Original scientific paper

Evaluation of Free Online Machine Translations for Croatian-English and English-Croatian Language Pairs

Sanja Seljan
Department of Information Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
sseljan@ffzg.hr

Marija Brkić
Department of Informatics, University of Rijeka
Omladinska 14, Rijeka, Croatia
mbrkic@uniri.hr

Vlasta Kučiš
Department of Translation Studies,
Faculty of Arts, University of Maribor
Koroška cesta 160, Maribor, Slovenia
asta.kucis@siol.net

Summary

This paper presents a study on the evaluation of texts from four domains (city description, law, football, monitors) translated from Croatian into English by four free online translation services (Google Translate, Stars21, InterTran and Translation Guide) and texts translated from English into Croatian by Google Translate. The aim of the paper is to conduct a machine translation evaluation of free translation services and to measure inter-rater agreement and the influence of error types on the criteria of fluency and adequacy. The evaluation is performed by students and the results are analyzed by Fleiss kappa and Pearson's correlation.

Key words: evaluation, free online translation services, English, Croatian, fluency, adequacy, error analysis, Fleiss, Pearson

Introduction

Although human-assisted machine translation (MT) still gains significantly better results than automatic translation (Bar-Hillel, 2003), the use of online translation tools has increased in recent years, even among less widely spoken languages. Despite the fact that it is desirable to have moderate to good quality

translations, there are cases where high quality translations are not of crucial importance (e.g. gist of a paragraph from a Web page, e-mail translation, basic information about a conference, a product, etc.). Evaluation from the user's perspective helps producers, but also examines the translation problems. However, most of the evaluation has been conducted only for widely spoken languages, which possess various language tools and resources.

The increase in the use of free online translation tools has caused an increased interest in the evaluation of these tools. The recent studies have raised questions regarding possible uses of online MT. Besides gisting, MT can be used in information retrieval, i.e. question-answering systems (Garcia-Santiago and Olvera-Lobo, 2010). As presented by this pilot research, free online MT tools are also used for homework translation, where, apart from pedagogical use, it is important to detect the inappropriate use (Hampshire and Porta Salvia, 2010).

In 1976 Systran launched its first MT system for the Commission of the European Communities. The first online free translation on the Internet appeared in 1997 by Babel Fish using Systran technology (Aiken et al, 2009). In 2007 Google Translate online translator appeared, relying more on the statistical approach and the comparison of matching probabilities, than on the rule-based approach. Ever since, it has been included in almost every evaluation study.

According to Kit and Wong (2008), misunderstanding of MT can be avoided by having realistic expectations and using appropriate text genre. Zervaki (2002) points out that in the case of simple sentences and SVO order, MT can produce acceptable terminology and syntax. However, in more complex sentences translations become incomprehensible. Online translation services mostly serve popular languages and there is a considerable difference in the quality of translation dependent on the language pair and the type of text being translated.

MT evaluation has been used not only for evaluation of different commercial or online systems, but also during system development.

This pilot research performs evaluation of free online translation services for less widely spoken languages, such as Croatian, and measures inter-rater agreement. The human evaluation, based on the criteria of fluency and adequacy, is enriched by error analysis, in order to examine the influence of error types on fluency and adequacy and to use it in further research.

Due to a small test set, the results should be taken as preliminary. After the related work section, the test set and evaluation procedure description is given. In addition to the evaluation results, error analysis is given and Pearson’s correlation and Fleiss kappa results are presented.

Related work

DARPA (Defense Advanced Research Projects Agency) has presented the outcome of the research on the evaluation of various MT systems, based on the black box methodology, as presented in White and O’Connell (1996). In order to avoid the subjectivity of human judgments, it has been suggested to

decompose judgments into adequacy, informativeness and fluency (White et al, 1995).

The study described in Garcia-Santiago and Olvera-Lobo (2010) analyzes the effectiveness of the translations from German and French into Spanish obtained through Google Translate, ProMT and WorldLingo. They emphasize the benefits that would be gained when research studies would use the same scale of human assessment.

In the study performed by Dis Brandt (2011) three popular online tools have been evaluated (Google Translate, Tungutorg, Apertium). Despite the fact that the number of errors significantly decreases after human editing, some tools show significantly better results.

In the study conducted by Aiken et al. (2006) random Spanish sentences from two introductory Spanish textbooks and two web sites have been translated into English by Systran and evaluated by undergraduate students.

