



Provided by the author(s) and NUI Galway in accordance with publisher policies. Please cite the published version when available.

Title	Asistent -- a machine translation system for Slovene, Serbian and Croatian
Author(s)	Arcan, Mihael; Popovic, Maja; Buitelaar, Paul
Publication Date	2016-09-29
Publication Information	Arcan, Mihael, Popovic, Maja, & Buitelaar, Paul. (2016). Asistent -- a machine translation system for Slovene, Serbian and Croatian. Paper presented at the 10th Conference on Language Technologies and Digital Humanities, Ljubljana, Slovenia, 29 September - 01 October.
Publisher	University of Ljubljana
Link to publisher's version	http://www.sdjt.si/wp/dogodki/konference/jdth-2016-english/
Item record	http://hdl.handle.net/10379/14893

Downloaded 2020-10-23T08:58:08Z

Some rights reserved. For more information, please see the item record link above.



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

Add text to be translated and specify the translation direction:

English Example

Slovene Example

Croatian Example

Serbian Example

Translate

This service is brought to you by
<http://nlp.insight-centre.org/>

ASISTENT (or `assistant@en`) was developed to translate text between English and the morphological complex south Slavic languages: Slovene, Serbian and Croatian.

Select translation table option/approach:

- phrase based factored hierarchical
 direct translation pivot translation

Please choose the translation direction:

- English → Slovene Slovene → English
 English → Croatian Croatian → English
 English → Serbian Serbian → English
 Slovene → Croatian Croatian → Slovene
 Serbian → Slovene Slovene → Serbian
 Croatian → Serbian Serbian → Croatian

META-NET Language White Paper Series (2012)

Machine Translation

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	<ul style="list-style-type: none">English	<ul style="list-style-type: none">FrenchSpanish	<ul style="list-style-type: none">CatalanDutchGermanHungarianItalianPolishRomanian	<ul style="list-style-type: none">BasqueBulgarianCroatianCzechDanishEstonianFinnishGalicianGreekIcelandicIrishLatvianLithuanianMalteseNorwegian (Bokmål, Nynorsk)PortugueseSerbianSlovakSloveneSwedishWelsh

What Do We Need?

Translation Toolkit:

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

Parallel data

- Europarl
- DGT
- OpenSubTitles

....



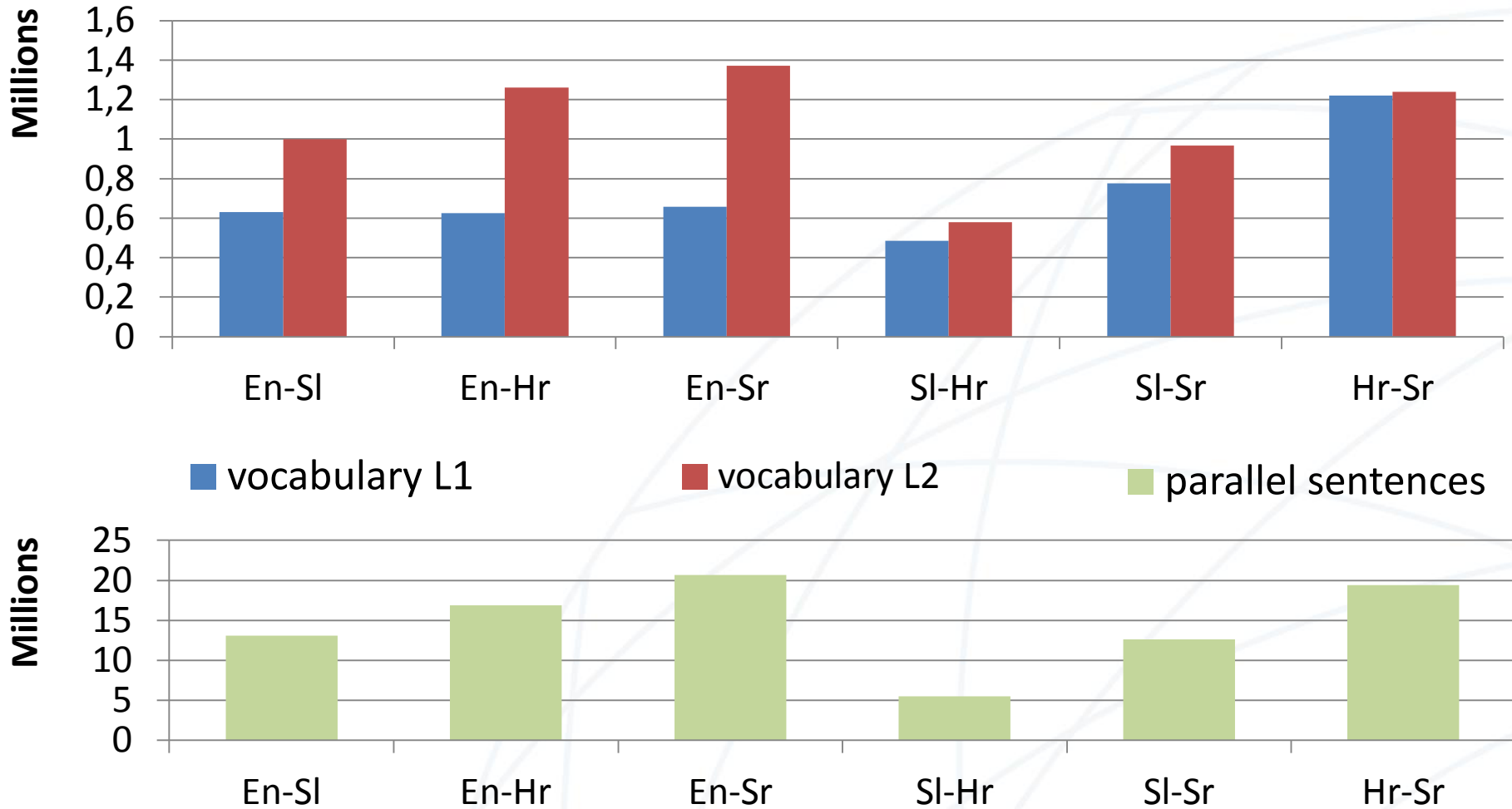
<http://server1.nlp.insight-centre.org/asistent/>

Parallel Corpora Used for Asistent

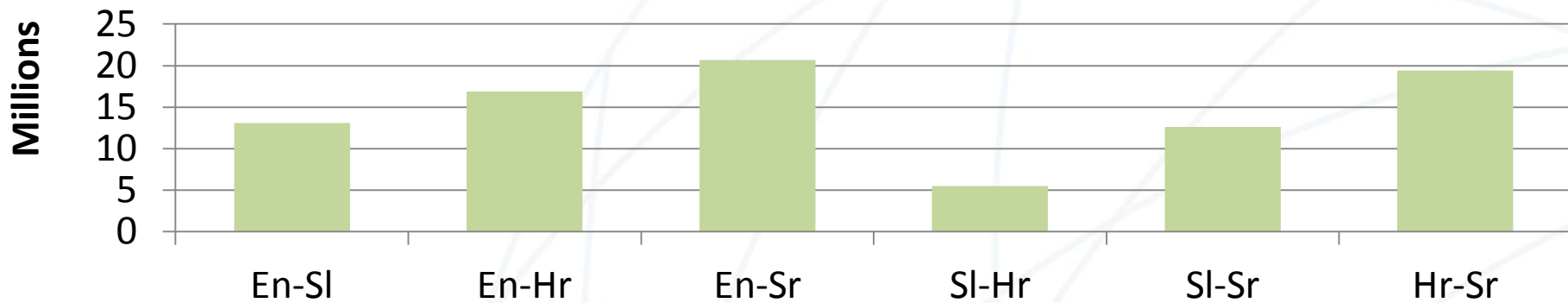
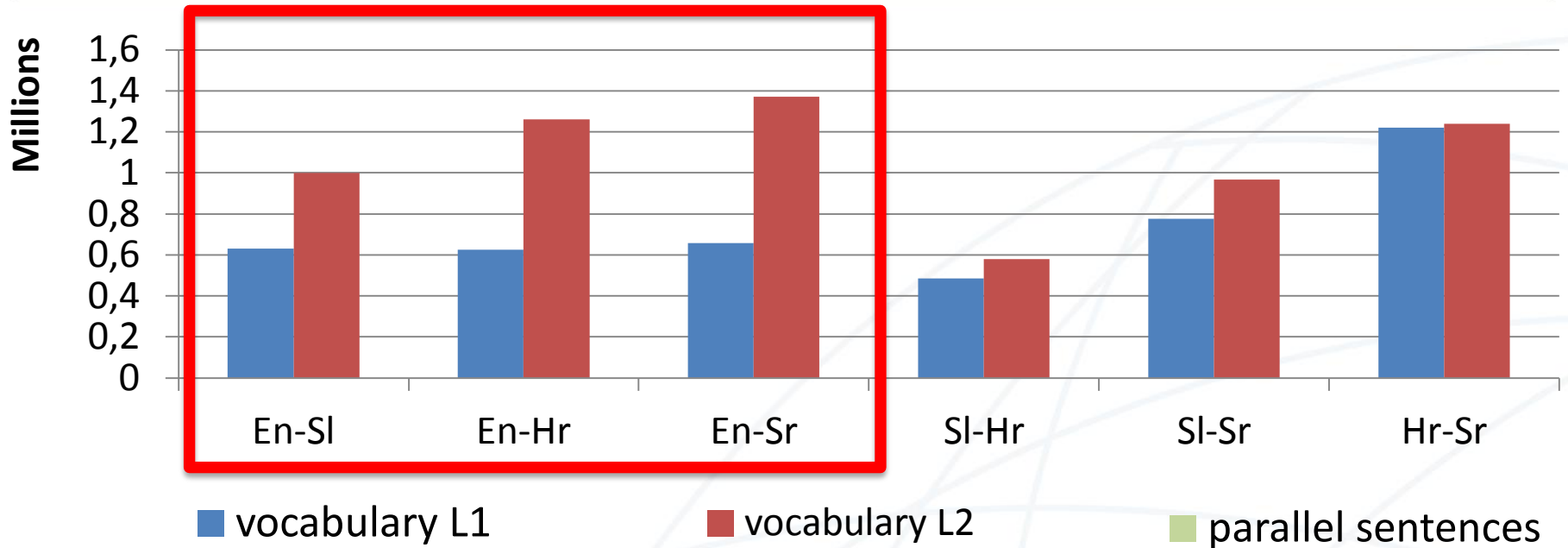
Name	Domain	En-Sl	En-Hr	En-Sr	Sl-Hr	Sl-Sr	Hr-Sr
DGT	legal	1.8M	196K				
ECB	finance	79K					
EMEA	medical	253K					
Europarl	EP proceedings	599K					
Gnome	IT	998	5K	126	4K	600K	300K
hrenWaC	Web Corpus		86K				
JRC-Acquis	legal	29K	38K				
KDE	IT	58K		32K	85K	49k	33.2k
LangCourse	education			3K			
PHP	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>

