Centre for
Data Analytics

Insight

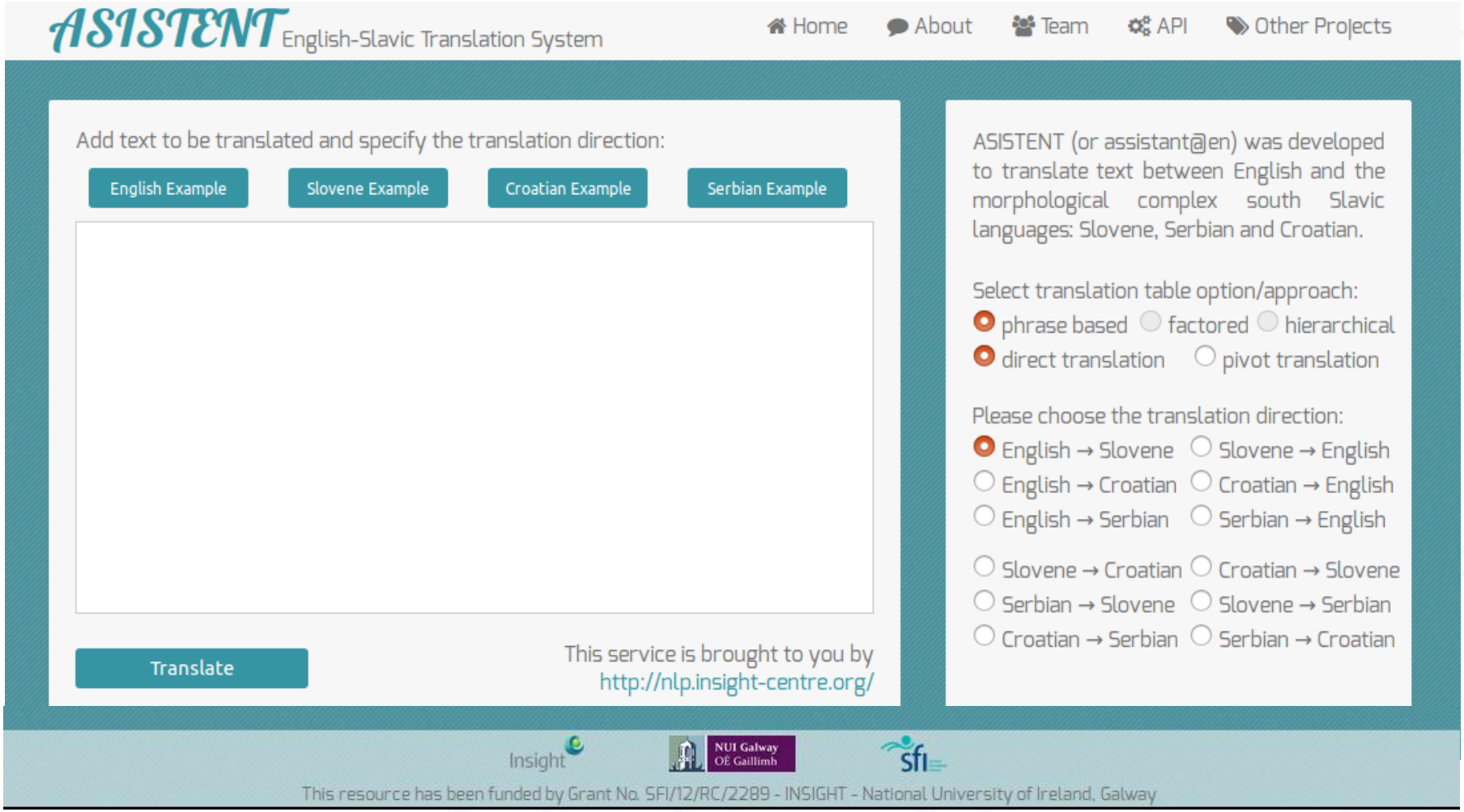
Asistent
A Machine Translation System for Slovene, Serbian and Croatian

Mihael Arcan, Maja Popovic, Paul Buitelaar

Insight Center for Data Analytics, NUI Galway, Ireland
Humboldt University of Berlin, Germany

DCU NUI Galway OÉ Gaillimh UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh UCD DUBLIN sfi

Asistent



http://server1.nlp.insight-centre.org/asistent/

Motivation

Insight

META-NET Language White Paper Series (2012)

Machine Translation

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	▪ English	▪ French ▪ Spanish	▪ Catalan ▪ Dutch ▪ German ▪ Hungarian ▪ Italian ▪ Polish ▪ Romanian	▪ Basque ▪ Bulgarian ▪ Croatian ▪ Czech ▪ Danish ▪ Estonian ▪ Finnish ▪ Galician ▪ Greek ▪ Icelandic ▪ Irish ▪ Latvian ▪ Lithuanian ▪ Maltese ▪ Norwegian (Bokmål, Nynorsk) ▪ Portuguese ▪ Serbian ▪ Slovak ▪ Slovene ▪ Swedish ▪ Welsh

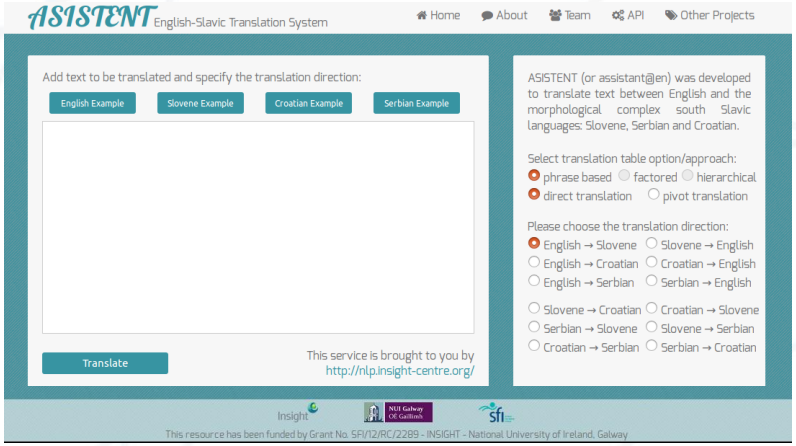
What Do We Need?

Translation Toolkit:

- Moses (text tokenisation, lower-casing)
- Giza++ (word alignment)
- KenLM (language model)

Parallel data

- Europarl
- DGT
- OpenSubTitles
- ....