In the study presented by Kit and Wong (2008), Babel Fish, Google Translate, ProMT, SDL free translator, Systran and WorldLingo have been evaluated using BLEU and NIST for translating from 13 languages into English. Systran and Babel Fish have proved to be the best for the majority of language pairs, while Google Translate has proved the best for translation from Arabic and Chinese into English. ProMT has proved better than the rest for Portuguese-English and WorldLingo for Swedish-English language pair.

The evaluation of MT systems is of high importance in MT research and product design. The evaluation is done either to measure system performance (Giménez and Màrquez, 2010; Lavie and Agarwal, 2007) or to identify weak points and/or to adjust parameter settings of different MT systems or of a single system through different phases (Denkowski and Lavie, 2010a; Agarwal and Lavie, 2008).

The evaluation of MT is done using different language independent algorithms, mostly BLEU (Papineni et al., 2002) or NIST (Doddington, 2002). In order to obtain metrics that give results closer to human evaluation results, there is a need for qualitative evaluation of different linguistic phenomena integrated with statistical approaches (Monti et al., 2011). In the study presented by (Denkowski and Lavie, 2010b) the factors that constitute "good" or "bad" translations are discussed, with the focus on difficult points of inter-rater agreement.

Experimental study

The evaluation of free online translation services has been performed from the user's perspective, by undergraduate and graduate students of languages, linguistics and information sciences, who have attended one or more courses on language technologies at the University of Zagreb, Faculty of Humanities and Social Sciences. The texts from four domains (city description, law, football,

monitors), have been translated from Croatian into English and from English into Croatian.

Translations from Croatian into English have been obtained from four Internet translation services with Croatian language support:

Google Translate (GT) - http://translate.google.com

Stars21 (S21) - http://stars21.com/translator

InterTran (IT) - http://transdict.com/translators/intertran.html

Translation Guide (TG) - http://www.translation-guide.com

GT is a web-based translation service provided by Google Inc. It is a statistical MT based on huge amount of corpora. It currently supports 57 languages. The Croatian language has been supported since 2008. Although S21 service is powered by GT, the translations are not always the same, probably due to different pre- or postprocessing techniques. Translation Experts Ltd. Company provides MT services using IT, which is powered by NeuroTran and WordTran, which translate sentence-by-sentence and word-by-word, respectively. NeuroTran is a hybrid system that uses a combination of linguistic rules, statistical methods and neural networks, as well as text analysis, to determine the context for the unique type of lexical selection. Although TG is powered by IT, the resulting translations differ. While Croatian-English translations have been obtained from four mentioned services, the English-Croatian translations have been obtained only from GT.

Test set description

The students have evaluated four illustrative excerpts translated from Croatian into English from the domains of city description, law, football and monitors, with 9, 9, 7, 9 sentences, respectively.

Table 1. Text statistics

		Total no. of words	Min sentence length	Max sentence length	Avg sentence length
En-Cro	city	160	11	27	17.7
	law	227	5	43	25.2
	football	160	3	32	17.7
	monitors	112	3	28	12.4
Cro-En	city	119	5	22	13.2
	law	196	9	46	21.7
	football	133	9	29	19.0
	monitors	123	3	22	15.4

Table 1 presents the average sentence length per domain and translation direction. The text about the city and the legal text are the same for both directions (indicating that English texts are on average 20-25% longer), while

the text on football and the text on LCD monitor differ from one direction to another.

Evaluation

The evaluation of Croatian-English (Cro-En) translations has been performed by 48 students (64.6% undergraduate, 35.4% graduate), who are either finalizing their studies of the English language and information sciences, or who have been learning English for about 10-12 years. 75% of the students have attended one or more language technology courses, such as Machine Translation, Translation Memories as Translation Tools or Computer-Assisted Language Learning. 83% of the students have declared to have had previous experience in translating, out of which 20% in professional translation, 52.5% for the faculty needs, and 27.5% for personal needs.

According to the answers, 39.6% of students use online translation tools to translate texts in the domain of technology, 31.3% in the domain of travelling (cities, countries), 8.3% for acquiring information on conferences, 20.8% for translating e-mails, 69.5% for translating domain specific texts, and 22.5% for translating texts related to the Internet services, games, wiki articles, literature and for homework translation tasks.

The English-Croatian (En-Cro) translations have been evaluated by 50 native speakers – students of languages, linguistics and information sciences (58% undergraduate, 42% graduate).

