АБУМАТРАН
ABUMATRAN

# Abu-MaTran

Automatic Building of Machine Translation

PIAP- GA-2012-324414

# D5.1c Evaluation of the MT systems deployed in the third development cycle

| | |
|---|---|
| **Dissemination level** | Public |
| **Delivery date** | 2015/12/31 |
| **Status and version** | Final, v1.0 |
| **Authors and affiliation** | Tommi Pirinen (DCU), Raphael Rubino (Prompsit), Víctor Sánchez-Cartagena (Prompsit), Filip Klubička (UZ) and Antonio Toral (DCU) |

Project funded by the European Community under the Seventh Framework Programme for Research and Technological Development

# Contents

# Executive Summary

This deliverable reports on the evaluation of the machine translation systems described in deliverable D4.1c, which have been developed between the second milestone (month 24) and the third milestone (month 36) of the project. Substantial improvements over the results of milestone 2 are observed for generic English→Croatian and Croatian→English systems. Furthermore, in the human evaluation we show that our current system results in translations of equivalent quality to those of the best performing third party system evaluated.

# 1 Introduction

This deliverable covers the evaluation of the machine translation (MT) systems built during the third development cycle of the project (as described in Deliverable D4.1c). All the experiments concern the language pair English–Croatian, in both directions. This deliverable is laid out as follows. We first detail the evaluation metrics and the experiments carried out in the remainder of this section and then consider each of the experiments performed separately for automatic evaluation measures (in Section 2) and subsequently for human evaluation (in Section 3).

## 1.1 Evaluation Metrics

Unless noted otherwise, each system is evaluated with two state-of-the-art automatic evaluation metrics: BLEU and TER. These metrics were introduced in Deliverable 5.1a (Toral et al., 2013, section 2).

On top of the automatic evaluation metrics, as a novelty in this milestone, most experiments are evaluated also manually. This human evaluation consists of ranking MT outputs with the Appraise[1] tool. For each experiment 100 segments were ranked. All the annotations were carried out by one person, who has native Croatian and advanced English.

Sentences are provided in randomised order to the evaluator. This is to avoid the evaluator (i) evaluating the same sentences again and again (i.e. in the different evaluation experiments conducted) and (ii) to augment the variety of the sentences evaluated for each experiment. The test set (from WMT13, as in the previous milestone) contains a set of news stories in sequential order. If they were provided in order, the evaluator would be evaluating sentences belonging to a small number of news stories, which may not be representative of the whole set. Conversely, when randomised, the evaluator will evaluate sentences coming from different newstories. Though the sentences are evaluated randomly, the evaluator is still shown the local context for each sentence (the previous and the next sentences).

The following guidelines were provided to the annotator:

> Given the translations by more than two MT systems, the task is to rank them:

[1] `https://github.com/cfedermann/Appraise`

- Rank a system A higher (rank1) than B (rank2), if the output of the first is better than the output of the second.
- Rank a system A higher (rank1) than B (rank2), if the output of the first is better than the output of the second.
- Use the highest rank possible, e.g. if you've three systems A, B and C, and the quality of A and B is equivalent and both are better than C, then do: A=rank1, B=rank1, C=rank2. Do NOT use lower rankings, e.g.: A=rank2, B=rank2, C=rank4.

In addition to these guidelines, if two MT outputs have different errors, the annotator was told that the one that conveys better the meaning of the source would be preferred, i.e. in the cases of "ties" of translation quality, then adequacy is preferred over fluency.

The annotator was also asked to take note of the translations when one MT output was strikingly better or worse than other(s). This will be used as a starting point in milestone 4 to try to derive conclusions on the impact brought by our different systems, which might then guide the development of milestone 4 systems.

Figure 1 provides a snapshot of Appraise for an evaluation task, in which the MT output of different systems for the language direction English-to-Croatian are being ranked in terms of translation quality.

## 1.2 Experiments

Here we provide the list of experiments performed during the current milestone. In each of them we concentrate on evaluating a different aspect of the MT system. The motivation for our sequence of experiments is to assess the performance of different approaches, settings and/or systems.

The list of experiments conducted is as follows:

- **Experiment 1**. Comparing different development sets for tuning the weights of the different components of the statistical MT system. In this milestone we have generated multiple translations for a development set (cf. Section 4 of D3.1c) relying on professional and amateur (crowdsourced) translators. In this experiment we aim to measure the impact of using either.