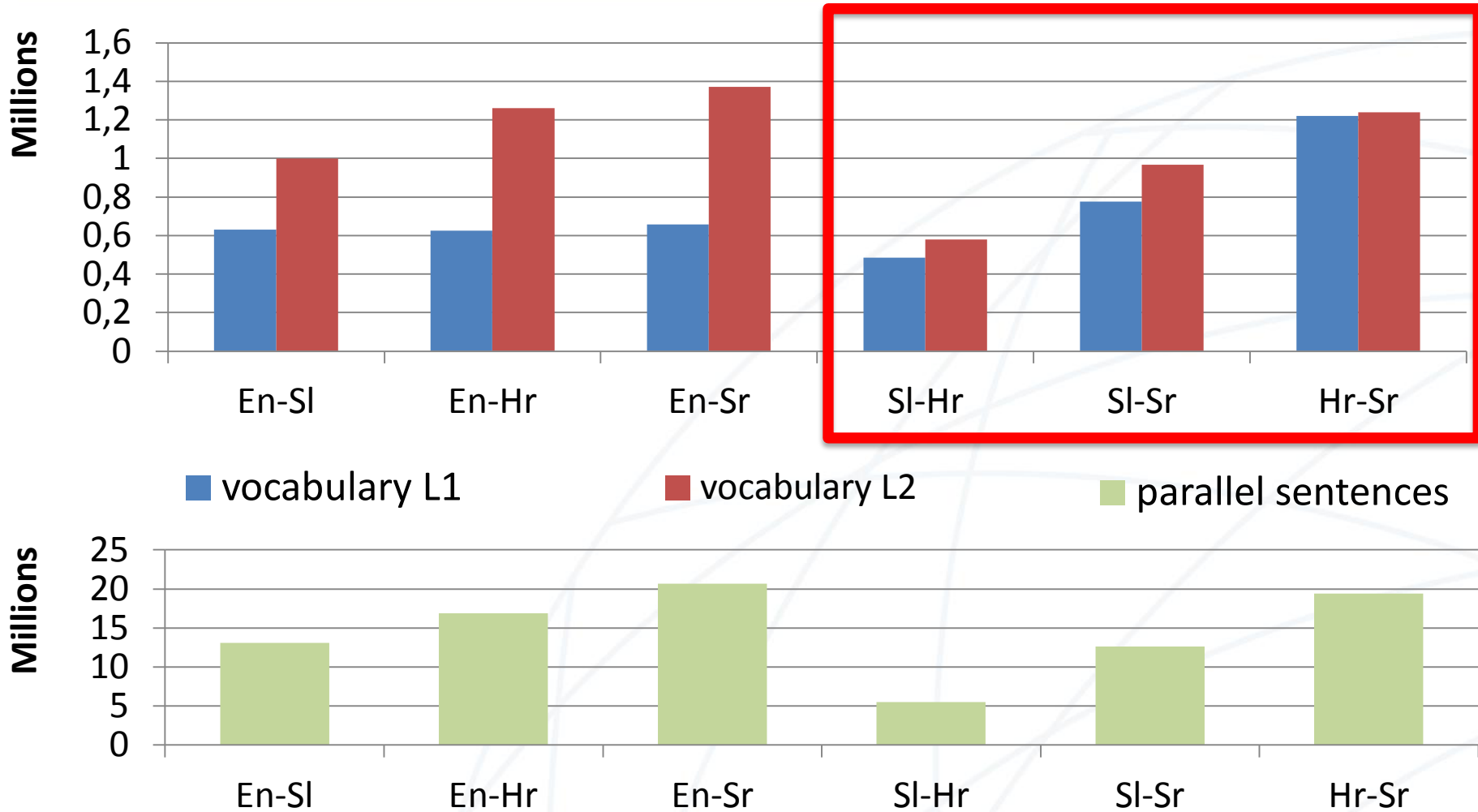
Concatenated Data for Asistent Training



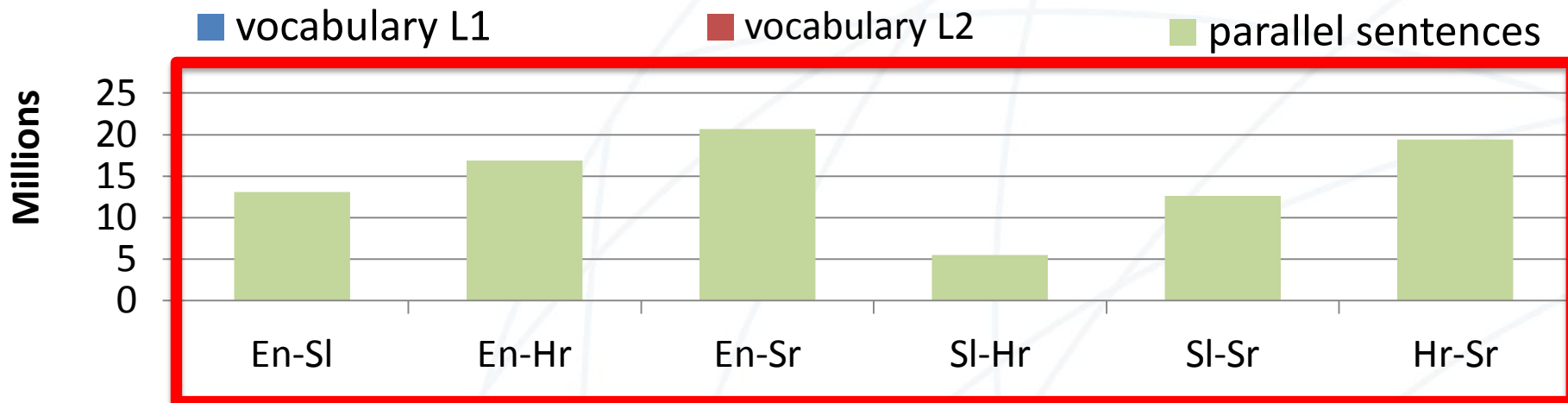
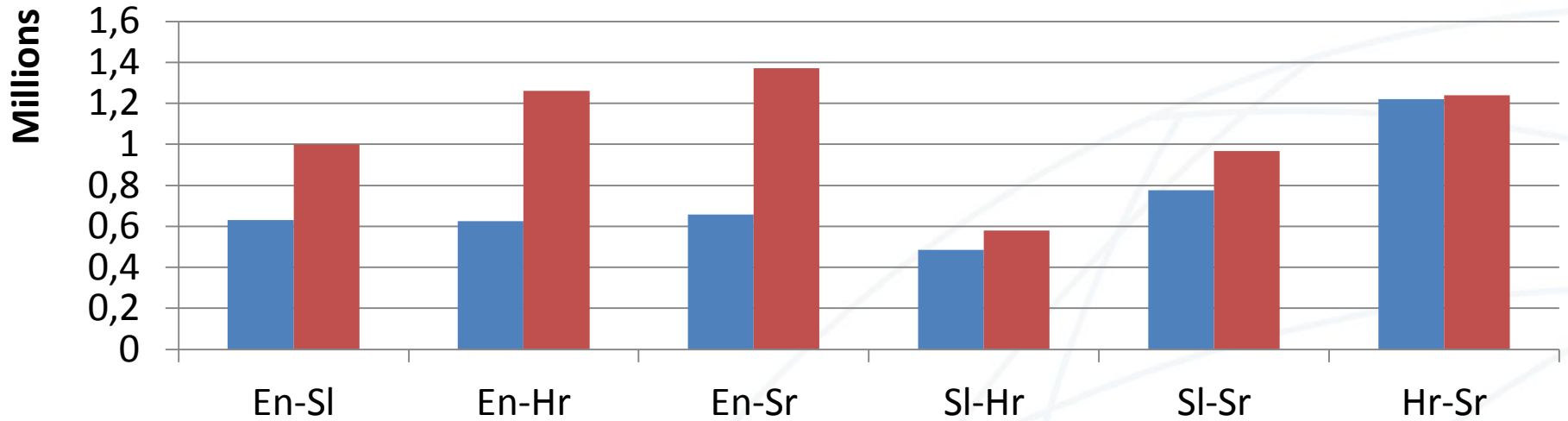
Concatenated Data for Asistent Training