Tools and resources

Fig. 1 presents the average grade assigned to Croatian language tools and resources on the Internet and language resources irrespective of the language, i.e. online dictionaries, terminology databases, glossaries, translation memories and translation tools. This grading has been made prior to the pilot research and is based solely on previous experience. The average grade for the Croatian language tools and resources is 3.00, and for language tools and resources in general 3.57.

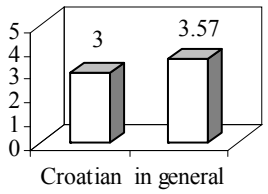

Fig. 1. Average grades for free language resources on the Internet

Fig. 2 presents the average grades for Croatian language tools and resources (online dictionaries, terminology databases, translation memories and GT). The

average grade is 2.90, being very close to the general perception of the quality of available tools and resources for Croatian (3.00) given in Fig. 1. The best grade is given to GT (3.54). Fig. 3 presents the average grades assigned to the selected online translation services. The average grade for all four services is 3.325, being close to the average grade 3.57 assigned to free online tools and resources presented in Fig. 1.

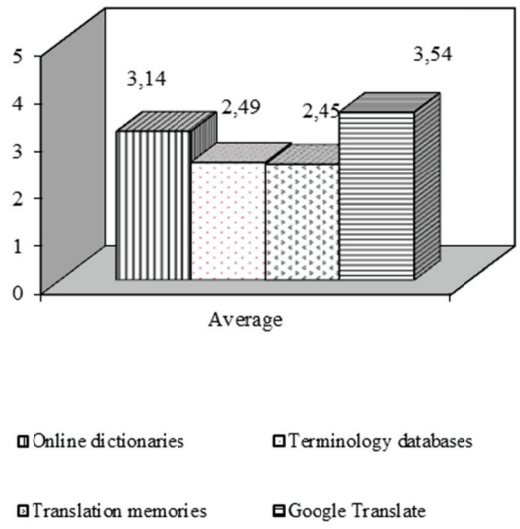

Fig. 2. Average grades for online free Croatian language tools and resources

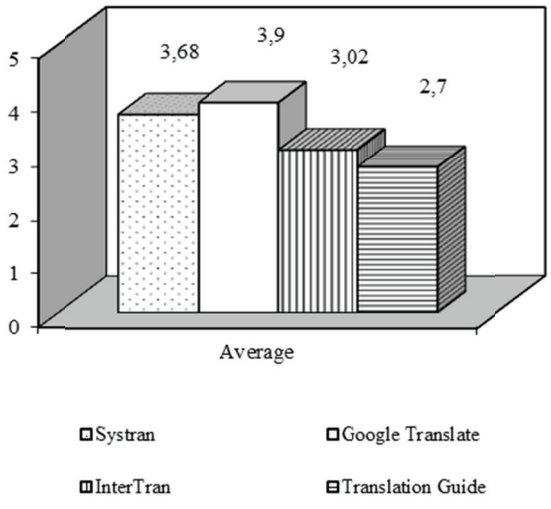

Fig. 3. Average grades for four elected online translation services

Among the interviewed students, 90% of them would like to use the following Croatian tools and resources of the appropriate quality: online dictionaries (90%), MT systems (78%), translation memories (38%), terminology databases (28%), glossaries (24%) and speech-to-text systems (14%) (Fig. 4).

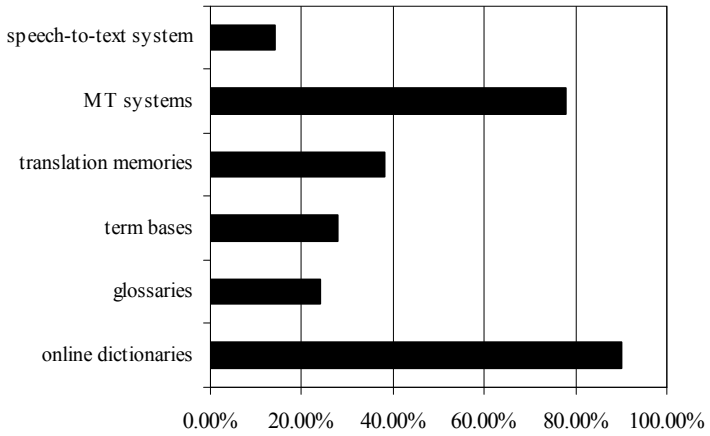

Fig. 4. Desirable resources of the appropriate quality

Results

The evaluation has been made according to the criteria of fluency (indicating how much the translation is fluent in the target language) and adequacy (indicating how much of the information is adequately transmitted). The Cro-En translations have been obtained from four translation services available for the Croatian language (GT, S21, IT, TG), while En-Cro translations have been obtained from GT.