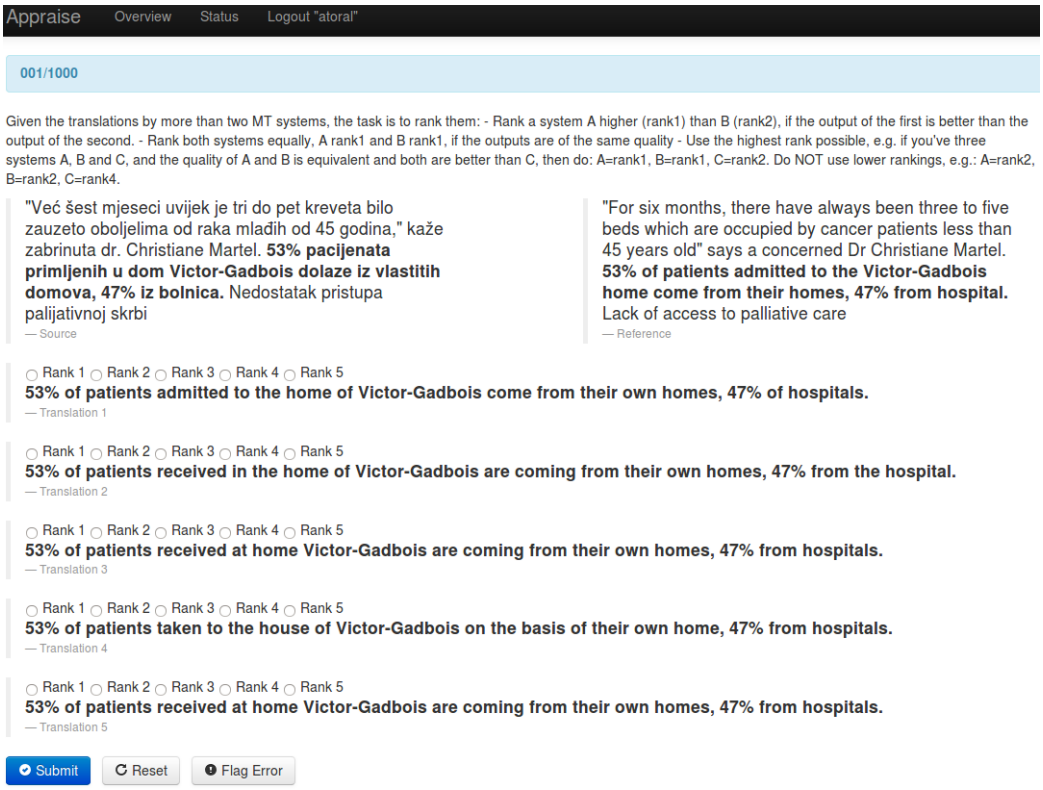


Figure 1: Snapshot of the Appraise tool, used for human evaluation of MT outputs for the Croatian-to-English language direction.

- **Experiment 2**. Comparing the use of three reordering models: word-based (the default in the Moses MT toolkit), phrase-based (Koehn et al., 2005) and hierarchical (Galley and Manning, 2008).
- **Experiment 3**. Measuring the impact of using additional models recently introduced in the statistical MT pipeline. Specifically, we consider the operation sequence model (OSM) (Durrani et al., 2011) and bilingual neural language models (BiNLM) (Devlin et al., 2014a).
- **Experiment 4**: Measuring the impact of adding monolingual data from a closely related language, Serbian.
- **Experiment 5**. Assessing the performance brought by using linguistic knowledge. We experiment with systems that incorporate morphological (morph segmentation, and factored models) and syntactic (syntax-based MT) knowledge.
- **Experiment 6**. Assessing the impact of data selection techniques.
- **Experiment 7**. Comparing our best systems to commercial alternatives.

# 2 Automatic Evaluation

In this section we report the automatic evaluation scores for each of the experiments conducted in milestone 3. The section is organised as follows: the results are divided into experiments as described in Section 1, i.e. from experiments 1 in subsection 2.1 to experiment 7 in subsection 2.6, and finally the resulting final milestone 3 system is evaluated in subsection 2.7.