Concatenated Data for Asistent Training



Concatenated Data for Asistent Training



Data 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

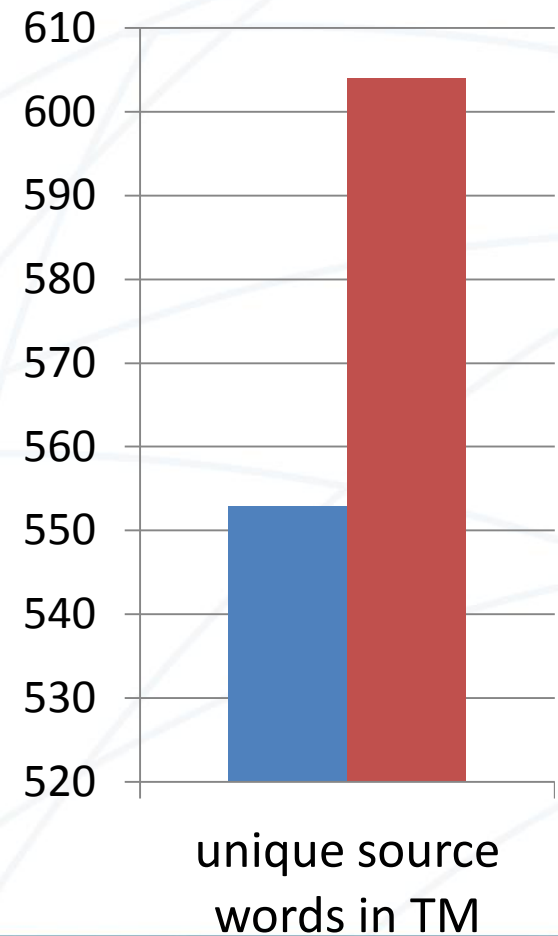
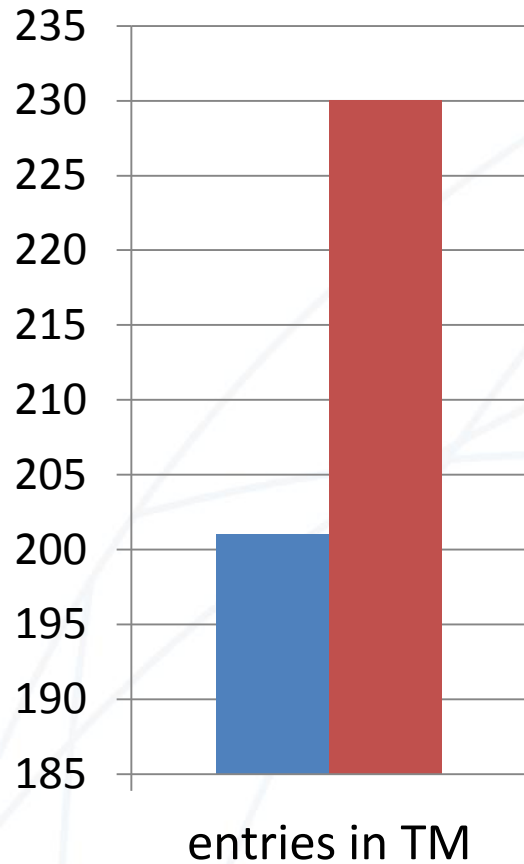
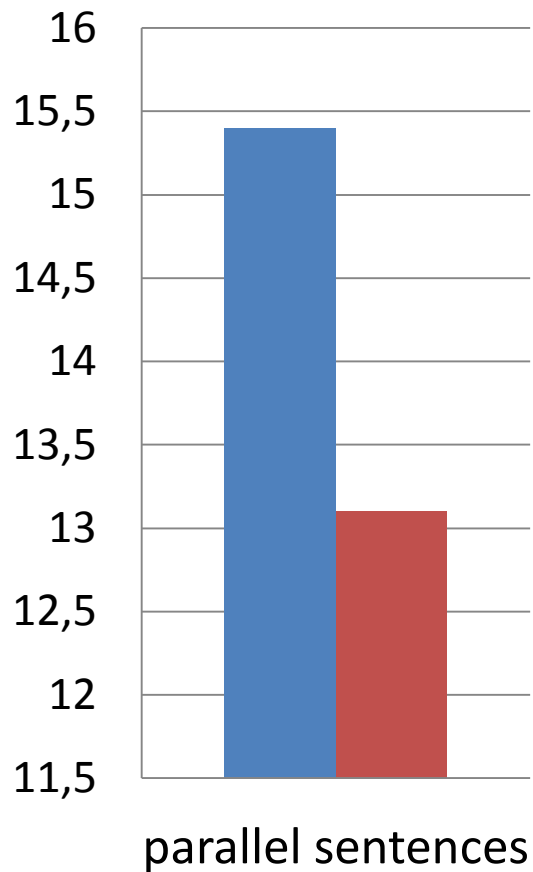
- too short and too long sentences were not included into the training set (average length +/- standard deviation)