Table 2 presents average grades per each domain in both translation directions (the average of fluency and adequacy).

Table 2. Average grades in Cro-En / En-Cro translations

	Cro-En					**Eng -Cro**
	GT	**S21**	**IT**	**TG**	**All**	**GT**
city	4.33	4.43	2.18	1.23	3.04	4.71
law	4.63	4.78	2.21	1.13	3.19	4.26
football	4.84	4.72	1.75	1.16	3.12	3.75
monitors	4.72	4.72	1.87	1.15	3.12	4.50
Average	**4.62**	**4.66**	**2.02**	**1.17**	**3.04**	**4.29**

Fig. 5 is a graphical representation of Cro-En translation grades presented in Table 2.

Cro-En translation have been given either low grades (TG and IT) or high grades (S21 and GT), in comparison to the average values. S21 service has obtained slightly better overall average grade than GT (4.66 versus 4.62), better grade in the domain of city description (4.43 versus 4.33), as well as in the legal domain (4.78 versus 4.63). GT has outperformed S21 in the domain of football (4.84 versus 4.72). The average grade for the domain of monitors (4.72) is equal for both services. The services IT and TG have obtained below the average grades in all domains. The best average results for all the services have been obtained in the domain of law (3.19), followed by the domains of monitors and football (3.12), while the lowest grades have been assigned to the city description (3.04), which has the most freedom in style.

Fig. 6 graphically represents grades assigned to En-Cro and Cro-En translations obtained from GT.
En-Cro translations have lower average results than the reverse direction (4.29 versus 4.62), especially in the football domain (3.75 versus 4.84), but also in the domain of law (4.26 versus 4.63) and monitors (4.50 versus 4.72). En-Cro translation direction has higher average grade only in the city description domain which contains shorter sentences, mostly nominative constructions and uses frequent phrase constructions. The most evident difference in the quality of the translation between En-Cro and Cro-En directions is in the text on football, which contains domain specific terms and mostly non-nominative constructions. This causes different types of errors.

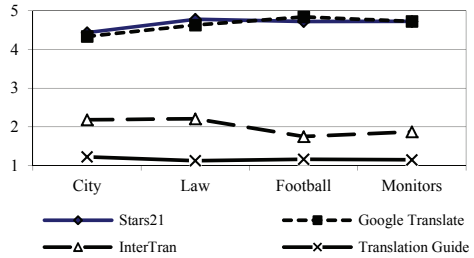

Fig. 5. Average scores for 4 free online translation services for Cro-En

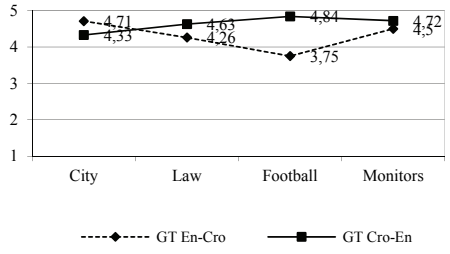

Fig. 6. Average scores for GT in En-Cro and Cro-En directions

Error analysis

The evaluation has been enriched by the analysis of translation errors, which have been divided into several categories:

- morphological errors, i.e. mistakes regarding wrong inflection/endings
- untranslated words
- lexical errors, i.e. wrong word choices, incorrect translations, typing errors, translations of words when not necessary, mistakes that stem from wrong word choices or forms of lexical entries, word omissions

- syntactic errors, i.e. incorrect word order or incorrect punctuation

The following excerpts present Cro-En translations in the domain of football, where there is a significant difference between translation directions, and in the domain of monitors, where the results are similar in both translation directions.

Cro-En

Football

Original: Ovaj Pravilnik utvrđuje globalna i obvezujuća pravila u svezi statusa igrača, njihovo pravo nastupa za sudjelovanje u organiziranom nogometu i njihove transfere između klubova koji pripadaju različitim savezima.

GT: These Regulations establish global and binding rules concerning the status of players eligible to play for their participation in organized football and their transfer between clubs belonging to different federations.