## 2.1 Experiment 1: Tuning Sets

For our English tuning test, we have generated three different translations into Croatian (cf. Section 4 of D3.1c), two of them produced using crowdsourcing methods, and one carried out by professional translators. This experiment aimed to select the best tuning set(s) from these three translations.

The results of all the different combinations are listed in the Table 1. Into Croatian we have systems tuned on each of the crowdsourced translations (rows *crowd1* and *crowd2*), the professional translation (*prof*), and

then tuning sets using multiple references, be them both the crowdsourced translations (*crowd1+2*) or all the three translations (*prof+crowd1+2*). Into English we have the same configurations. It should be noted though, that due to the fact that we have one single reference into English, tuning sets that use multiple Croatian references are built in this direction by means of concatenation.

This needs to be explained better: more than 1 source with the same reference translation?

| Tuning Set | Direction | dev crowd1 | dev crowd2 | dev pro | test BLEU | test TER |
|---|---|---|---|---|---|---|
| **none** | en→hr | 0.2309 | 0.2098 | 0.2444 | 0.2125 | 0.6458 |
| **crowd1** | | **0.2614** | 0.2331 | 0.2692 | 0.2362 | 0.6250 |
| **crowd2** | | 0.2595 | **0.2358** | 0.2660 | 0.2344 | **0.6242** |
| **crowd1+2** | | 0.2596 | 0.2347 | 0.2691 | 0.2348 | 0.6287 |
| **professional** | | 0.2449 | 0.2213 | 0.2721 | 0.2273 | 0.6457 |
| **3 references** | | 0.2562 | 0.2326 | **0.2731** | **0.2363** | 0.6303 |
| **none** | hr→en | 0.2819 | 0.269 | 0.2946 | 0.3013 | 0.5494 |
| **crowd1** | | **0.3307** | 0.3106 | 0.3230 | 0.3322 | 0.5353 |
| **crowd2** | | 0.3304 | 0.3155 | 0.3220 | 0.3335 | 0.5336 |
| **crowd1+2** | | 0.3294 | **0.3166** | 0.3232 | 0.3352 | 0.5326 |
| **professional** | | 0.3241 | 0.3043 | **0.3398** | **0.3392** | **0.5211** |
| **3 references** | | 0.3302 | 0.3148 | 0.3300 | 0.3386 | 0.5264 |

Table 1: Automatic evaluation of translation systems using different tuning sets. Best results in bold.

Translating from English to Croatian, the results reveal a small advantage when using all three translations, which was the selected tuning set to be used in the rest of the experiments as well as the final system. For Croatian to English, however, using only professional translation yields the best results for the test set (with the use of the three references coming in a close second position), and for that reason in further experiments we keep this as the tuning set for this translation direction.

## 2.2 Experiment 2: Re-Ordering Models

In this experiment we assess the performance of using additional reordering models. As baseline we use the best system from experiment 1, which uses a word-based reordering model, which is the default in Moses. The two additional models used are phrase-based and hierarchical. Both the word- and the phrase-based models use three different orientations (monotone, swap and discontinuous) while the hierarchical model uses four orientations (non merged discontinuous left and right orientations). All the three models are trained bidirectionally. The results are shown in Table 2.

| Reordering | Direction | dev crowd1 | dev crowd2 | dev pro | test BLEU | test TER |
|---|---|---|---|---|---|---|
| **word-based** | en→hr | 0.2562 | 0.2326 | 0.2731 | 0.2363 | 0.6303 |
| **3 models** | | **0.2593** | **0.2339** | **0.2784** | 0.2355 | 0.6336 |
| **word-based** | hr→en | 0.3241 | 0.3043 | 0.3398 | 0.3392 | 0.5211 |
| **3 models** | | **0.3268** | **0.3074** | **0.3411** | **0.3404** | **0.5202** |

Table 2: Automatic evaluation of translation systems with reordering models

Overall, using the three reordering models results in better scores in terms of the automatic metrics. They are therefore employed by our systems in the remaining experiments.