Evaluation / Development set creation

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

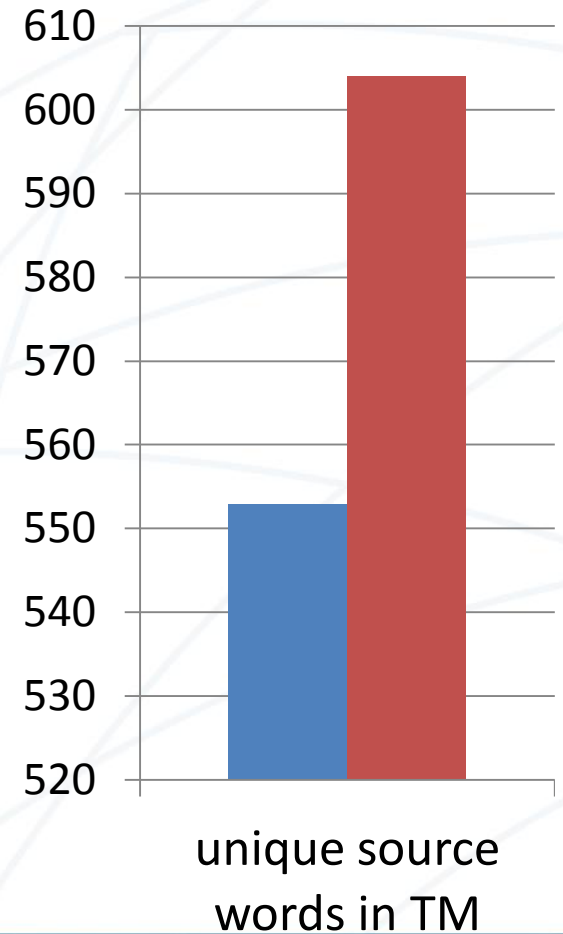
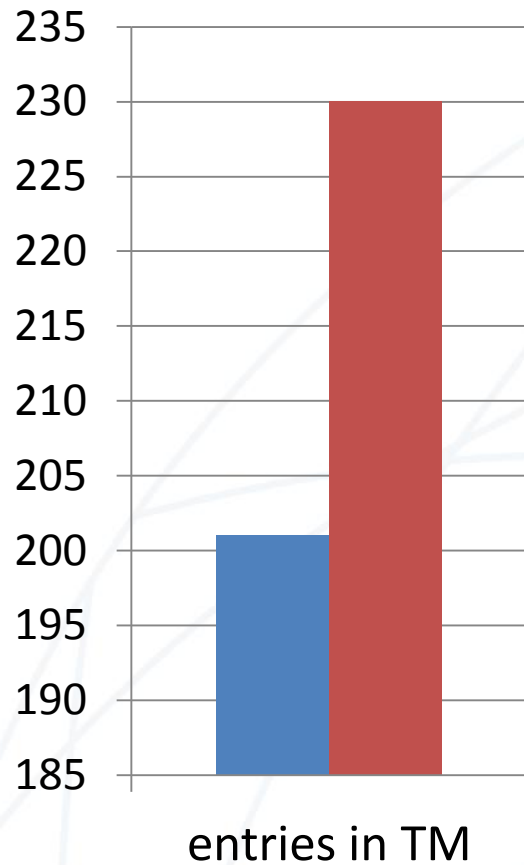
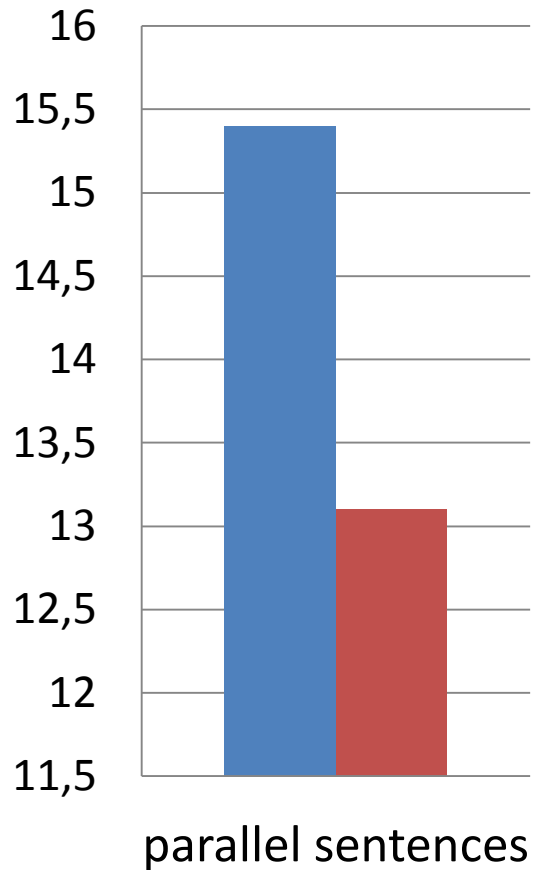
Data Preparation (example English-Slovene)

■ non-preprocessed ■ preprocessed

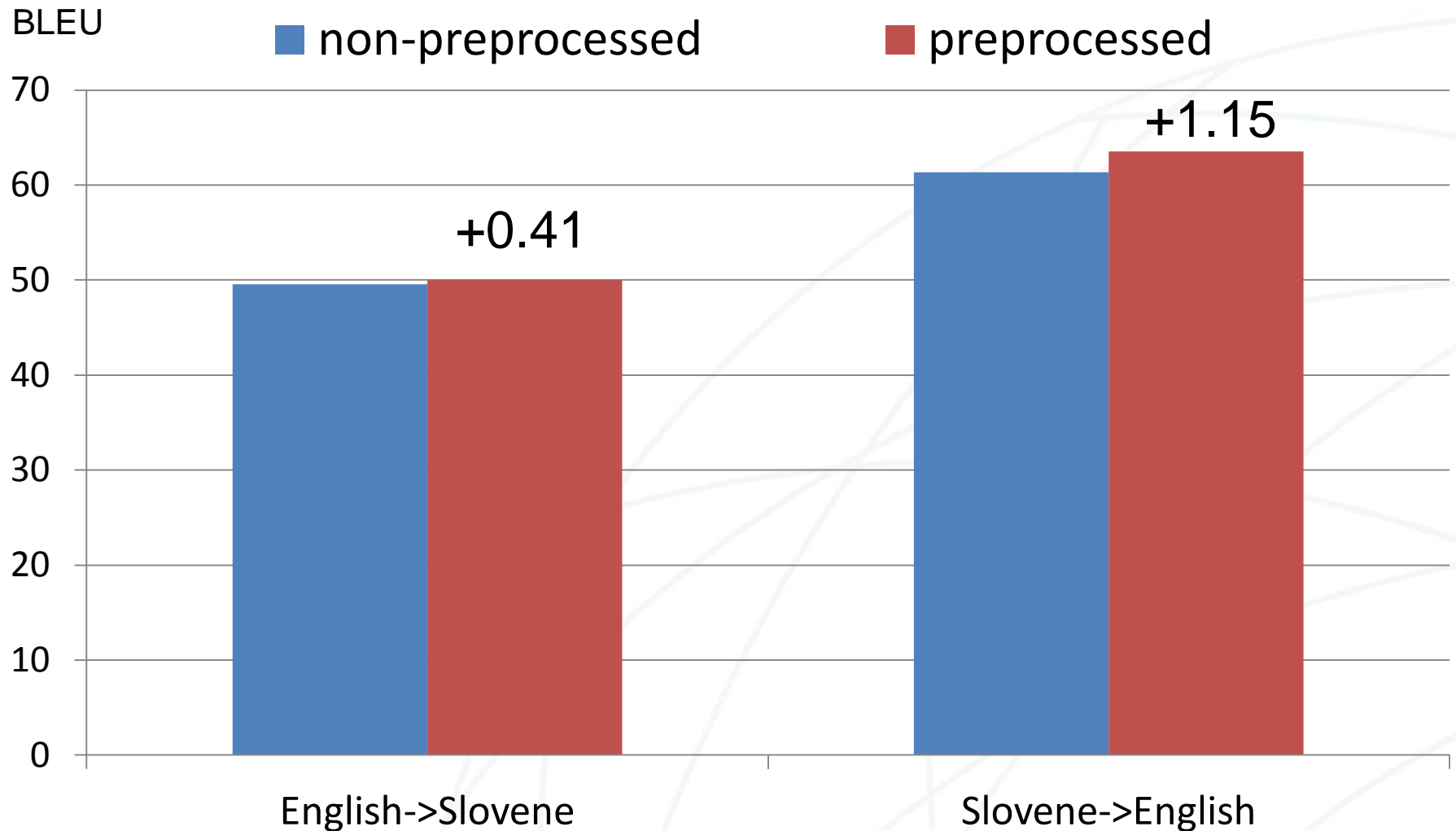


Data Preparation (example English-Slovene)