http://server1.nlp.insight-centre.org/asistent/

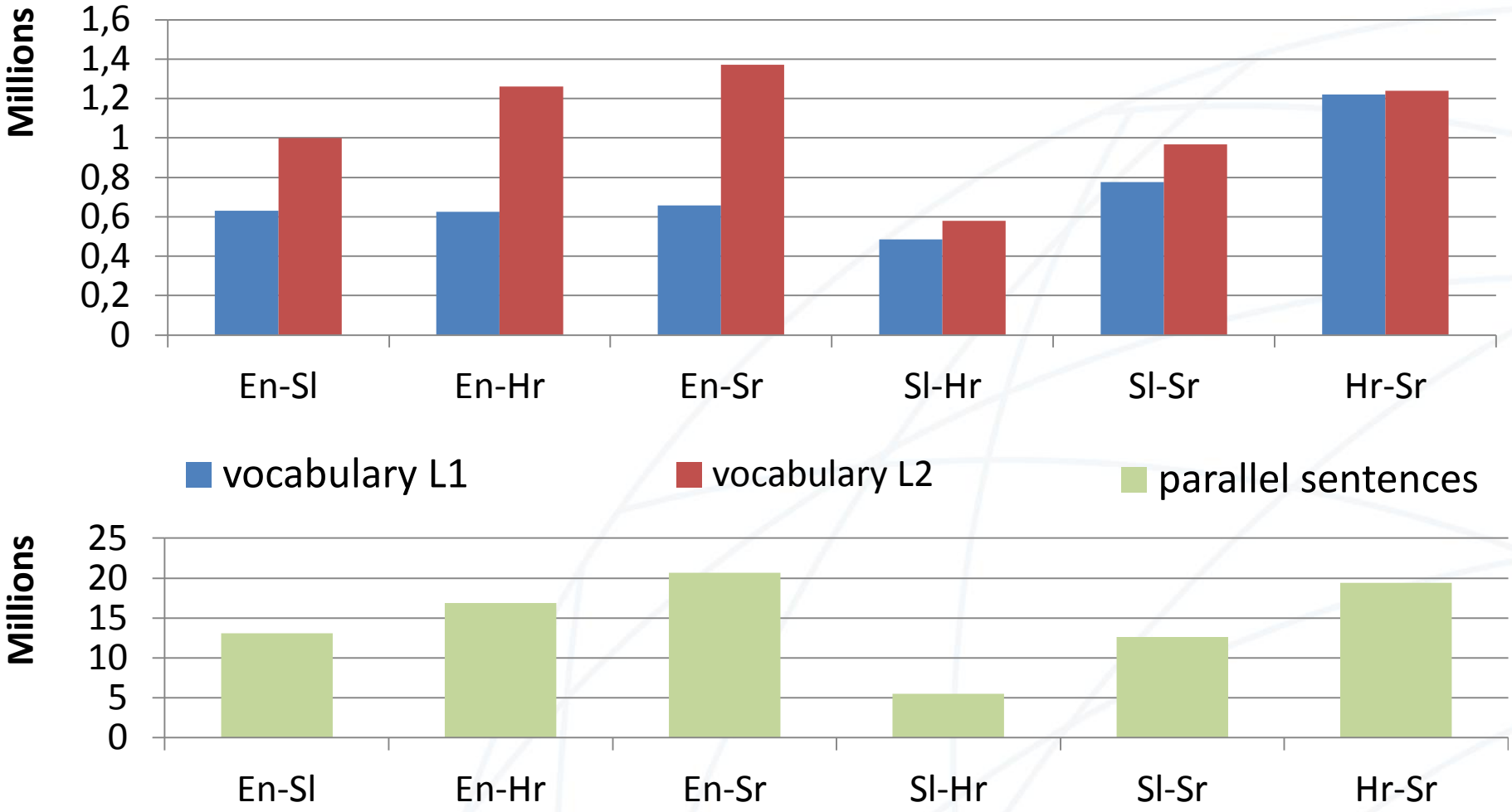
Parallel Corpora Used for Asistent

Insight

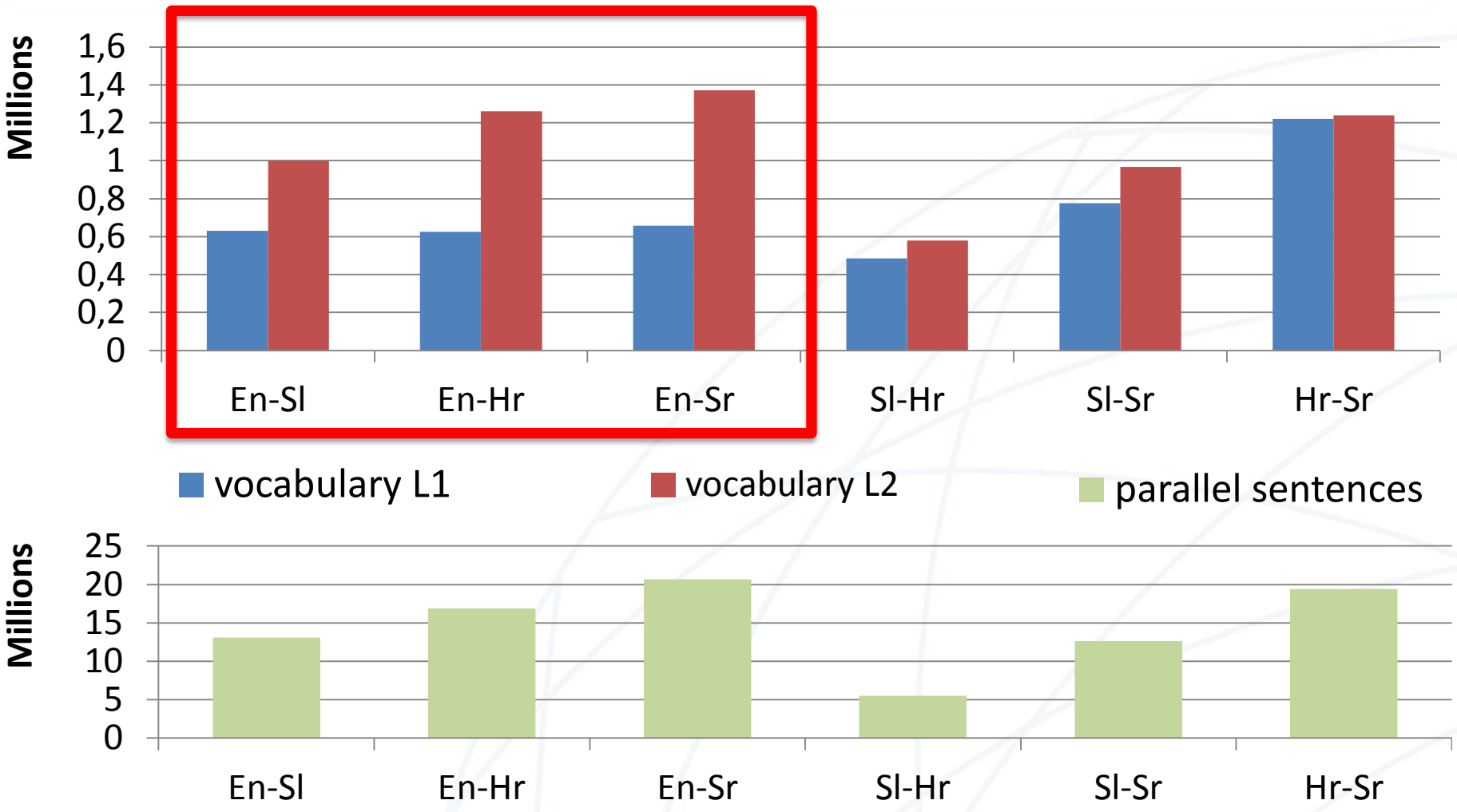
Name	Domain	En-Sl	En-Hr	En-Sr	Sl-Hr	Sl-Sr	Hr-Sr
DGT	legal	1.8M	196K				
ECB	finance	79K					
EMEA	medical	253K					
Europarl	EP proceedings	599K					
Gnome	IT	998	5K	126	4K	600K	300K
hrenWaC	Web Corpus		86K				
JRC-Acquis	legal	29K	38K				
KDE	IT	58K		32K	85K	49k	33.2k
LangCourse	education			3K			
PHP	IT	1K					
OpenSubtitles	subtitles	10.1M	16.3M	20.5M	6.1M	13.3M	22.3M
SETimes	Web Corpus		198K	209K			
Tatoeba	education		777	633			200K
TED	education	13K	76K	1K			
Ubuntu	IT		8K		557	86K	51K