## 2.3 Experiment 3: Additional Components

In this experiment we aim to measure the impact of adding additional models recently introduced in the statistical MT framework. In the specific, we consider the Operation Sequence Model (OSM) (Durrani et al., 2011) and the Bilingual Neural LM (BiNLM) (Devlin et al., 2014b). Both models are trained on the same parallel data used to learn the translation and reordering models. The results are given in Table 3, where we apply both additional models to both translation directions separately and also jointly.

| LM | Direction | dev crowd1 | dev crowd2 | dev pro | test BLEU | test TER |
|---|---|---|---|---|---|---|
| **None** | en→hr | 0.2593 | 0.2339 | 0.2784 | 0.2355 | 0.6336 |
| **OSM** | | 0.2669 | 0.2396 | 0.2849 | 0.2408 | 0.6265 |
| **BiNLN** | | 0.2659 | 0.2404 | 0.2856 | 0.2379 | 0.6310 |
| **OSM+BiNLN** | | **0.2706** | **0.2443** | **0.2881** | **0.2457** | **0.6198** |
| **None** | en→hr | 0.3268 | 0.3074 | 0.3411 | 0.3404 | 0.5202 |
| **OSM** | | 0.3298 | 0.3094 | 0.3472 | 0.3460 | **0.5088** |
| **BiNLN** | | 0.3304 | 0.3136 | 0.3461 | 0.3471 | 0.5138 |
| **OSM+BiNLN** | | **0.3337** | **0.3162** | **0.3535** | **0.3499** | 0.5090 |

Table 3: Automatic evaluation of translation systems using additional language models

The results shows notable improvement when using both methods together, which is the selected configuration for the rest of the experiments.

## 2.4 Experiment 4: Data from a Related Language

In this experiment we test whether using data from Serbian, a closely related language to Croatian, is beneficial as-is, or if it is better to translate the

Serbian data intro Croatian using a rule-based machine translation system. Details on the set-up can be found in Section 4 of Pirinen et al. (2016). The results are shown in Table 4. We only consider the English to Croatian direction in this experiment.

| Serbian data | Direction | dev crowd1 | dev crowd2 | dev pro | test BLEU | test TER |
|---|---|---|---|---|---|---|
| **as is** | en→hr | 0.2851 | **0.2562** | 0.3018 | 0.2486 | 0.6148 |
| **MT'ed** | | 0.2851 | 0.2554 | **0.3039** | **0.2494** | **0.6144** |

Table 4: Automatic evaluation of translation with additional Serbian data

The results show very small differences between using the Serbian data as is or machine-translated ("MT'ed") into Croatian with a rule-based system. As pointed out in Deliverable D4.1c (Pirinen et al., 2016), the amount of Serbian data is rather small compared to the size of Croatian data. Therefore, whichever way we add the Serbian data, it is expected that the impact would be limited.

## 2.5 Experiment 5: Linguistic Processing

In experiment 5 we compared a number of linguistic pre-processing schemes and approaches to the SMT pipeline, using a number of different underlying systems. The two methods compared in Table 5 are morphological segmentation (cf. Section 3.1 of Deliverable D4.1c (Pirinen et al., 2016)) methods and factored models (cf. Section 3.2 of Deliverable D4.1c).

For morphological segmentation we compare three methods: Morfessor Baseline 2.0 (Morfessor in the table), Morfessor Flatcat (Flatcat in the table) and morphs induced from the rule-based morphology of the Apertium rule-based MT system (Apertium in the table). For factored models we examine two systems for generating POS-tagged data: Hunpos and CRF Suite.

In the results we can observe some clear winners. For systems based on factored models, CRF results in better scores than Hunpos. Regarding morph segmentation, unsupervised methods beat supervised ones in general for English to Croatian while the opposite is true for the Croatian→English direction.

Following on with linguistically motivated approaches, we also built syntax-based MT systems (cf. Section 3.3 in D4.1c), whose results are shown in Table 6. We tested empirically the best performing relaxation parameter (i.e. right-relaxation) on our syntactic parsers. This resulted in a small gain over the equivalent system without relaxation (e.g. 0.2799 vs 0.2724 for the

| Model | Direction | test BLEU | test TER |
|---|---|---|---|
| **Morfessor** | | **0.2368** | **0.6439** |
| **Flatcat** | | 0.234 | 0.6448 |
| **Apertium** | en→hr | 0.2192 | 0.6480 |
| **Hunpos** | | 0.2332 | **0.6292** |
| **CRF** | | **0.2361** | 0.6303 |
| **Morfessor** | | 0.3237 | 0.5576 |
| **Flatcat** | hr→en | 0.3143 | 0.5617 |
| **Apertium** | | **0.3342** | **0.5442** |

Table 5: Automatic evaluation of linguistically augmented systems (morph segmentation and factored models).