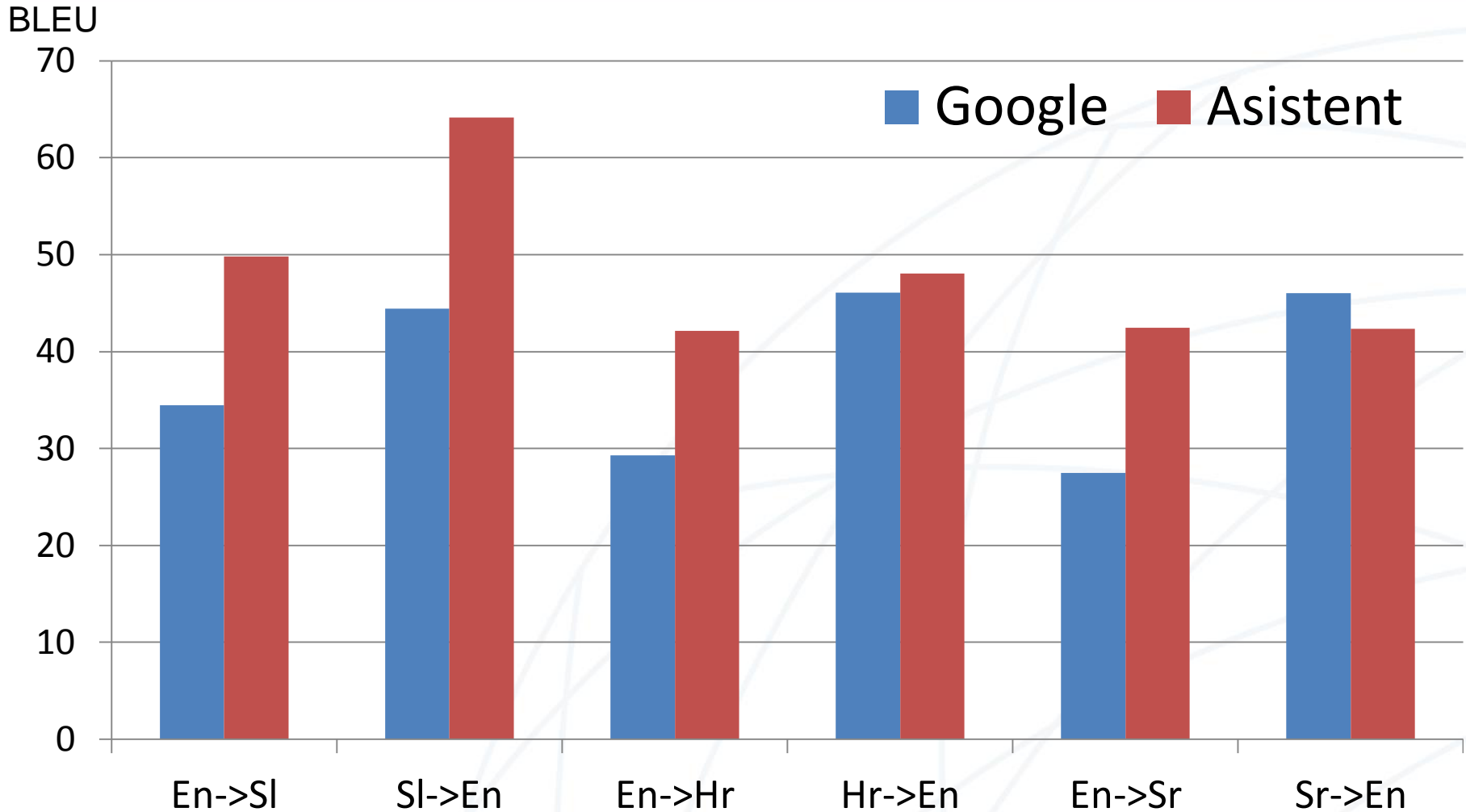
■ non-preprocessed ■ preprocessed



Data Preparation

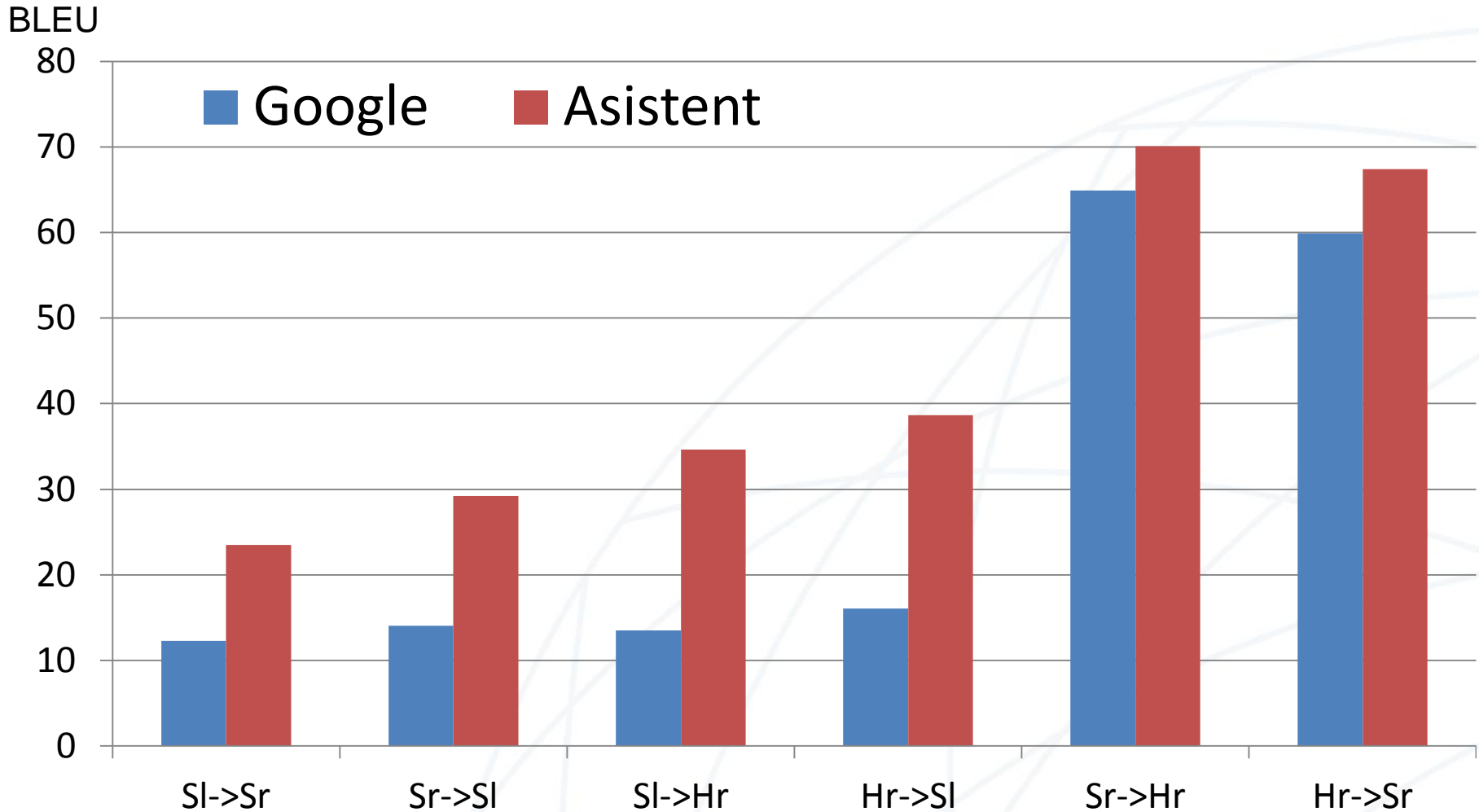


Automatic Evaluation, In-Domain (English-Slavic)



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

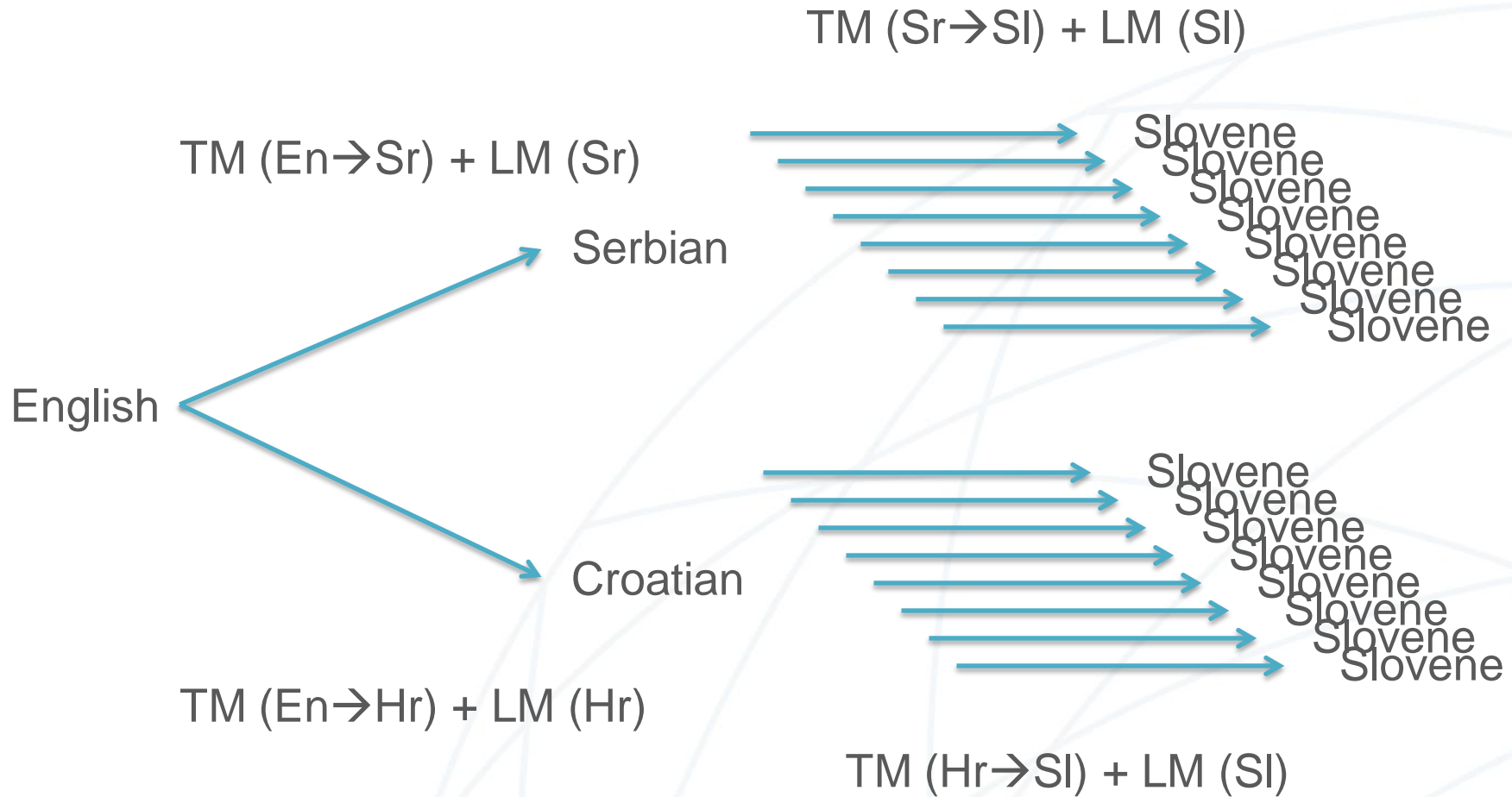
... can enable a bridge between languages, when existing parallel corpora are under-resourced ...

