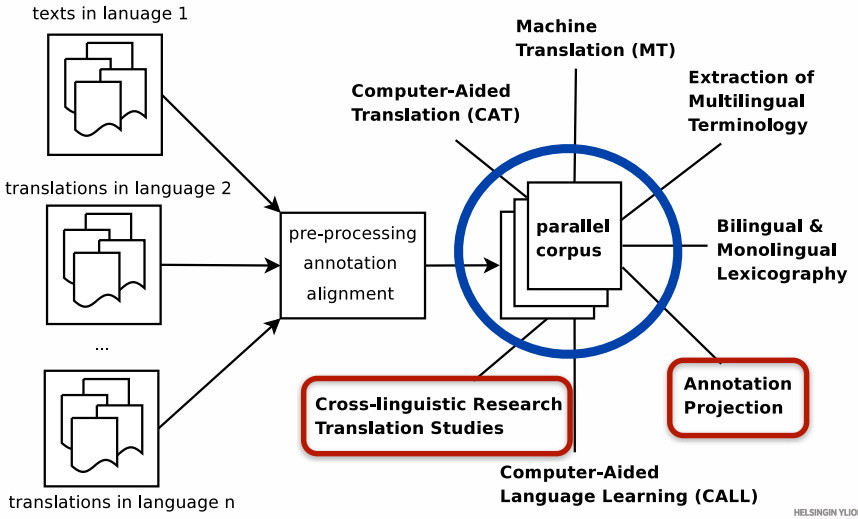
Languages without Treebanks

Unsupervised learning
- not yet practically useful

Hand-Written Rule-Based Systems
- requires experts and is time-consuming
- issues with robustness

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

The Amazing Utility of Parallel Corpora



Linguistic Solidarity ...

Resource-rich languages support resource-poor languages!

Parsing with Universal Dependencies

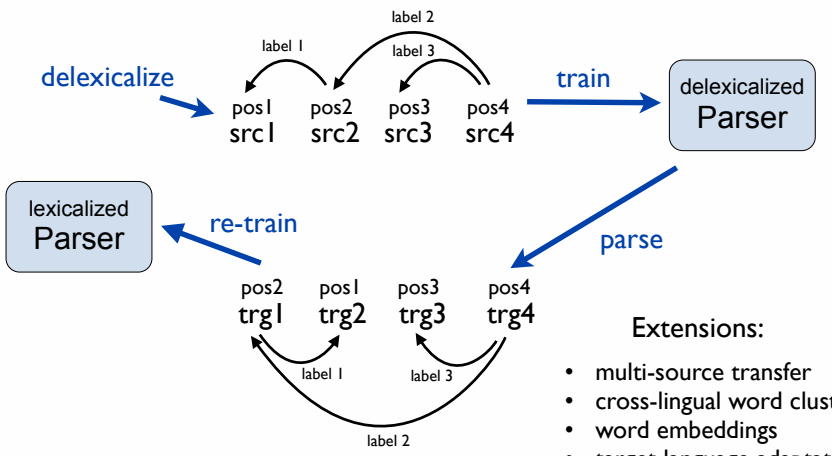
Cross-lingually harmonized annotation

- *http://universaldependencies.org*

baseline models

language	size	lemma	morph.	LAS	UAS	LACC
CS	60k	X	X	85.74	90.04	91.99
DE	14k			79.39	84.38	90.28
EN	13k		(X)	85.70	87.76	93.29
ES	14k			84.05	86.77	92.90
FI	12k	X	X	84.51	86.51	93.53
FR	15k			81.03	84.39	91.02
GA	0.7k	X		72.73	78.75	84.74
HU	1k	X	X	83.19	85.28	92.73
IT	9k	X	X	89.58	91.86	95.92
SV	4k		X	82.66	85.66	91.06

Cross-Lingual Methods I



Extensions:

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

Delexicalized Parsing Across Languages

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	

Gold vs. Predicted PoS/Morphology

Monolingual parsing:

LAS	CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
gold PoS & morphology	85.74	—	85.70	—	84.51	—	72.73	83.19	89.58	82.66
gold coarse PoS	80.75	79.39	84.81	84.05	74.62	81.03	71.39	73.39	88.25	81.02
delexicalized & gold PoS	70.36	71.29	76.04	75.47	59.54	74.19	66.97	66.57	79.07	66.95
predicted PoS & morphology	82.67	—	81.36	—	80.59	—	66.74	75.78	87.16	78.76
predicted coarse PoS	79.41	74.39	80.33	80.16	70.25	78.73	65.93	68.04	85.08	76.42
delexicalized & predicted PoS	62.44	61.82	67.40	69.03	49.79	68.60	55.33	58.90	72.92	61.99
coarse PoS tagger (accuracy)	98.28	93.19	94.89	95.13	95.69	95.99	91.97	94.69	97.63	96.79
morph. tagger (accuracy)	93.47	—	94.80	—	94.53	—	91.92	91.06	97.50	95.26

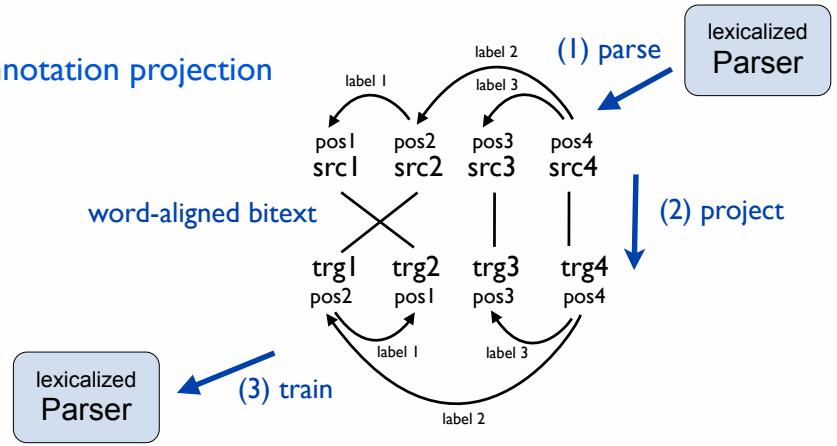
Delexicalized Parsing Across Languages

... with predicted PoS labels:

Δ LAS	CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
CS		-9.30	-7.73	-10.27	-7.17	-8.53	-8.85	-4.36	-10.59	-4.05
DE	-6.69		-6.22	-7.28	-6.62	-5.18	-7.77	-8.22	-5.26	-5.09
EN	-3.94	-5.93		-8.42	-5.37	-6.27	-6.99	-2.87	-7.96	-4.87
ES	-3.99	-7.05	-5.46		-4.58	-5.59	-7.28	-4.63	-4.86	-2.31
FI	-2.47	-7.72	-3.94	-3.80		-1.70	-5.39	-5.68	-1.59	-2.28
FR	-4.24	-7.62	-5.24	-7.68	-4.95		-9.50	-4.73	-7.61	-3.51
GA	-2.15	-2.38	-1.42	-6.91	-2.25	-3.57		-3.12	-7.13	-3.01
HU	-2.81	-5.29	-3.14	-2.50	-5.63	-1.64	-2.41		-2.05	-1.62
IT	-8.81	-7.15	-6.19	-6.98	-5.33	-5.84	-8.61	-8.08		-3.98
SV	-2.64	-10.18	-6.13	-14.78	-3.12	-13.11	-10.83	-6.68	-14.09	

Cross-Lingual Methods II

Annotation projection



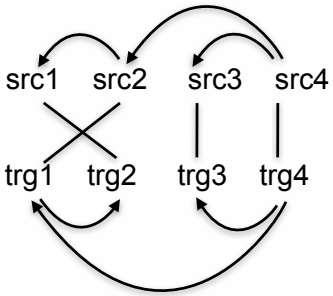
Direct Annotation Projection

One-to-one mapping of dependency relations

source language

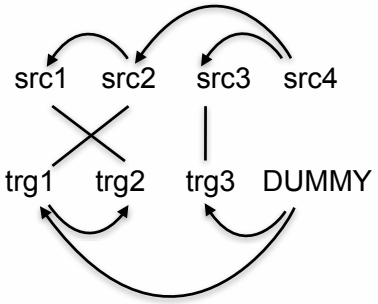
word alignment

target language



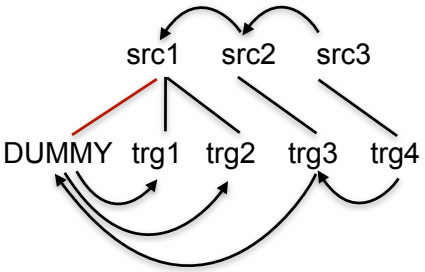
Direct Annotation Projection

Unaligned source words:



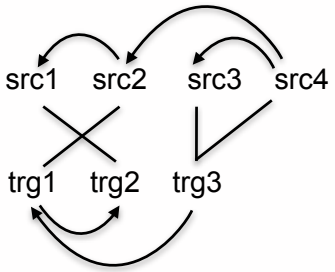
Direct Annotation Projection

One-to-many alignments:



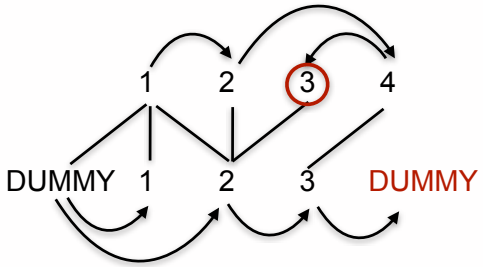
Direct Annotation Projection

Many-to-one alignments:



Direct Annotation Projection

Many-to-many alignments:



Annotation Projection Results

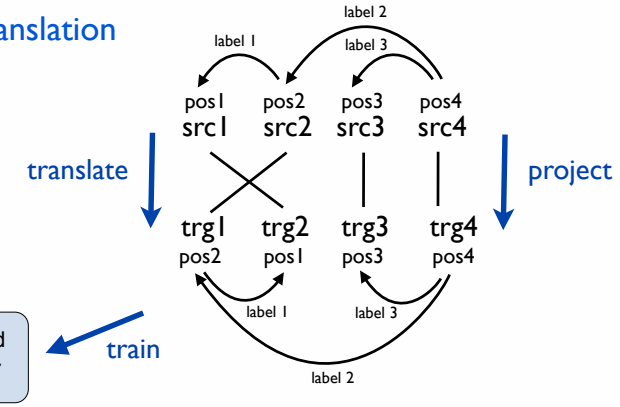
Example: Spanish as target language

tagger trained on projected data

	delexicalized		annotation projection		
PoS	**gold**	**predicted**	**gold**	**predicted**	**projected**
cs	43,82	33,55	49,17	46,83	36,85
de	53,63	46,35	63,49	61,31	53,15
en	60,94	52,52	65,07	62,62	56,69
es	*75,47*	*69,03*	*84,05*	*80,16*	*80,16*
fi	30,14	26,03	42,37	40,96	23,50
fr	66,42	58,74	69,33	66,18	61,81
hu	31,17	28,67	48,97	47,36	26,82
it	64,96	57,98	65,76	63,31	55,98
sv	51,93	37,15	59,06	57,43	52,06

Cross-Lingual Methods III

Treebank translation



Treebank Translation Results

Example: Spanish as target language

	annotation projection			treebank translation		
PoS	**gold**	**predicted**	**projected**	**gold**	**predicted**	**projected**
cs	49,17	46,83	36,85	49,81	48,07	40,02
de	63,49	61,31	53,15	64,88	62,34	53,30
en	65,07	62,62	56,69	67,20	64,48	56,18
es	*84,05*	*80,16*	*80,16*	*84,05*	*80,16*	*80,16*
fi	42,37	40,96	23,50	36,11	34,45	26,86
fr	69,33	66,18	61,81	71,15	67,70	63,77
hu	48,97	47,36	26,82	43,16	41,07	25,81
it	65,76	63,31	55,98	68,74	66,10	61,82
sv	59,06	57,43	52,06	59,80	57,41	51,26

Open Questions

How much data do we need?