*http://opus.lingfil.uu.se/index.php

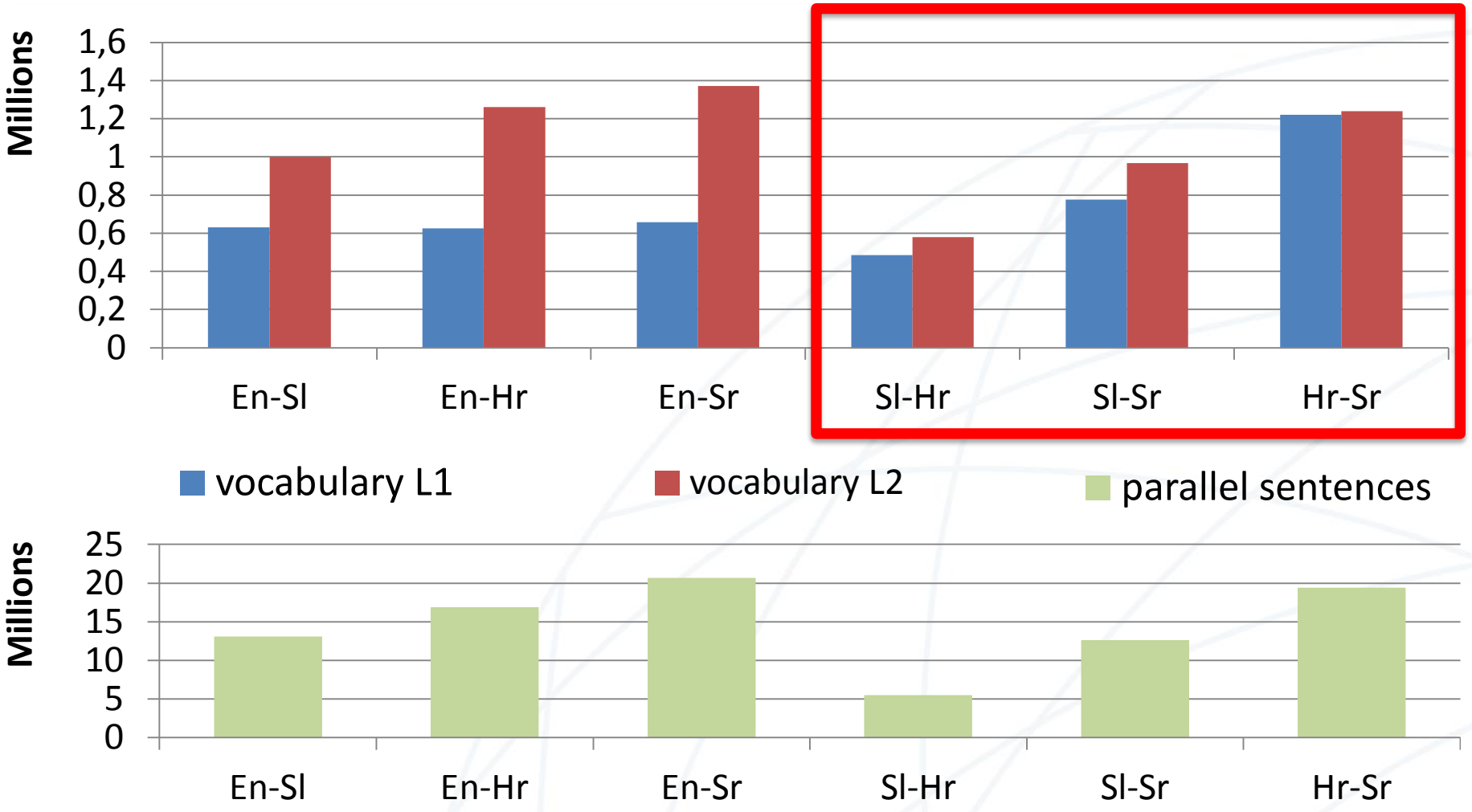
Concatenated Data for Asistent Training



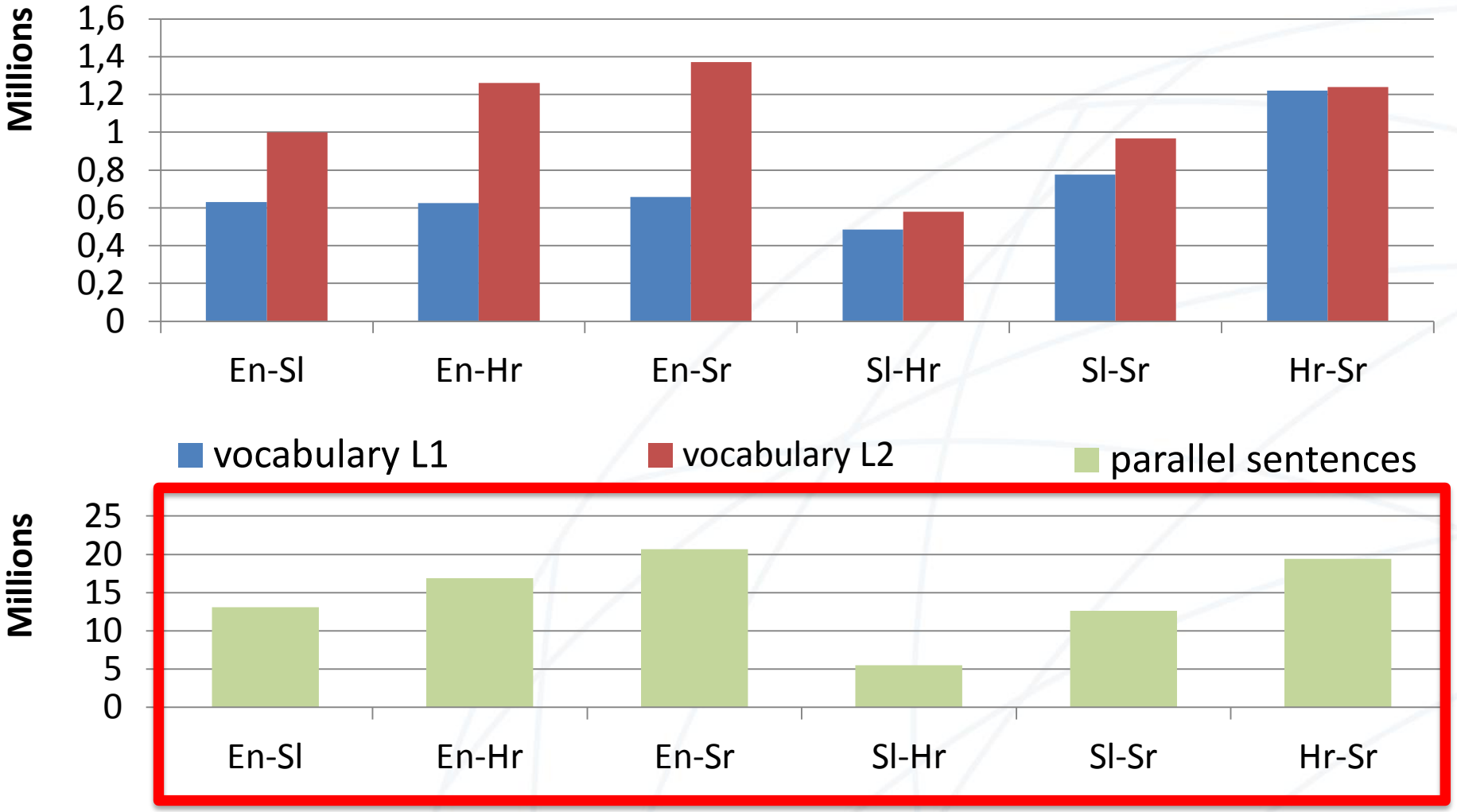
Concatenated Data for Asistent Training



Concatenated Data for Asistent Training



Concatenated Data for Asistent Training



Data Preparation

Insight

Data cleanup

- Falsely encoded special characters removed/fixed
- Cyrillic → Latin (Serbian)
- removing special symbols, as ”#”, ”%” and ”@”
- Slovak bi-text was removed from the Tatoeba corpus

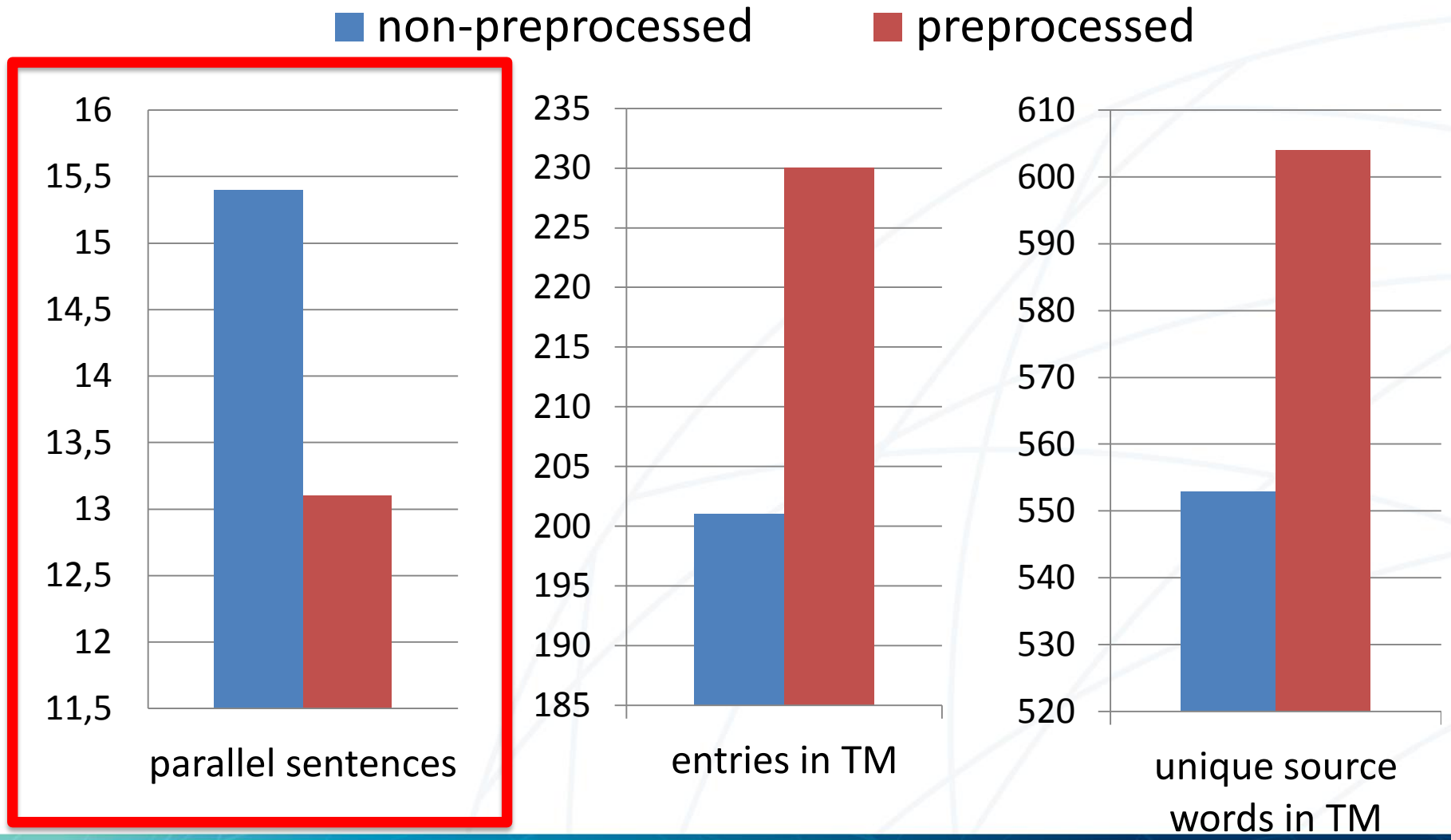
Corpora filtering based on the sentence length proportions

- too short and too long sentences were not included into the training set (average length +/- standard deviation)

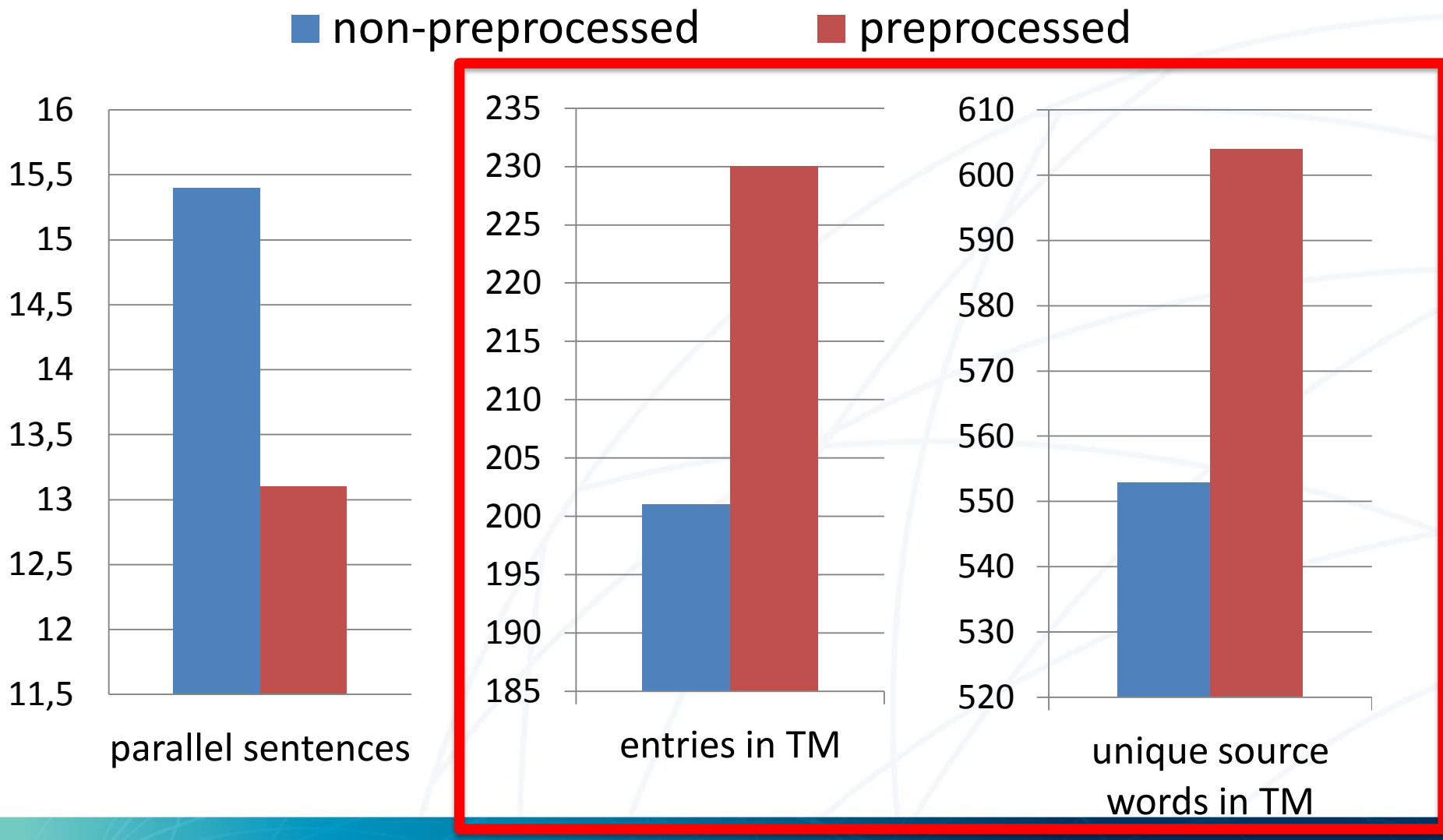
Evaluation / Development set creation

- avoiding too short/long sentences (10 and 40 words / 5 to 15 words)