S21: These Regulations establish a global and binding rules concerning the status of players, their right to participate to participate in organized football and their transfer between clubs belonging to different federations.

IT: Ovaj Pravilnik in band their is performing, the law for taking part to organized football which are belonging and their the different coalitions the transfers between clubs.

TG: Ovaj Legality utvrđuje fine plus obvezujuća Pravda into a union fit on playing njihovo pravo intervention for coagency into a organized football plus their transfere između klubova who appurtenant različitim savezima.

Monitors

Original: Pročitajte do kraja ove upute prije nego počnete koristiti ovaj uređaj.

GT: Read these instructions before using this device.

S21: Read these instructions before using this device.

IT: Read our in full ove instructions before počnete are you using this unit.

TG: Pročitajte in full ove instructions before počnete are you using this uređaj.

Translations offered by GT and S21 are very similar, although not identical. GT uses the following phrase *...players eligible to play for their participation* and S21 the phrase *...players, their right to participate to participate in*. IT and TG services have not correctly translated isolated words, such as *Regulation*, but have left the source language word *Pravilnik* or have used the incorrect translation *Legality*; the word *federations* has been translated as *coalitions* or left in the source language *savezima*. Instead of the phrase *establish global and binding rules*, the IT has offered *in band their is performing, the law*. TG has produced the translation *utvrđuje fine plus obvezujuća Pravda* with some of the words untranslated (*utvrđuje, obvezujuća)*. TG translations are completely incomprehensible and inadequate.

In the domain of monitors, GT and S21 have produced the same output and obtained the highest grades. IT service has produced *before počnete are you using* instead of *before using*. Not only that part of the phrase has not been translated (*počnete*), but there are also syntactic errors. Translation by TG is even worse, because everyday words such as *read, these, device* have not been translated.

Although the translations offered by TG are powered by IT, they differ in the number of untranslated words, i.e. TG does not recognize words with diacritics.

The following excerpts present En-Cro translations in all four domains performed by GT.

En-Cro
City description
Original: It lies on the intersection of important routes between the Adriatic coast and Central Europe.
GT: Nalazi se na sjecištu važnih prometnica između jadranske obale i srednje Europe.
Law
Original: Pursuant to Article 88 of the Constitution of the Republic of Croatia, I hereby issue this DECISION ON THE PROCLAMATION OF THE ELECTRONIC SIGNATURE ACT
GT: Na temelju članka 88. Ustava Republike Hrvatska, donosim ovaj ODLUKA O proglašenju Zakona o elektroničkom potpisu
Football
Original: Registration Period: a period fixed by the relevant Association in accordance with Article 6.
GT: Registracija razdoblje: razdoblje koje utvrdi relevantne Udruge u skladu s člankom 6.
Monitors
Original: Read these instructions completely before using the equipment.
GT: Pročitajte ove upute prije korištenja u potpunosti opreme.

The sentence on the city description has been correctly translated. In the law domain, morphological errors have been made (*na temelju Ustava Republike Hrvatsk<u>a</u>*), i.e. the genitive form should have been used (*Hrvatsk<u>e</u>*). In the phrase *donosim ov<u>aj</u> ODLUK<u>A</u>*, the transitive verb should take the direct object in accusative (*donosim ov<u>u</u> ODLUK<u>U</u>*). There is also a mismatch in gender. The phrase *Registration Period* in the text on football might have been translated as a multiword unit (*Registracij<u>sko</u> razdoblje*), in order to avoid a morphological error (*Registracij<u>a</u> razdoblje*). Subject-verb gender agreement errors also fall into the category of morphological errors (*razdoblje koje utvrd<u>i</u> relevantn<u>e</u> Udruge)* instead of (*razdoblje koje utvrđ<u>uje</u> relevantn<u>a</u> Udrug<u>a</u>* or *razdoblje koje <u>je</u> utvrdil<u>a</u> relevantn<u>a</u> Udrug<u>a</u>*). The verb *fixed* could have been translated in the present (*utvrđuje*) or in the past (*je utvrdila*). In any case, there is either a morphological error or the omission of the auxiliary verb. In the translation of the text on monitors, the translation of *completely* as *u potpunosti* is correct, but stands in the wrong place, which is an example of a syntactic error.
The error analysis in En-Cro translations has shown the highest number of lexical errors, including also errors in style (average 2.44), followed by untranslated words (1.83), morphological (1.75) and syntactic errors (1.38).
The highest number of errors has been found in the text from the football domain (mostly lexical errors and untranslated words). That text has also gained the lowest average grade (3.75).
The best score has gained the translation from the city description domain, although it has the second highest number of all error types. In the city description domain prevail lexical errors.