MST parser in the Croatian-to-English direction). However, the best results are obtained by the unsupervised hierarchical system (rows "hierarchical"), which being unsupervised does not include any external linguistic information at all.

| System | Direction | relaxation | test BLEU | test TER |
|---|---|---|---|---|
| hierarchical | en→hr | none | **0.2412** | **0.6136** |
| Berkeley (tree-to-string) | | rightbin | 0.1183 | 0.7531 |
| MST Parser (string-to-tree) | | none | 0.2388 | 0.6231 |
| | | rightbin | 0.2396 | 0.6234 |
| hierarchical | hr→en | none | **0.3502** | **0.5381** |
| MST parser (tree-to-string) | | none | 0.2724 | 0.6126 |
| | | rightbin | 0.2799 | 0.6011 |
| Berkeley (string-to-tree) | | rightbin | 0.2692 | 0.6261 |

Table 6: Automatic evaluation of linguistically augmented systems (syntax-based).

## 2.6 Experiment 6: Data selection for Translation Models

In this experiment we assess different ways of selecting and combining parallel data. The aim is to find the setup that leads to the highest score.

In terms of data selection, we have a pool of 7 parallel training corpora available (cf. Section 2.1 in D4.1c (Pirinen et al., 2016)). We concatenate the 7 corpora and rank their parallel sentences in terms of bilingual cross-entropy, by scoring the sentences against language models (perplexity) that are in-domain with respect to the tuning set and other language models that are out-of-domain. For in-domain language models we use SETimes, as this corpus regards news, the same domain as our tuning and test sets. For out-of-domain, we pick a random subset of the 7 corpora whose size in number of tokens is that of SETimes (4,834,182 for English and 4,503,428 for Croatian).

Once the data is ranked we split it in two subsets: the highest ranked 25% (top) and the remaining 75% (bottom). On top of this, we experiment with the vocabulary saturation filter to reduce the size of the latter subset (bottom_vsf).

Besides data selection, we also experiment with different ways of combining the parallel training data when there is more than one corpus. We try two approaches: concatenation and linear interpolation. Because the latter performs better, we use this approach in our experiments.

| Dataset | Direction | dev crowd1 | dev crowd2 | dev pro | test BLEU | TER |
|---|---|---|---|---|---|---|
| **top** | en→hr | **0.2751** | 0.2477 | 0.2842 | 0.2487 | **0.6115** |
| **bottom** | | 0.2322 | 0.2115 | 0.2337 | 0.2168 | 0.6467 |
| **top bottom** | | 0.2722 | **0.2489** | **0.2856** | 0.248 | 0.612 |
| **top bottom_vsf** | | 0.2736 | 0.2469 | 0.283 | **0.2497** | 0.6118 |
| **7 phrase tables** | | 0.263 | 0.2369 | 0.2787 | 0.2445 | 0.6147 |
| **top** | hr→en | **0.3527** | **0.3453** | 0.3507 | 0.3636 | 0.4942 |
| **bottom** | | 0.3063 | 0.2921 | 0.3111 | 0.3172 | 0.5348 |
| **top bottom** | | 0.2763 | 0.2718 | 0.2877 | 0.2974 | 0.5731 |
| **top bottom_vsf** | | 0.3439 | 0.3387 | **0.3626** | 0.3646 | **0.4906** |
| **7 phrase tables** | | 0.3501 | 0.3367 | 0.3616 | **0.3721** | **0.4878** |

Table 7: Automatic evaluation of translation with data selection

The results appear rather mixed and there is no clear winner among the different systems evaluated. That said, if one considers a trade-off between translation quality and parallel data size, then systems "top" (trained on 25% of the sentence pairs) and "top bottom_vsf" (trained on 41% of the sentence pairs) seem to be the best choices. In the next experiments we use system "top", with reordering models, OSM and BiNLM components trained also on the "top" parallel dataset.