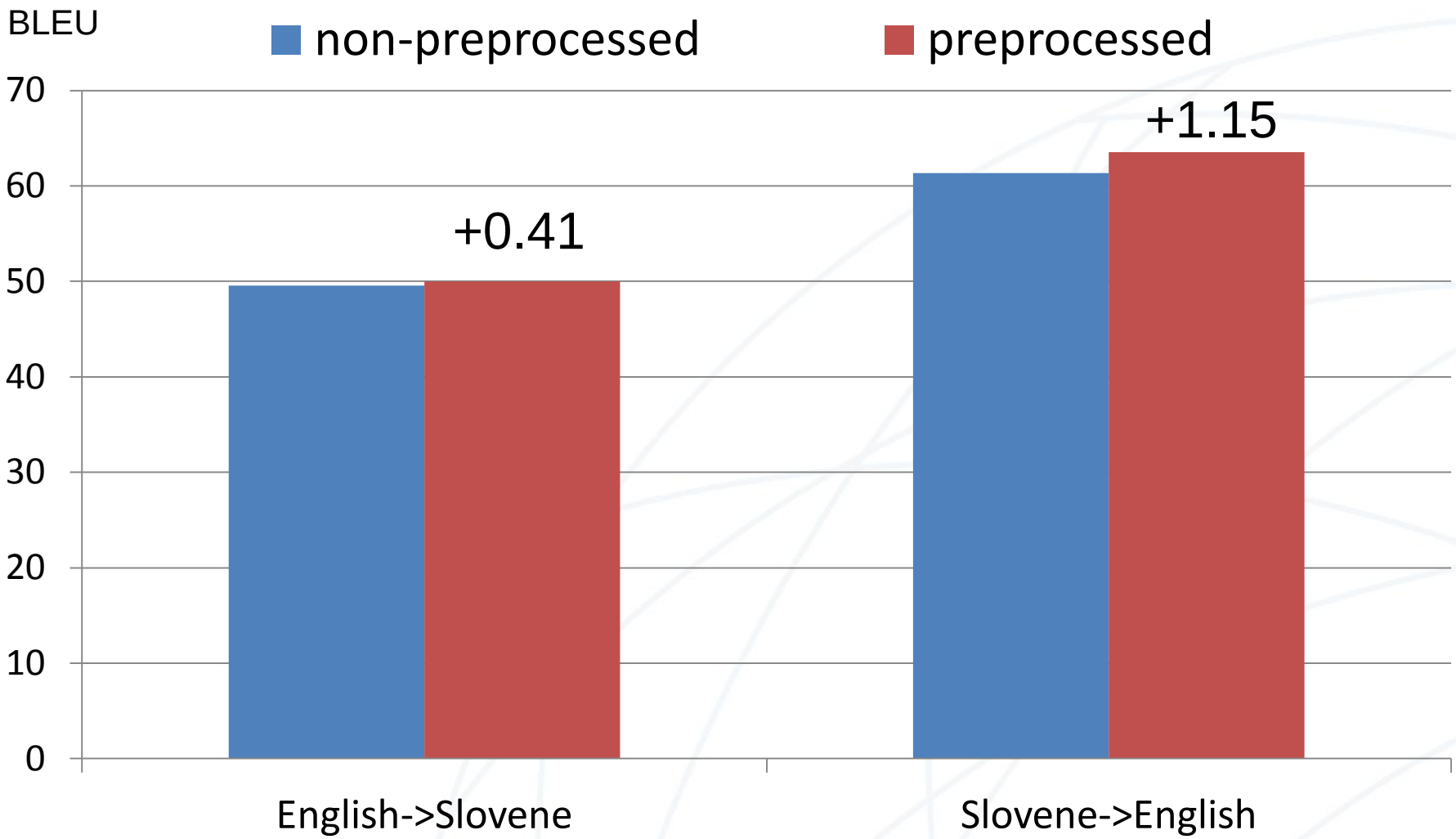
Data Preparation (example English-Slovene)



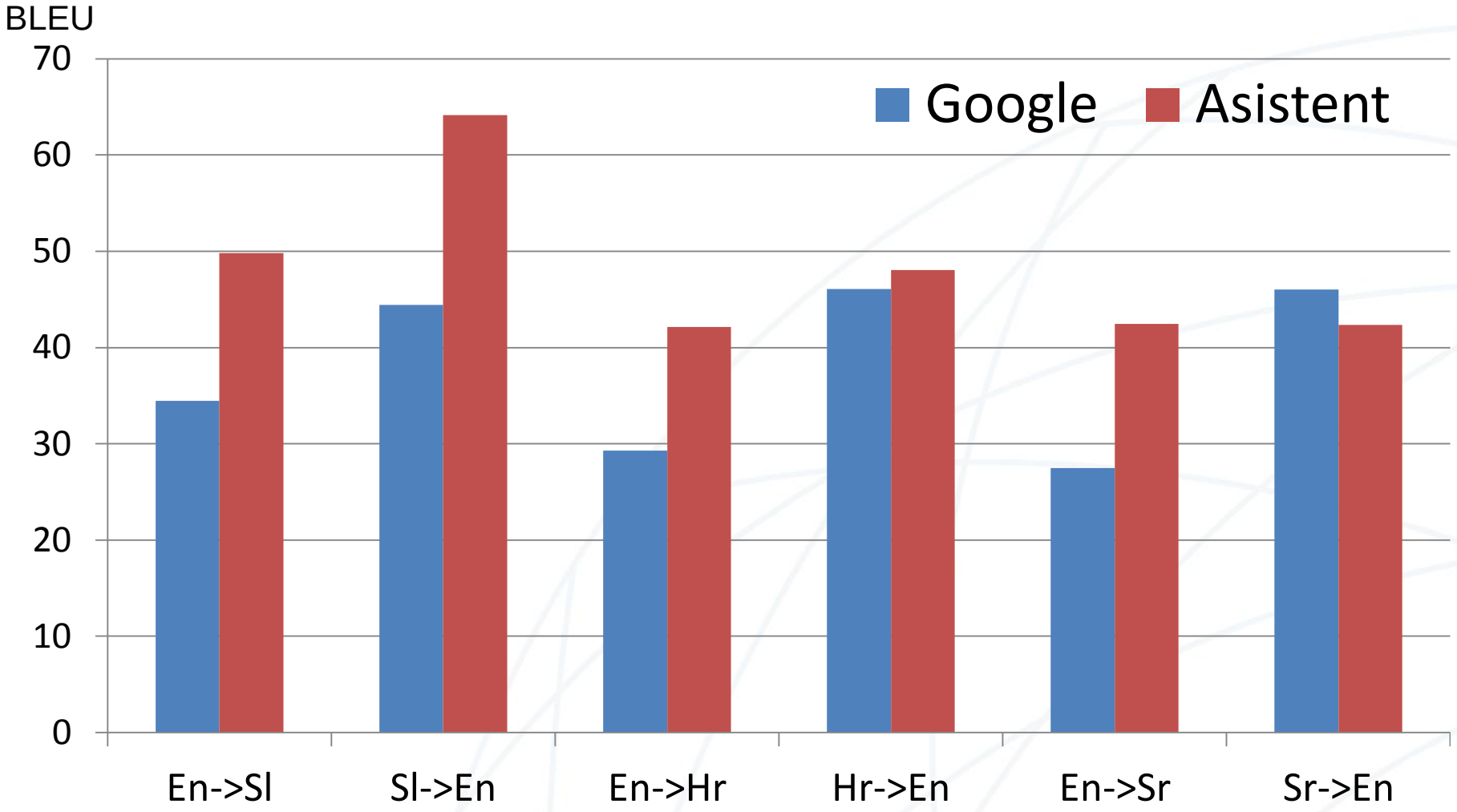
Data Preparation (example English-Slovene)