The lowest number of errors has been found in the legal domain, where all types of errors are evenly distributed.
Morphological errors have been mostly found in the domain of monitors, which has, despite of this, gained the second highest average grade (4.5). The smallest number of morphological errors has been found in the city description domain.
Untranslated words are by far mostly found in the football domain which has gained the lowest average grade.
The dominant value in En-Cro translations, i.e. the number of errors according to the majority, is 1 morphological error in the domains of city description and monitors, while the dominant value is 3 in the legal and football domains. The dominant value with regard to the lexical errors in the city description is 1 and between 1 and 3 in other domains. The dominant value with regard to the number of untranslated words is 1 in all domains. The dominant value with regard to the syntactic errors is 1 in all domains but football, where it is evenly distributed between 1, 2, and 3.

Pearson's correlation
The Pearson's correlation between the number of errors and the average grade is negative, indicating that smaller number of errors augments the average grade, which is mostly reflected in the correlation between untranslated words and average grades.
The correlation between errors types and the criteria of fluency and adequacy has shown that the criterion of fluency is more affected by the increase of lexical and syntactic errors, while adequacy is more affected by untranslated words.
Fleiss' kappa
Fleiss' kappa has been used for assessing the reliability of agreement among raters when giving ratings to the sentences (1). It indicates the extent to which the observed amount of agreement among raters exceeds what would be expected if all the raters made their ratings completely randomly.

$$\overline{P} = \frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2 - Nn\right) \qquad (1)$$

The score is between 0 and 1. A value of 1 implies perfect agreement while values less than 1 imply less than perfect agreement. Interpretation of values used in this case is as follows: < 0 poor agreement, 0.0-0.20 slight agreement, 0.21-0.40 fair agreement, 0.41-0.60 moderate agreement, 0.61-0.80 substantial agreement, and 0.81-1.00 almost perfect agreement.
There is a relatively high level of the agreement among raters per domain and per system in Cro-En translations, as given in Table 3. It varies from moderate

(mainly for IT translation service), through substantial agreement (S21 and GT), up to almost perfect agreement (TG).

Table 3. The level of agreement per domain and per system for Cro-En services

Cro-En	city	law	football	monitors
S21	0.61	0.65	0.73	0.70
GT	0.73	0.58	0.61	0.70
TG	0.94	0.99	0.94	0.94
IT	0.52	0.34	0.52	0.51

Table 4. The level of agreement per domain with regard to the criteria of fluency and accuracy for En-Cro translations by GT

EN-CRO	fluency	adequacy
city description	0.58	0.67
monitors	0.53	0.59
law	0.40	0.49
football	0.35	0.37

Table 4 presents the inter-rater agreement for En-Cro translations. The lowest level of agreement has been detected in the domains of football and law, which contain longer and more complex sentences. For the domain of football fair agreement has been detected. The agreement is moderate for the domains of law and monitors and substantial for the city description domain. The level of inter-rater agreement is lower for En-Cro translations in all domains.

Conclusion

This paper presents an evaluation study of machine translations from four domains. Cro-En translations have been obtained from four free online translation services. En-Cro translations have been obtained only from GT. The results regarding the use of freely available language resources indicate that there is a high interest in their use for the Croatian language.

Fleiss kappa shows substantial, even perfect agreement in the evaluation of four translation services. It shows almost perfect agreement in the ranking of TG as the worst translation service. Substantial agreement is achieved for S21 and GT services, which have gained the highest grades. Moderate agreement is shown for IT, which has performed slightly better than TG.

In Cro-En translations the average evaluation results for S21 and GT range from 4.63 to 4.84 for the domains of football, law and monitors. The average grade for the city description in Cro-En translation is lower than in the opposite direction due to more freedom in style.

In En-Cro direction, the translations have been performed by GT and have obtained lower grades than in the opposite direction. This is true for all but the city description domain, which contains mostly nominative constructions, frequent words, and no domain specific terms.