Direct Translation: source language → target language

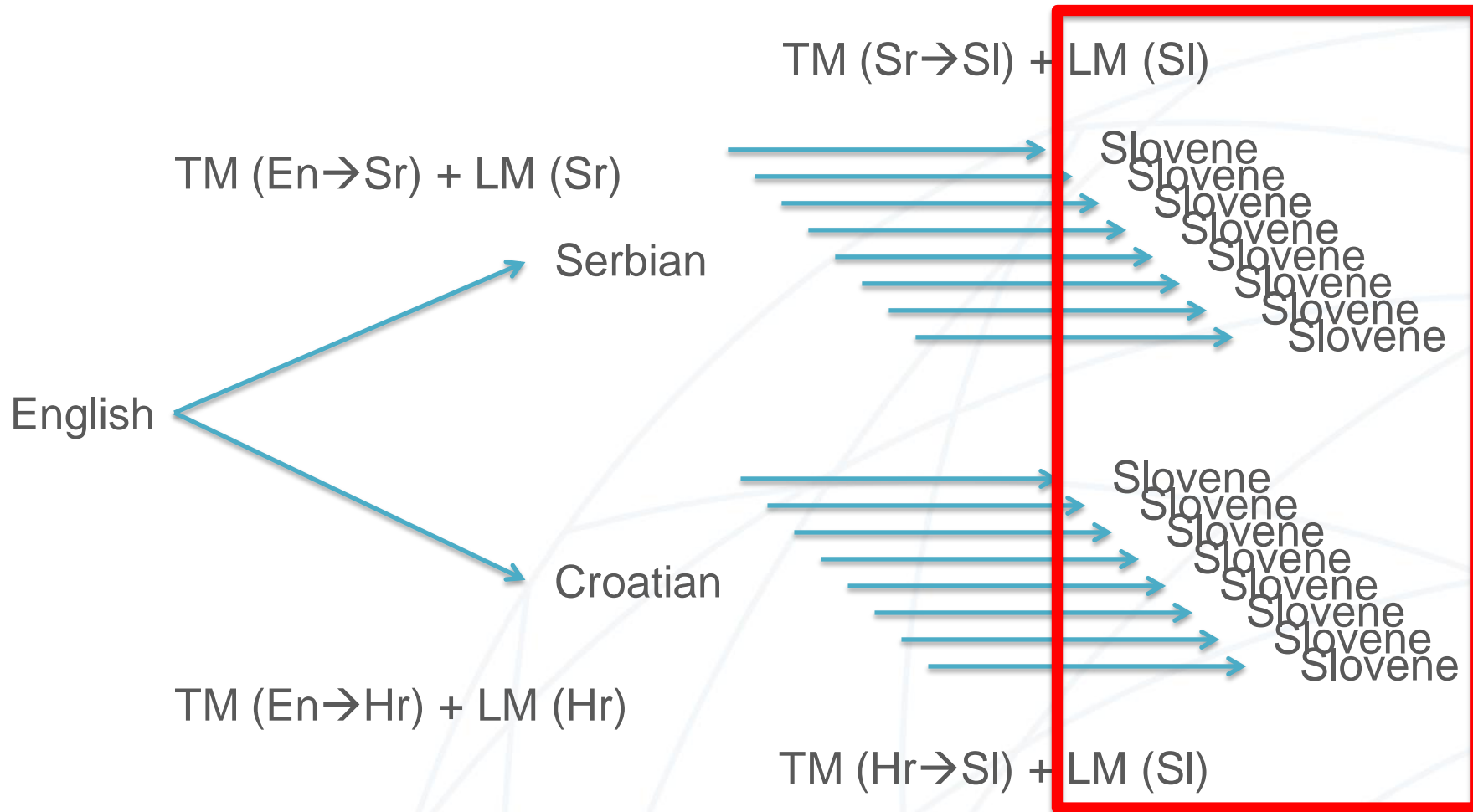
Pivot Translation: source language → pivot language → target language