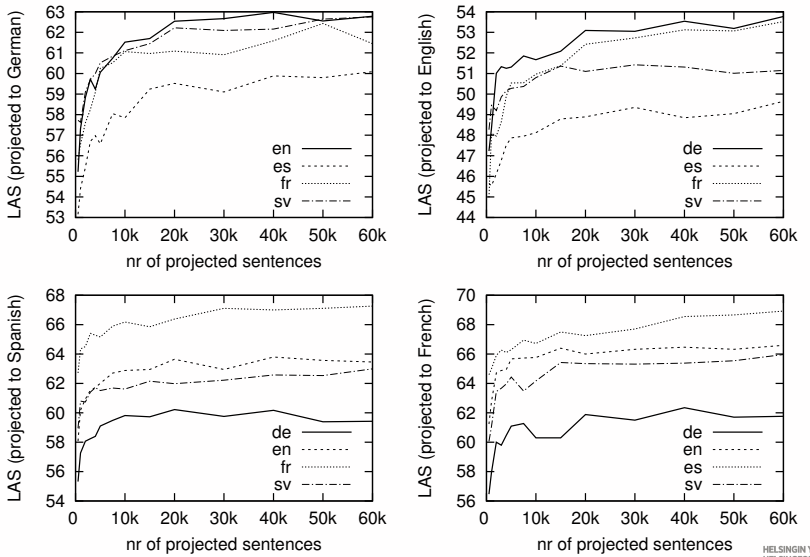
How much does it depend on ...

- PoS tagging accuracy?
- Word alignment / translation quality?
- Relatedness of source and target language?

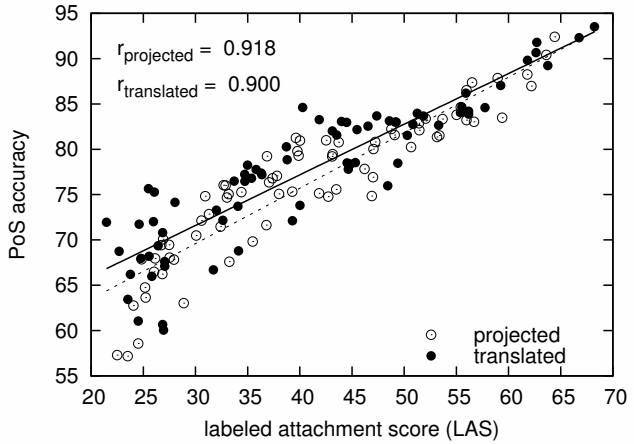
Does it make sense for truly under-resourced languages?

- parallel data available?
- MT available?

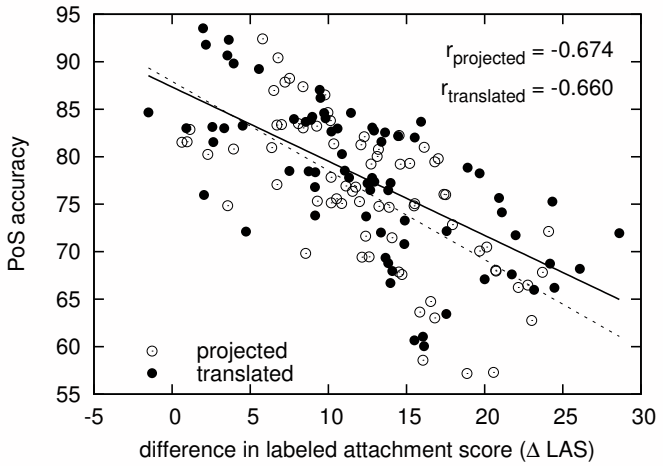
How Much Data Do We Need?