Data Preparation

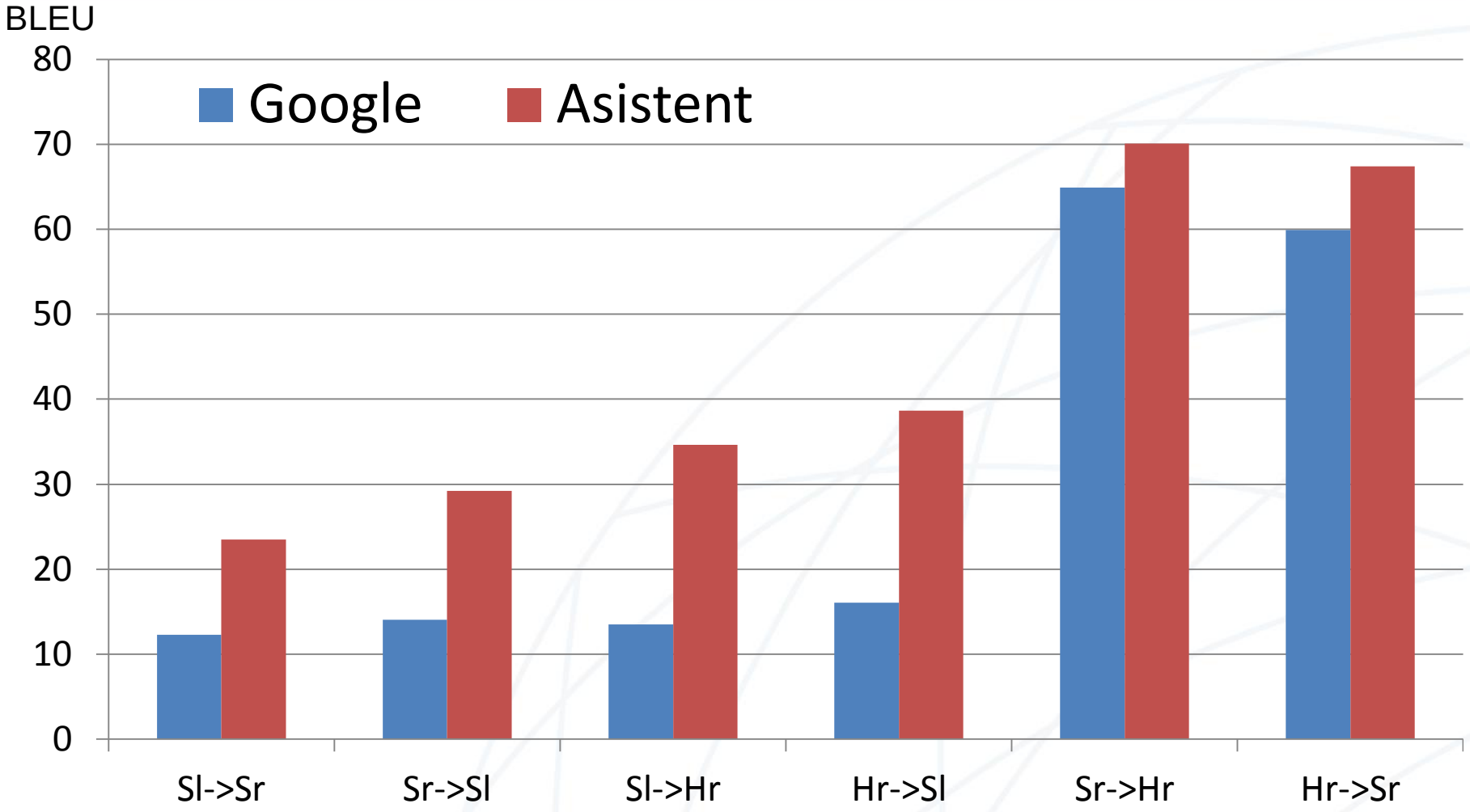


Automatic Evaluation, In-Domain (English-Slavic)



http://server1.nlp.insight-centre.org/asistent/data/asisten_evaluation_set.tar.gz

Automatic Evaluation, In-Domain (Slavic-Slavic)



http://server1.nlp.insight-centre.org/asistent/data/asisten_evaluation_set.tar.gz

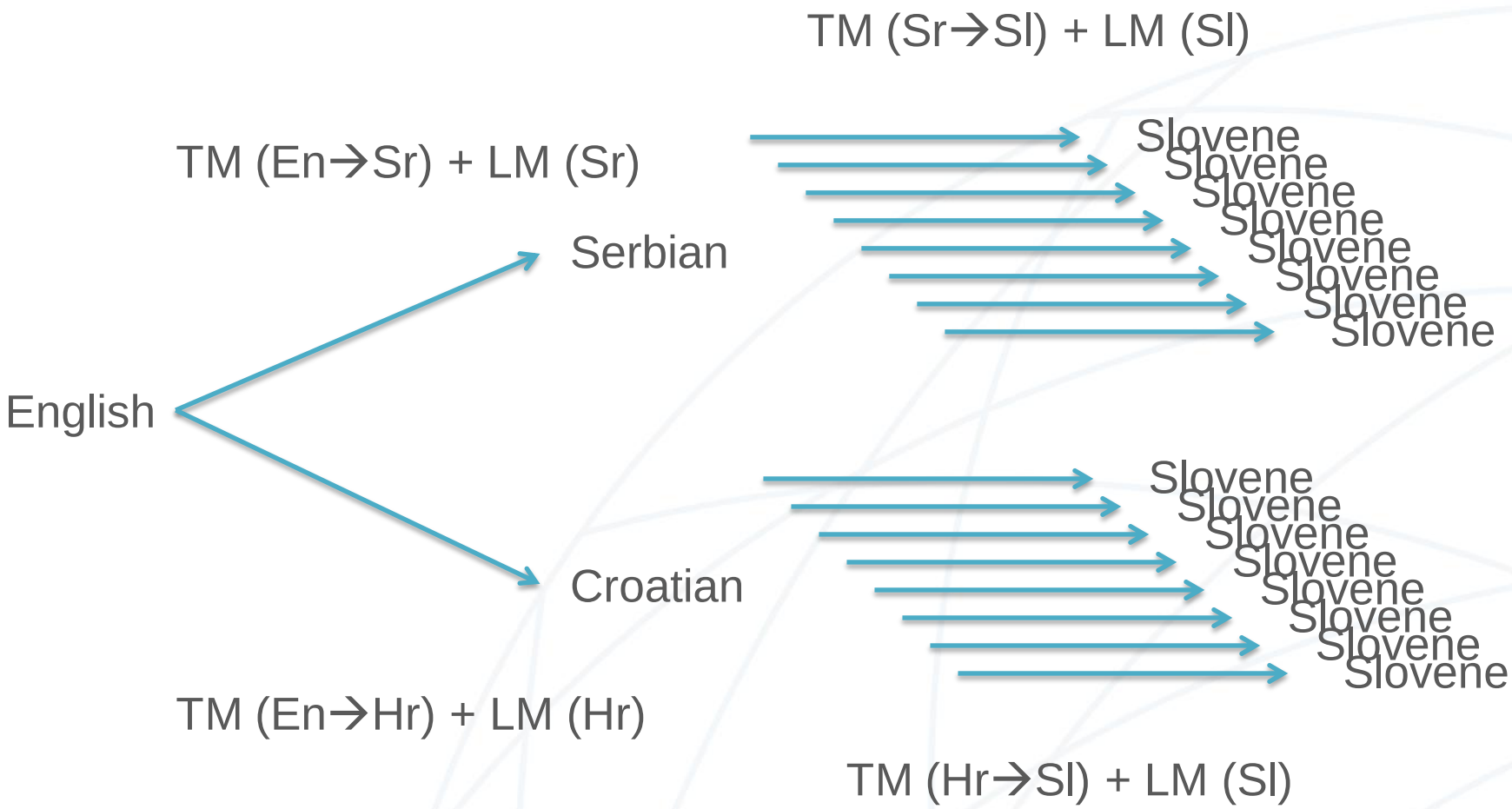
Pivot Translation

... can enable a bridge between languages, when existing parallel corpora are under-resourced ...

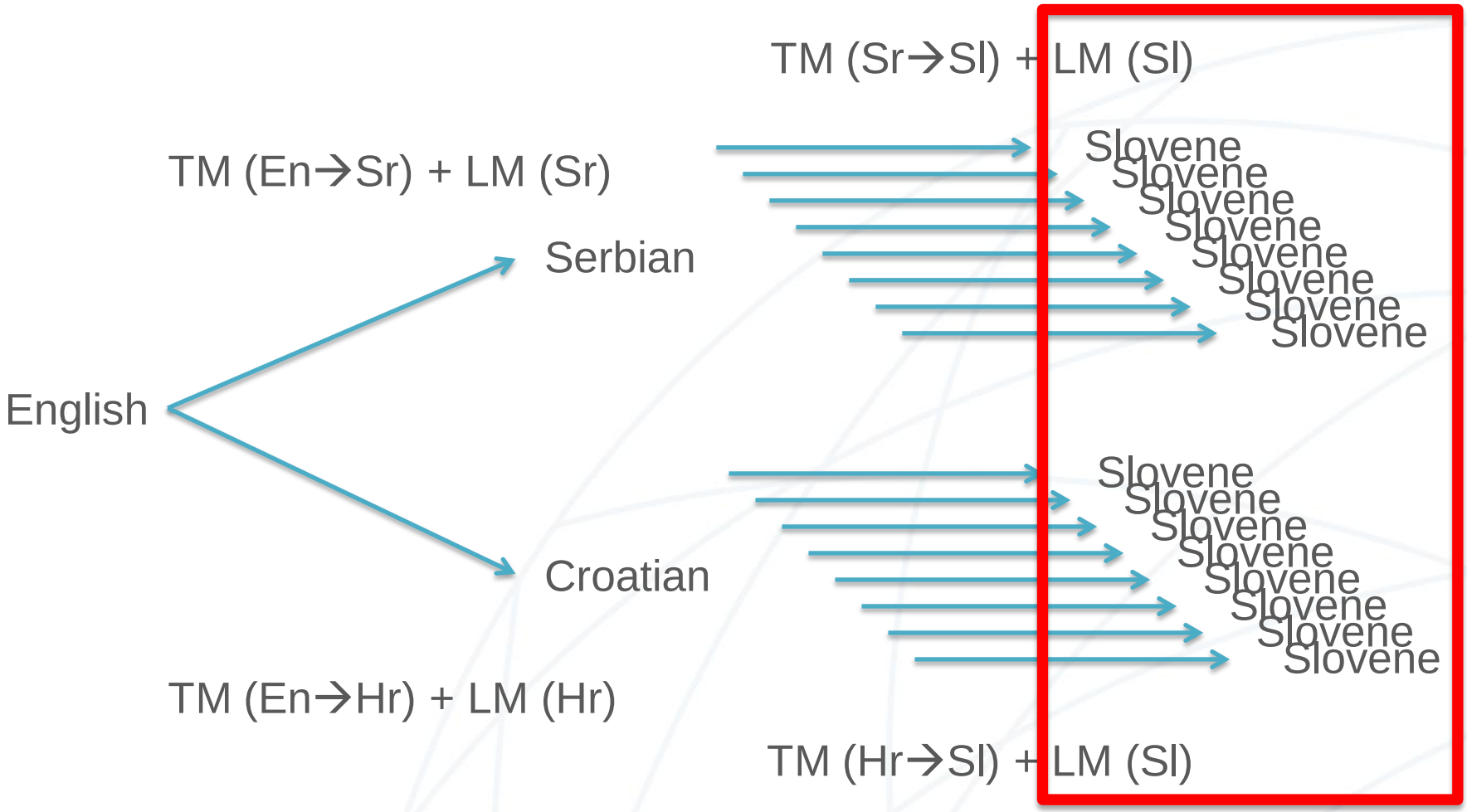
Direct Translation: source language → target language