### 2.7 Experiment 7: Final Milestone 3 System

The final milestone 3 systems are composed of the best systems of the experiments in previous subsections. The results are given in table 8. We evaluate two of our systems against the commercial online $3^{rd}$ party systems[2] by Yandex, Bing and Google.

| System | Direction | BLEU | TER |
|---|---|---|---|
| *Google* | en→hr | 0.2673 | 0.5946 |
| *Bing* | | 0.2281 | 0.6263 |
| *Yandex* | | 0.2030 | 0.6801 |
| **Abu-MaTran milestone 3** | | 0.2544 | 0.6081 |
| *Google* | hr→en | 0.4099 | 0.4635 |
| *Bing* | | 0.3658 | 0.5199 |
| *Yandex* | | 0.3463 | 0.5311 |
| **Abu-MaTran milestone 3** | | 0.3852 | 0.4819 |

Table 8: Automatic evaluation of the final milestone 3 system

Overall, we can say our system obtains a better BLEU score than those of Yandex's and Bing's, but slightly worse than Google's.

## 3 Human Evaluation

The human evaluation involves asking human judges to rank the MT outputs produced for a sentence by anonymised systems. From those rankings we then derive a human score for each system with the TrueSkill method adapted to MT evaluation (Sakaguchi et al., 2014) following its usage at WMT15.[3] Namely, we run 1,000 iterations of rankings followed by clustering ($p = 0.05$). If two systems are placed in different clusters (column "range" in results' tables) then the one with lower range is considered significantly better.

In the following subsections we provide the human evaluation for the experiments carried out. The order of the experiments is the same as in the previous section for the automatic evaluation. Human evaluation was carried out for all the experiments but two: using languages for a related

[2]Translations of the test set were obtained with these third-party systems as of December 22nd 2015

[3]`https://github.com/mjpost/wmt15`

language (experiment 4) and data selection (experiment 6). This is due to the limitations already discussed in the automatic evaluation for experiment 4, and to the very similar scores across systems obtained in experiment 6.

## 3.1 Experiment 1: Tuning Sets

| # | Direction | Score | Range | System |
|---|---|---|---|---|
| 1 | | 0.298 | 1–4 | professional |
| 2 | | 0.082 | 1–4 | crowd2 |
| 3 | en→hr | -0.073 | 2–5 | 3 references |
| 4 | | -0.082 | 1–5 | crowd1 |
| 5 | | -0.225 | 3–5 | crowd1+2 |
| 1 | | 0.271 | 1–3 | crowd1+2 |
| 2 | | 0.087 | 1–5 | crowd1 |
| 3 | hr→en | 0.086 | 1–4 | crowd2 |
| 4 | | -0.184 | 1–5 | 3 references |
| 5 | | -0.260 | 3–5 | professional |

Table 9: Human ranking for experiment 1

The results are shown in Table 9. Similarly to what we observed for the automatic evaluation, the results are mixed also for the human evaluation. This may indicate that the impact of the development set in our setup is too small to be of importance.

## 3.2 Experiment 2: Reordering Models

The results are shown in Table 10. For English to Croatian a single word-based model results in a higher score according to the human evaluation, while the opposite is true for the Croatian to English direction. However none of these differences is significant (note that all the systems are in the same range, 1–2).

## 3.3 Experiment 3: Additional Components

The results are shown in Table 11. Using both components jointly results in the highest score according to the human evaluation, as it happened also for

| # | Direction | Score | Range | System |
|---|---|---|---|---|
| 1 | en→hr | 0.092 | 1–2 | word-based |
| 2 | | 0.092 | 1–2 | 3 models |
| 1 | hr→en | 0.091 | 1–2 | 3 models |
| 2 | | -0.091 | 1–2 | word-based |

Table 10: Human ranking for experiment 2

the automatic evaluation. In addition, it is worth noting that this system is significantly better than the baseline (i.e. using no additional components) as the systems are placed in different ranges: 1–2 for both and 3–4 for none for both translation directions.

| # | Direction | Score | Range | System |
|---|---|---|---|---|
| 1 | en→hr | 0.409 | 1–2 | Both |
| 2 | | 0.393 | 1–2 | OSM |
| 2 | | -0.321 | 3–4 | None |
| 2 | | -0.481 | 3–4 | BiNLM |
| 1 | hr→en | 0.432 | 1–2 | Both |
| 2 | | 0.339 | 1–3 | OSM |
| 3 | | -0.187 | 2–4 | BiNLM |
| 4 | | -0.584 | 3–4 | None |