The Impact of PoS Tagging Performance

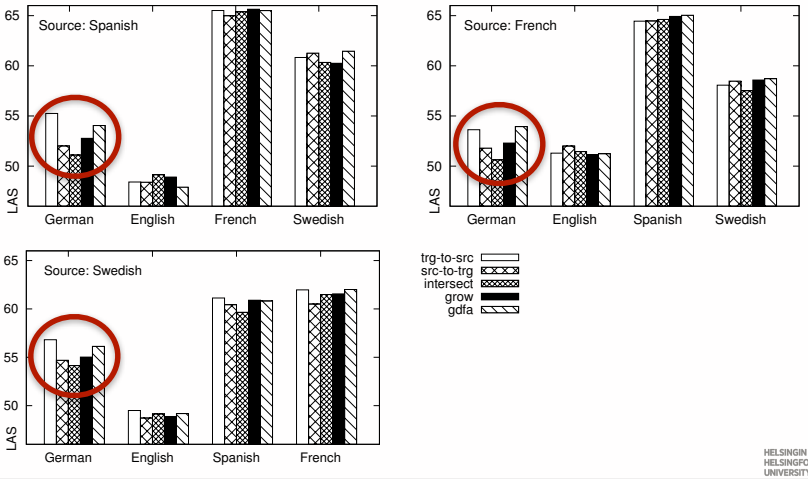


Performance Drop When Predicting PoS



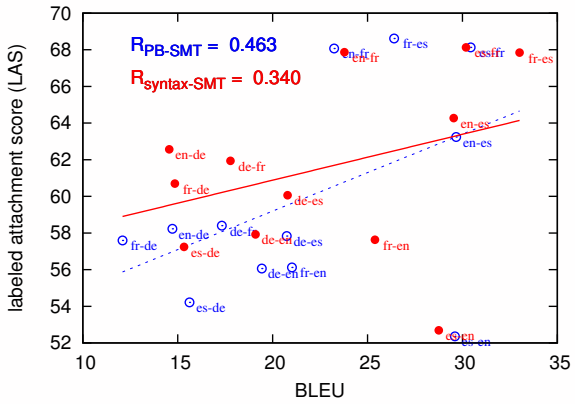
The Impact of Word Alignment

Different symmetrization algorithms:



Translation Quality vs. Parsing Quality

Treebank translation approach:



Multi-Source System Combinations

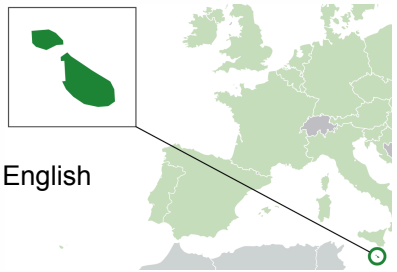
	DE	EN	ES	FR	SV
monolingual baseline with gold PoS	78.38	91.46	82.30	82.30	84.52
delexicalized monolingual with gold PoS	70.84	82.44	71.45	73.71	74.55
best delexicalized cross-lingual with gold PoS	52.53	48.24	62.66	62.39	59.42
best cross-lingual model with gold PoS	**67.60**	**61.56**	**69.36**	**72.78**	**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)

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

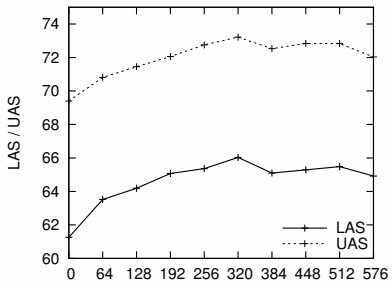
Resources and tools

- 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

Adding projected data to 64 manually annotated trees:



predicted PoS

Test-Case Two: Ingush

Nakh-Daghestanian language with ca 300,000 speakers

- no tagger
- no parser
- no parallel data

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?

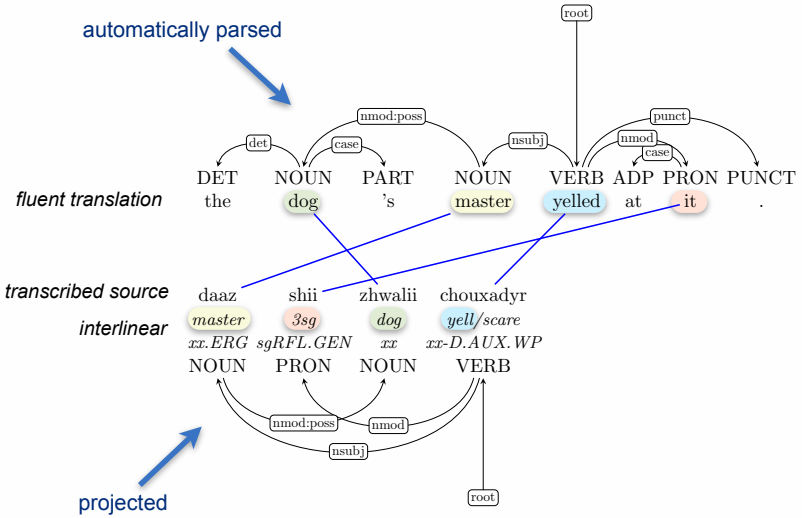
Step 1: Build an Interlinear Tagger

delexicalized

reference	predicted	including xx P	including xx R	without xx P	without xx 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

		ambiguous		
(scores in %)	unambiguous	(train)	(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



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?