Pivot Translation: source language → pivot language → target language

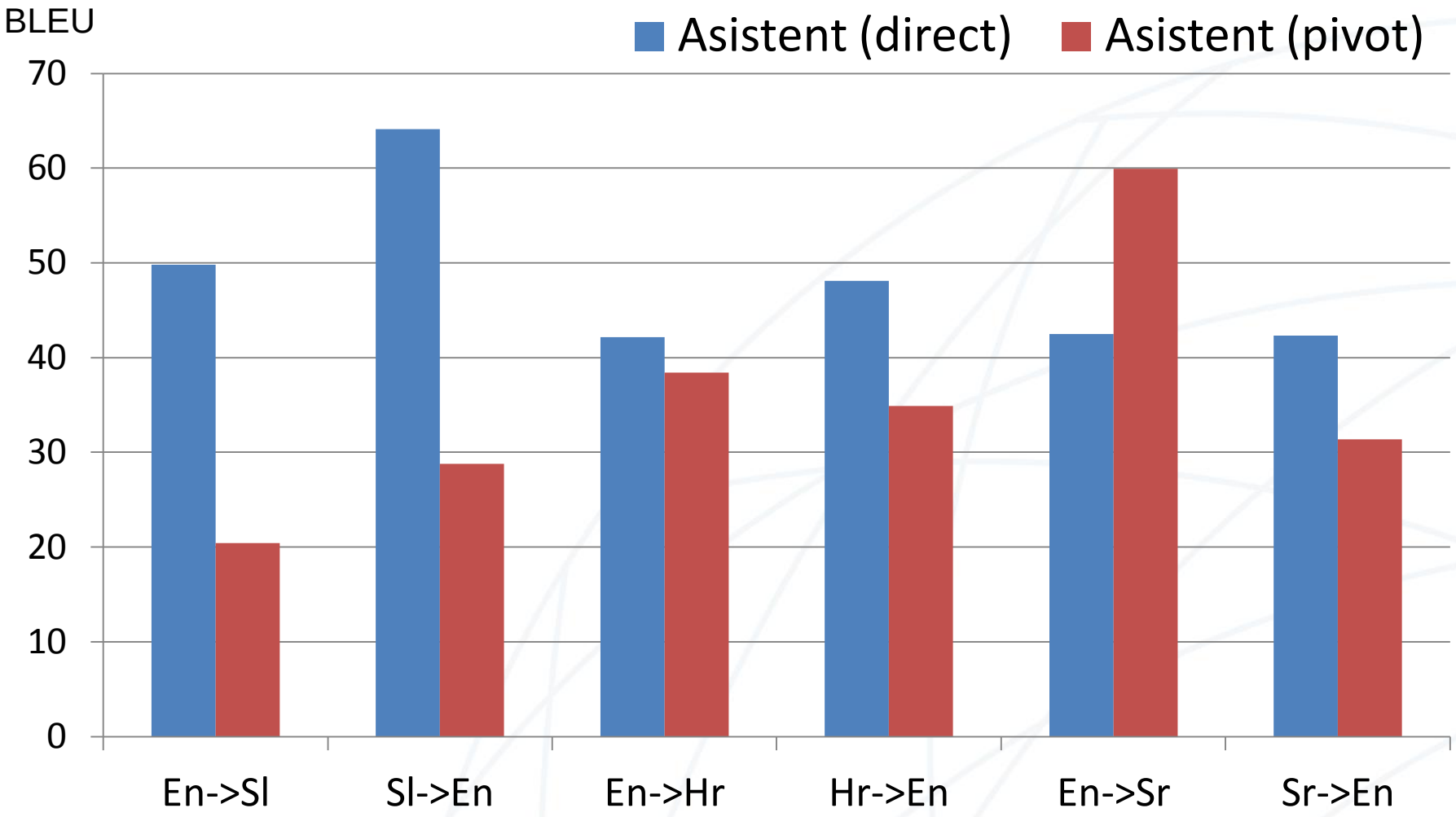
Pivot Translation



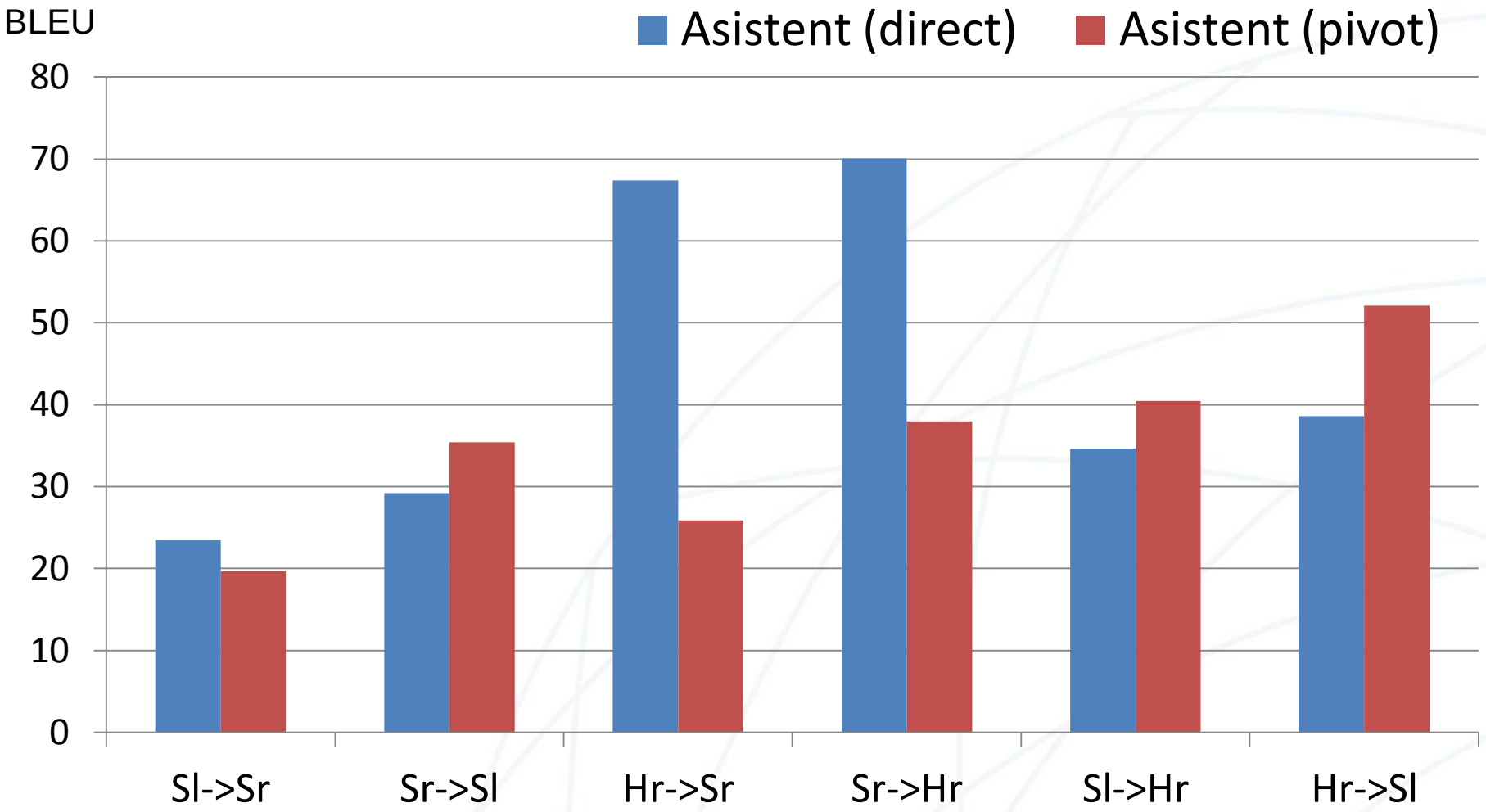
Pivot Translation



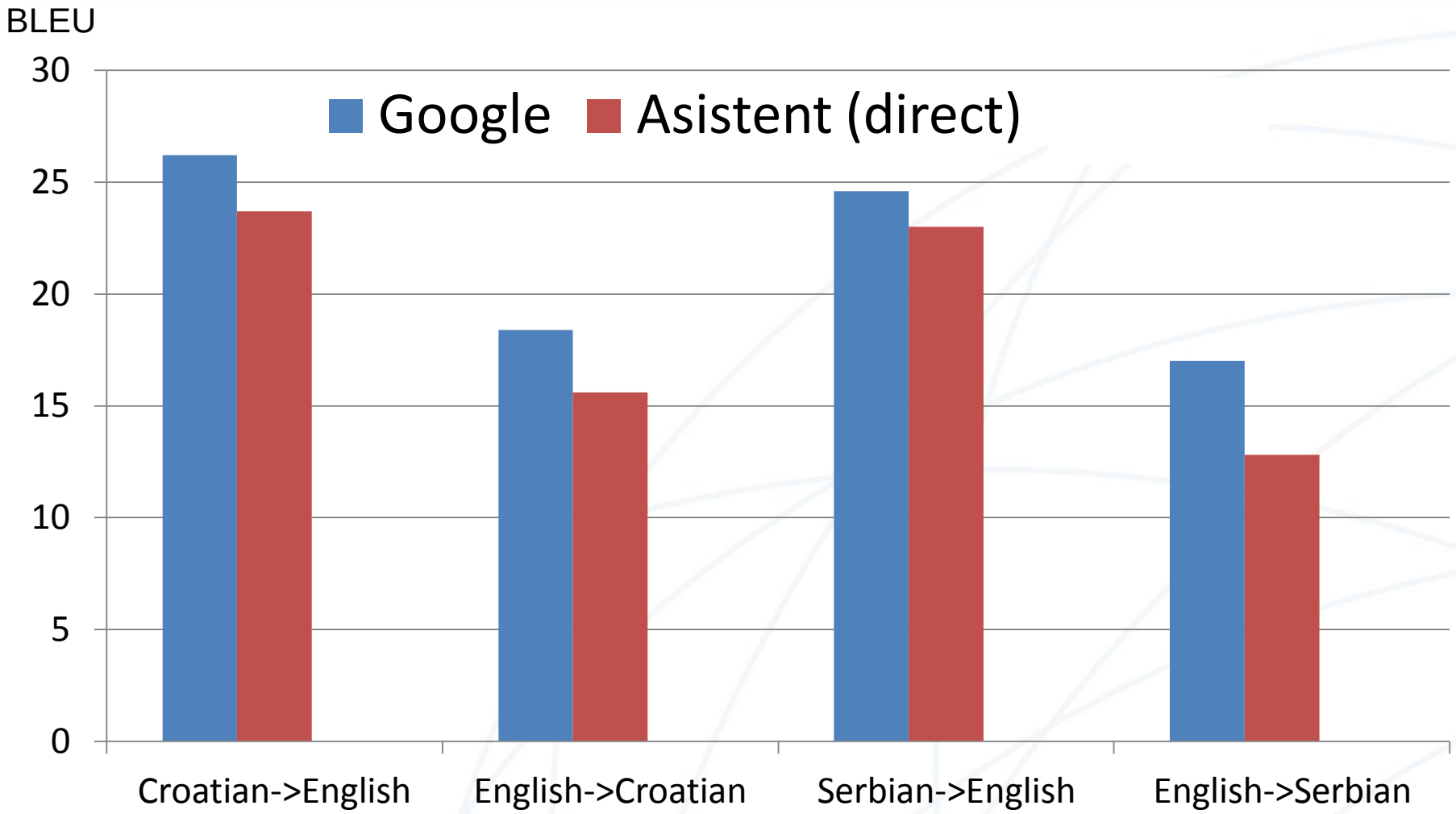
Evaluation for Pivot Translation, In-Domain (English-Slavic)