Table 11: Human ranking for experiment 3

## 3.4 Experiment 5: Linguistic Processing

The results are shown in Table 12. Flatcat (unsupervised morph segmentation) and Apertium (rule-based morph segmentation) come out on top for translations into Croatian and into English, respectively, with their improvements being significant over the other systems that add linguistic knowledge. The fact that these 2 systems are the ones that come out on the bottom for the opposite direction seems to indicate that segmentation is sensitive to the translation direction.

| # | Direction | Score | Range | System |
|---|---|---|---|---|
| 1 | en→hr | 0.566 | 1–2 | Flatcat |
| 2 | | 0.209 | 1–4 | None |
| 3 | | 0.12 | 2–5 | Morfessor |
| 4 | | 0.073 | 2–5 | CRF |
| 5 | | -0.23 | 3–5 | Hunpos |
| 6 | | -0.738 | 6–6 | Apertium |
| 1 | hr→en | 0.788 | 1–1 | Apertium |
| 2 | | -0.131 | 2–4 | None |
| 3 | | -0.283 | 2–4 | Morfessor |
| 4 | | -0.373 | 2–4 | Flatcat |

Table 12: Human ranking for experiment 5

## 3.5 Experiment 7: Final Milestone 3 System

The results are shown in Table 13. As in the automatic evaluation, Google comes on top for both directions, but in the human evaluation it is not significantly better than our system, which on its turn is significantly better than both Microsoft and Yandex for English-to-Croatian, and than Yandex only for Croatian-to-English.

| # | Direction | Score | Range | System |
|---|---|---|---|---|
| 1 | en→hr | 0.961 | 1–2 | Google |
| 2 | | 0.506 | 1–2 | Abu-MaTran |
| 4 | | -0.664 | 3–4 | Bing |
| 5 | | -0.884 | 3–4 | Yandex |
| 1 | hr→en | 0.696 | 1–2 | Google |
| 2 | | 0.220 | 1–3 | Abu-MaTran |
| 4 | | 0.104 | 2–3 | Bing |
| 5 | | -1.246 | 4–4 | Yandex |

Table 13: Human ranking for the final milestone 3 systems

# 4 Conclusions

This deliverable has reported the evaluation of the MT systems developed during the period of the third milestone of the project (M24–M36). Our generic MT system for milestone 3 outperforms the generic system we built for milestone 2 for both translation directions (English to Croatian and vice versa). Furthermore, in the human evaluation we have showed that our system results in translations of equivalent quality to those of the best performing third party system evaluated.

In the next development cycle, we will go beyond overall system evaluation by carrying out an error analysis of our current best performing system.

# References


Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., and Makhoul, J. (2014a). Fast and Robust Neural Network Joint Models for Statistical Machine Translation. In *Proceedings of ACL*, pages 1370–1380.

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., and Makhoul, J. (2014b). Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1370–1380, Baltimore, Maryland. Association for Computational Linguistics.

Durrani, N., Schmid, H., and Fraser, A. (2011). A Joint Sequence Translation Model with Integrated Reordering. In *Proceedings of ACL/HLT*, pages 1045–1054.

Galley, M. and Manning, C. D. (2008). A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 848–856. Association for Computational Linguistics.

Koehn, P., Axelrod, A., Birch, A., Callison-Burch, C., Osborne, M., Talbot, D., and White, M. (2005). Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceesings of the International Workshop on Spoken Language Translation*, pages 68–75.

Pirinen, T., Rubino, R., Cartagena, V. M. S., and Toral, A. (2016). Abumatran deliverable d4.1b mt systems for the third development cycle. Technical report.

Sakaguchi, K., Post, M., and Van Durme, B. (2014). Efficient Elicitation of Annotations for Human Evaluation of Machine Translation. In *Proceedings of WMT*, pages 1–11.

Toral, A., Cortés-Vaíllo, S., Ramírez-Sánchez, G., Klubička, F., and Ljubešić, N. (2013). Abu-matran deliverable D5.1a: Evaluation for the first development cycle. Technical report.