Error analysis shows that the translation grades are mostly influenced by untranslated words (especially the criteria of adequacy), while morphological and syntactic errors reflect grades in smaller proportion.
GT service, which has been used in both translation directions, harvesting data from the Web, seems to be well trained and suitable for the translation of frequent expressions. However, it does not perform well where language information is needed, e.g. gender agreement. The use of background terminology database of multiword expressions and/or translation memory database, would probably improve results, especially translations of specific terms and idiomatic expressions.
Further research in a specific domain would enable a more detailed analysis of specific language phenomena and error types and would enable identifying language-specific problems in automatic MT.

Acknowledgments

This work has been supported by the Ministry of Science, Education and Sports of the Republic of Croatia, under the grant 130-1300646-0909.

References

Agarwal, A.; Lavie, A. METEOR, M-BLEU and M-TER: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output. // *Proceedings of the ACL 2008 Workshop on Statistical Machine Translation*. 2008, pp. 115-118

Aiken, M.; Ghosh, K.; Wee, J.; Vanjani, M. An Evaluation of the Accuracy of Online Translation System. // *Technology, Communication of the IIMA*. 2009.

Aiken, M.; Vanjani, M. B.; Wong, Z. Measuring the Accuracy of Spanish to English Translations. // *Issues in Information Systems*. 7 (2006), 2, pp. 125-128.

Amancio, D. R.; Nunes, M. G. V.; Oliveira Jr., O. N.; Pardo, T. A. S.; Antiqueira, L.; Costa, L. da F. Using Metrics from Complex Networks to Evaluate MT. // *Physica A*. 390 (2011), 1, pp. 131-142.

Bar-Hillel, Y. The Present Status of Automatic Translation of Languages. // *Readings in Machine Translation*. Boston: MIT Press, 2003, pp. 45-77.

Denkowski, M.; Lavie, A. Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level. // *Proceedings of NAACL/HLT*. 2010, pp. 250-253.

Denkowski, M.; Lavie, A. Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks. // *Proceedings of AMTA*. 2010.

Dis Brandt, M. Developing an Icelandic to English Shallow Transfer Machine Translation System, Ms. Thesis. Reykjavik University, 2011.

Doddington, G. Automatic Evaluation of Machine Translation Quality Using Ngram Co-occurrence Statistics. // *Proceedings of the 2nd International Conference on Human Language Technology Research*. 2002, pp. 138–145.

Kit, C.; Wong, T. M. Comparative Evaluation of Online Machine Translation Systems with Legal Texts. // *Law Library Journal*. Vol. 100 (2008), 2, pp. 299–321.

Garcia-Santiago, L.; Olvera-Lobo, M. D. Automatic Web Translators as Part of a Multilingual Question-Answering (QA) System: Translation of Questions. // *Translation Journal*. Vol. 14 (2010), 1.

Giménez, J.; Màrquez, L. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-) Evaluation. // *The Prague Bulletin of Mathematical Linguistics*. 2010, 94, pp. 77–86.

Hampshire, S.; Porta Salvia, C. Translation and the Internet : Evaluating the Quality of Free Online Machine Translators. // *Quaderns: revista de traducció*. 2010, 17, pp. 197-209.

Lavie, A.; Agarwal, A. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. // *Proceedings of the ACL 2007 Workshop on Statistical Machine Translation*. 2007, pp. 228-231.

Monti, J.; Barreiro, A.; Elia, A.; Marano, F.; Napoli, A. Taking on new challenges in multi-word unit processing for machine translation. // *Proceedings of the 2nd Workshop on Free/Open-Source Rule-Based Machine Translation*. 2011, pp. 11-19.

Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation. // *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. 2002, pp. 311–318.

Somers, H.; Gaspari, F.; Niño, A. Detecting Inappropriate Use of Free Online Machine Translation by Language Students - A Special Case of Plagiarism Detection. // *Proceedings of the 11th Annual Conf. of the European Association for Machine Translation*. 2006, pp. 41-48.

White, J. S.; O’Connell, T. A. Adaptation of the DARPA Machine Translation Evaluation Paradigm to End-to-End Systems. // *Proceedings of AMTA-96*. 1996.

White, J. S.; O'Connell, T.; O'Mara, F. 1995. Evaluation Methodologies in the ARPA Machine Translation Initiative. // *Proceedings of AIPA95*. Automatic Information Processing Association Steering Group, 1995.

Zervaki, T. Online Free Translation Services. // *Proceedings of the 24th Int. Conf. on Translating and the Computer.* London: Aslib, 2002.