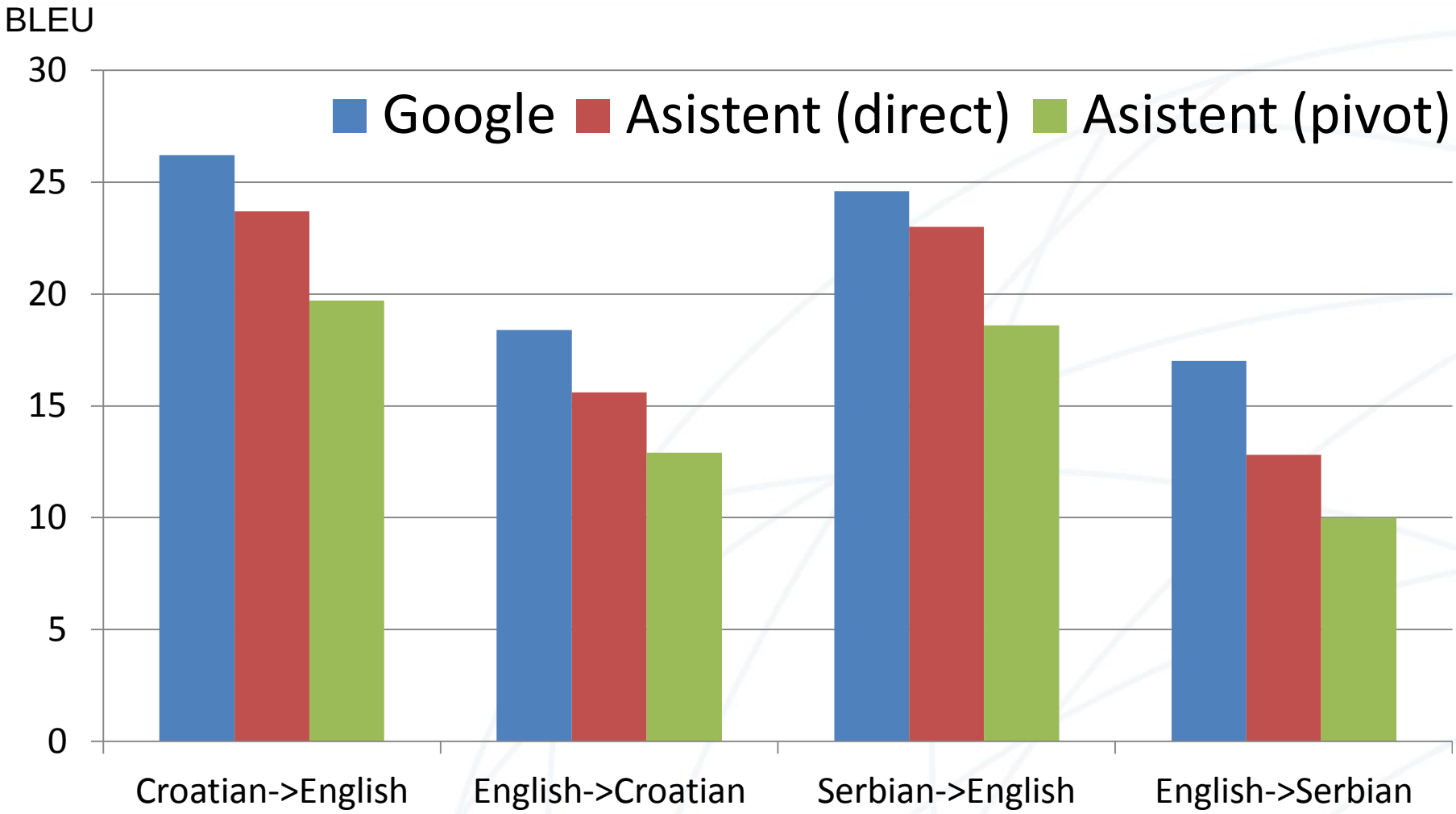
Evaluation for Pivot Translation, In-Domain (Slavic-Slavic)



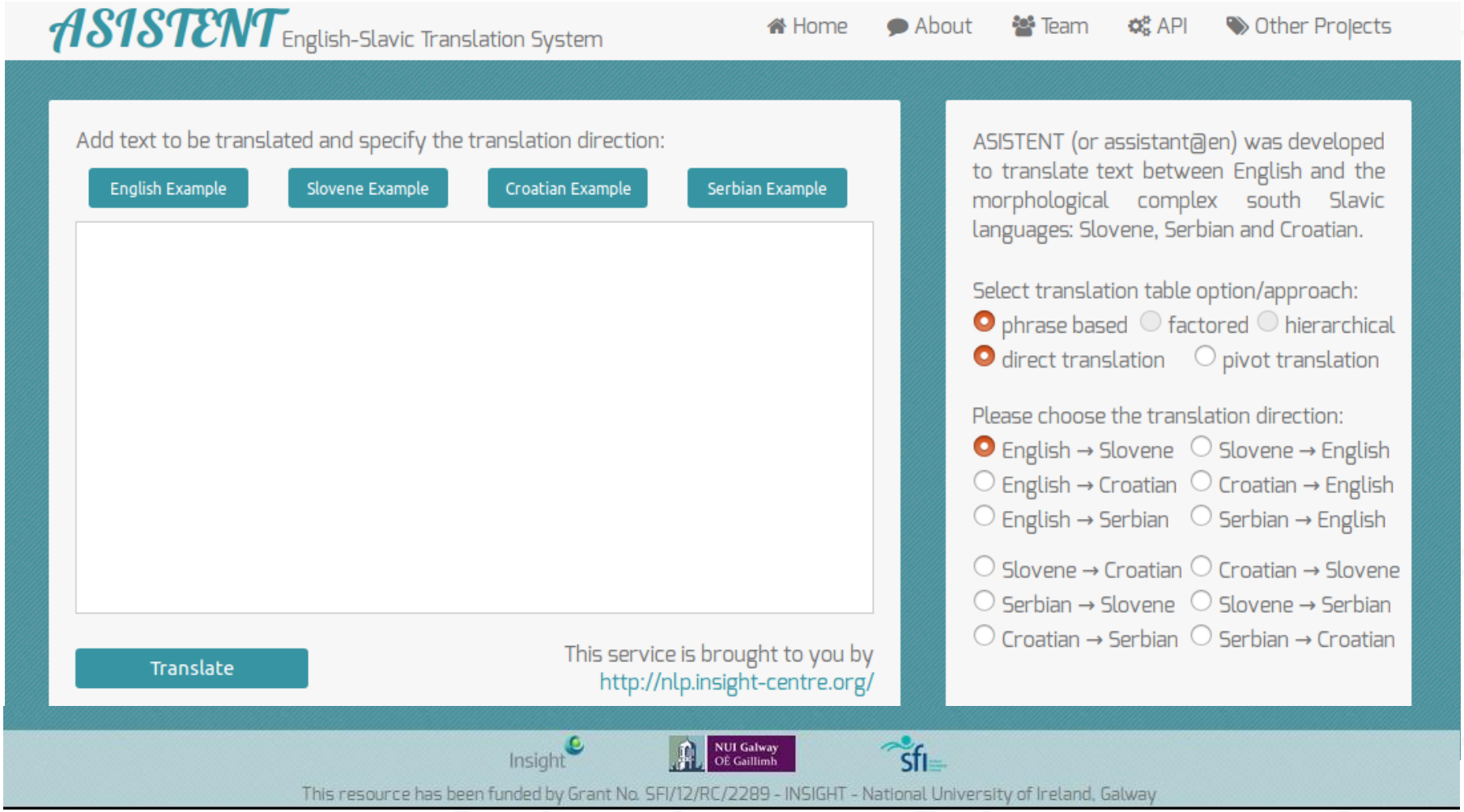
Automatic Evaluation, Out-Domain
(Massive Open Online Courses MOOCs)



Automatic Evaluation, Out-Domain
(Massive Open Online Courses MOOCs)