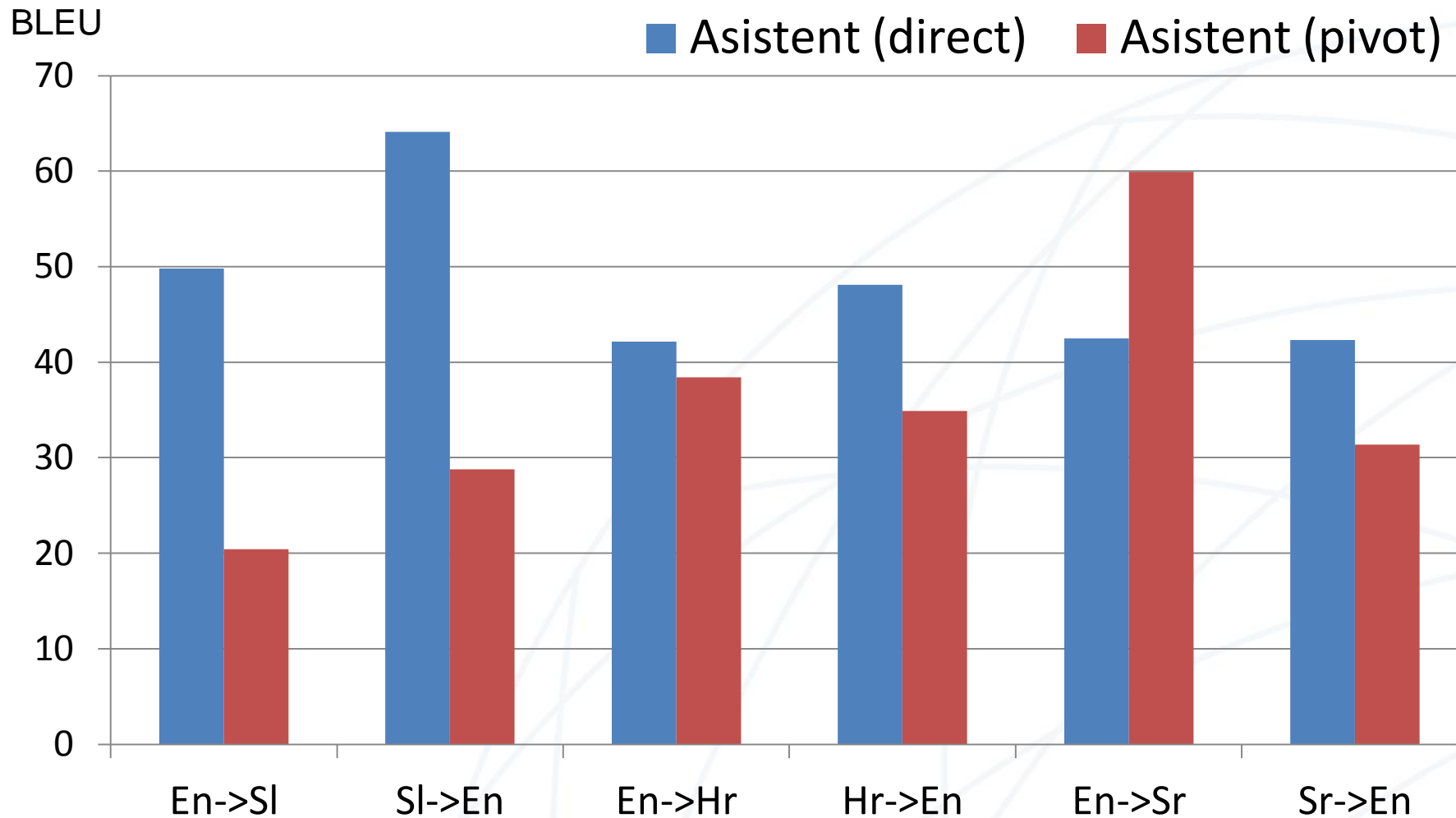
Pivot Translation



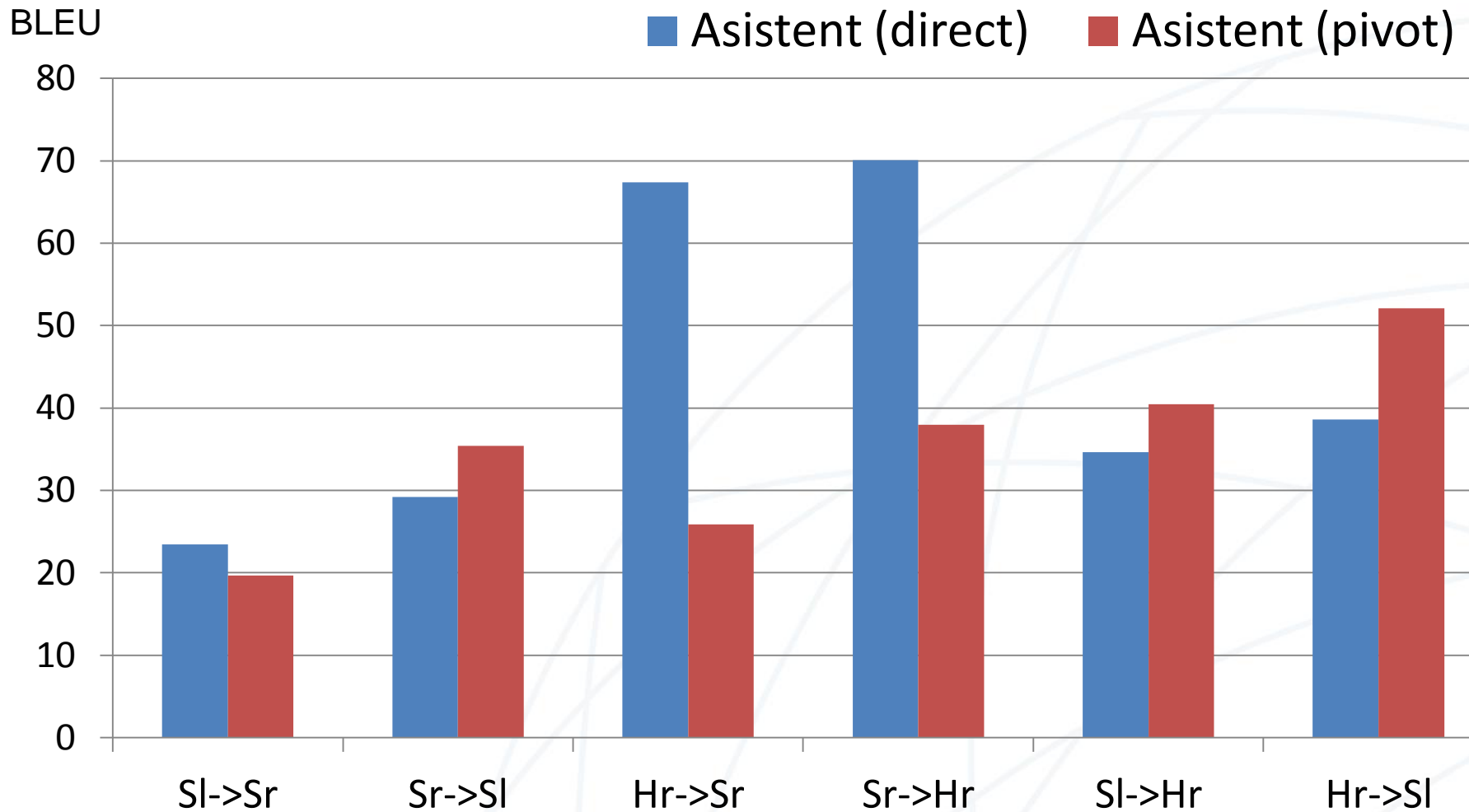
Pivot Translation



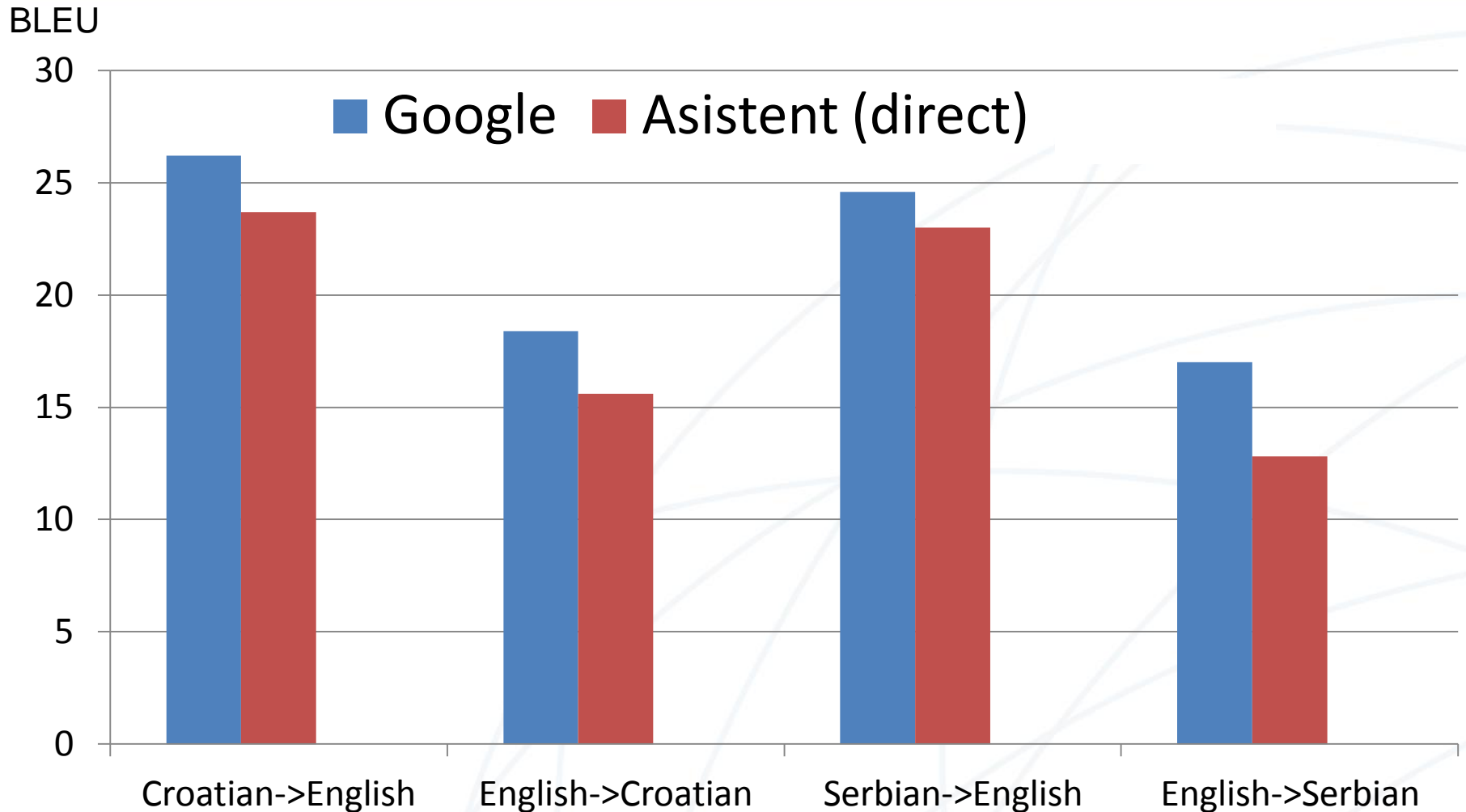
Evaluation for Pivot Translation, In-Domain (English-Slavic)