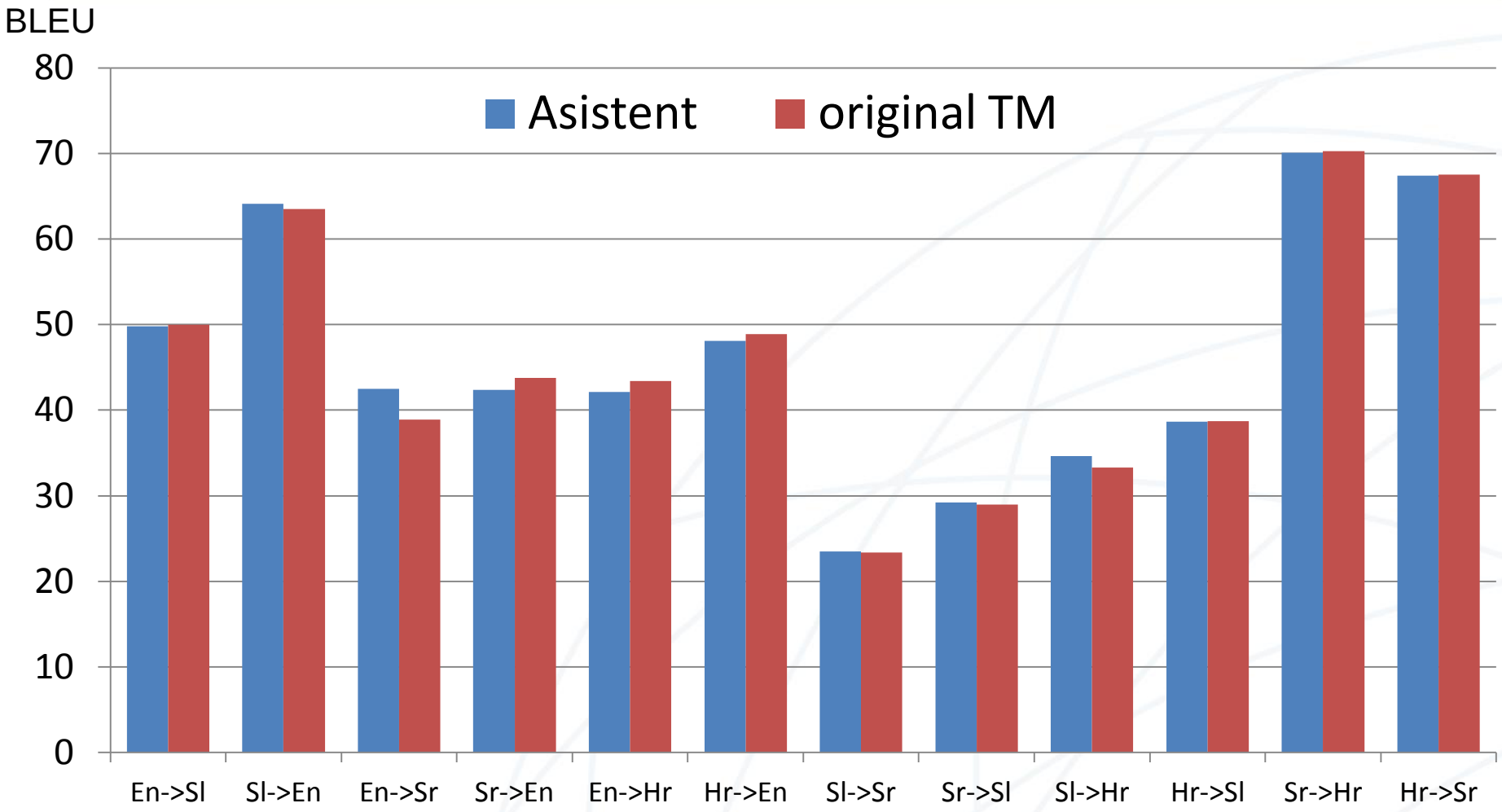


Asistent



http://server1.nlp.insight-centre.org/asistent/

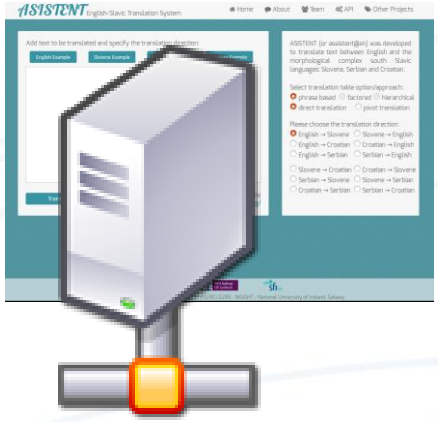
Filtered vs Original Translation Models



Asistent's API request

Translation request (json)

Provided translation by Asistent



```
{    "nbest":"5",
     "translation_direction":"en_sl",
     "method":"phrase_based",
     "approach":"direct",
     "text2translate":[
          {"source":"Accusations of witchcraft are also common in other African
               countries."}
     ]
}
```

http://server1.nlp.insight-centre.org/asistent/rest_service.html

Asistent's API request


```
{
    "time":"6 wallclock secs ( 0.02 usr 0.01 sys + 5.16 cusr 0.42 csys = 5.61 CPU)",
    "translation_direction":"en_sl",
    "nbest":"3",
    "method":"phrase_based",
    "approach":"direct",
    "text2translate":[
        {
            "source":"Accusations of witchcraft are also common in other African countries.",
            "possible_translations":{
                "obtožbe so pogosti tudi v čarovništva , druge afriške države . ":"-9.741",
                "obtožbe čarovništva so pogosti tudi v drugih afriških državah . ":"-9.644",
                "obtožbe o čarovništvu so pogosti tudi v drugih afriških državah . ":"-9.706"
        },
        "best":"obtožbe čarovništva so pogosti tudi v drugih afriških državah . "
    }
    ],
}
```

Future Work

Comparison between different SMT methods

- hierarchical models
- factored models (added linguistic information)
- neural machine translation

Manual Evaluation

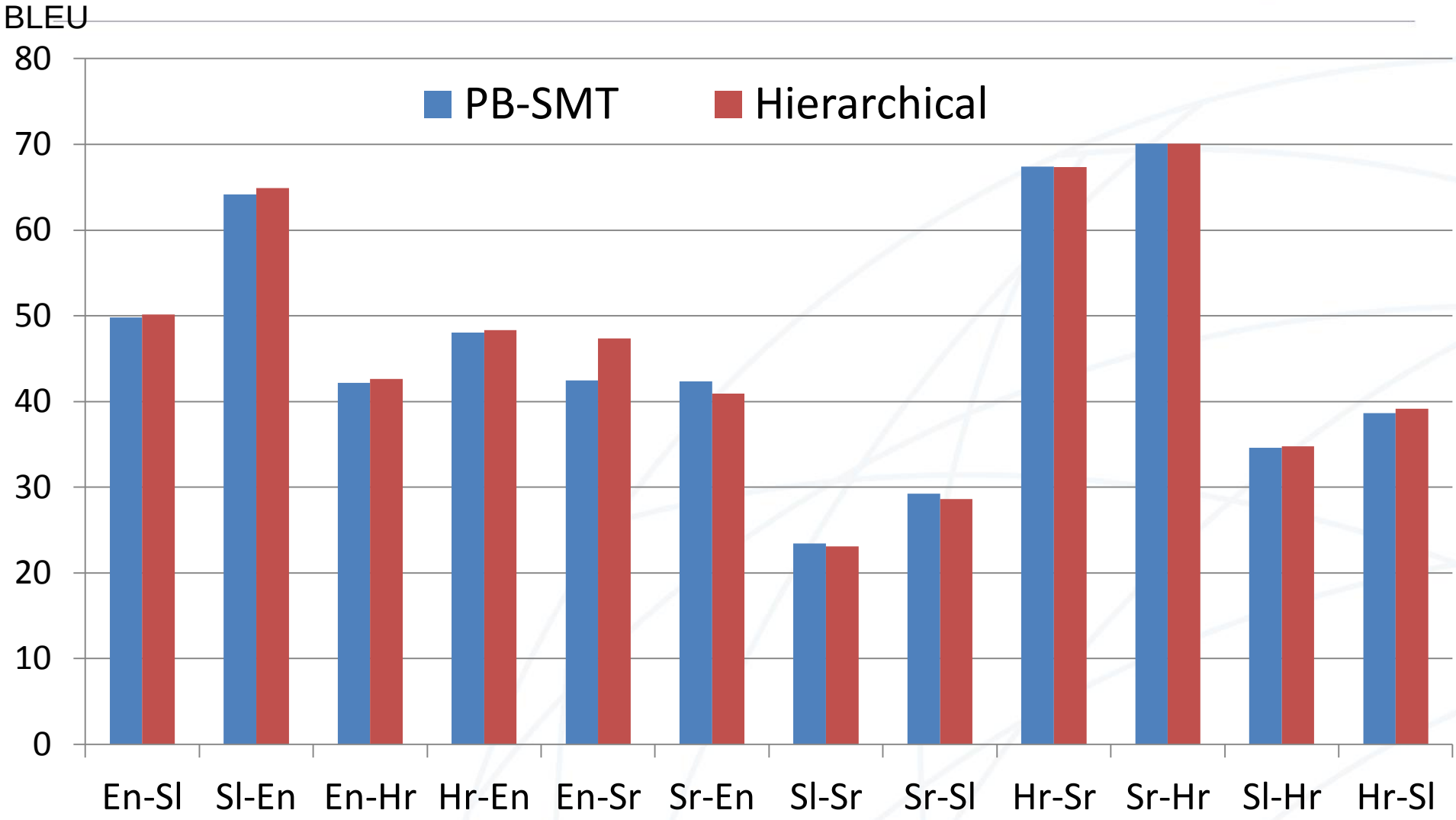
- pivot translations
- phrase-based vs hierarchical models

Making Asistent/TM (more) accessible

Asistent - A Machine Translation System for Slovene, Serbian and Croatian

Mihael Arcan

mihael.arcan@insight-centre.org

Thank You

Hierarchical Models Evaluation (BLEU)