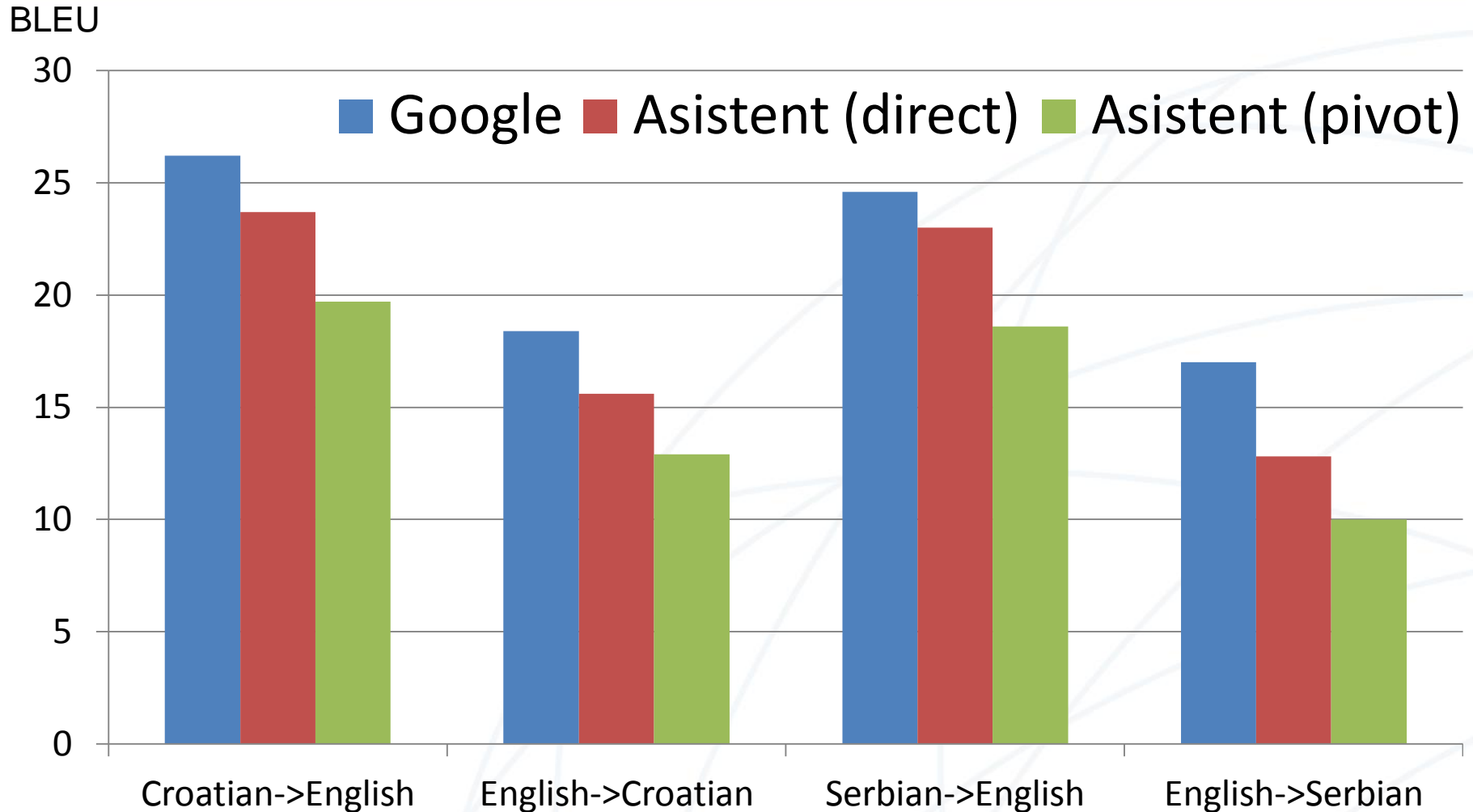
Evaluation for Pivot Translation, In-Domain (Slavic-Slavic)



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



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



Add text to be translated and specify the translation direction:

English Example

Slovene Example

Croatian Example

Serbian Example

Translate

This service is brought to you by
<http://nlp.insight-centre.org/>

ASISTENT (or `assistant@en`) was developed to translate text between English and the morphological complex south Slavic languages: Slovene, Serbian and Croatian.

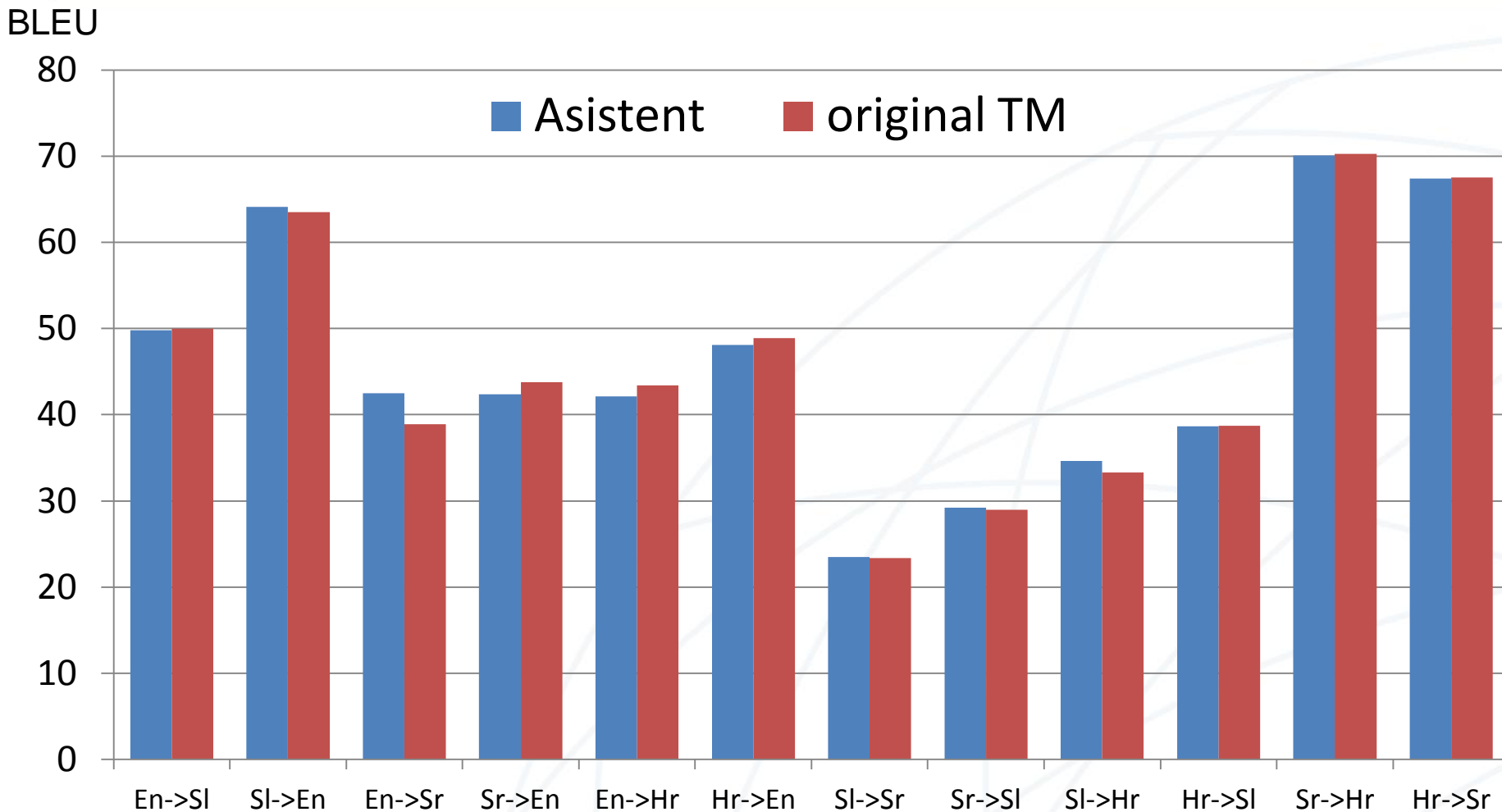
Select translation table option/approach:

- phrase based factored hierarchical
 direct translation pivot translation

Please choose the translation direction:

- English → Slovene Slovene → English
 English → Croatian Croatian → English
 English → Serbian Serbian → English
 Slovene → Croatian Croatian → Slovene
 Serbian → Slovene Slovene → Serbian
 Croatian → Serbian Serbian → Croatian

Filtered vs Original Translation Models Insight



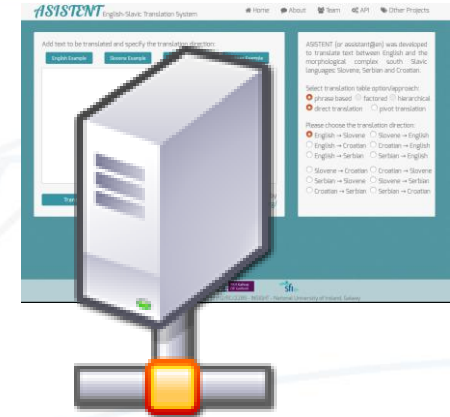
Asistent's API request



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 . "
    }
  ],
}
```

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)

