The VoiceTRAN Speech Translation Demonstrator

Jerneja Žganec Gros[1], Stanislav Gruden[1], France Mihelič[2], Tomaž Erjavec[3], Špela Vintar[4], Peter Holozan[6], Aleš Mihelič[1], Simon Dobrišek[2], Janez Žibert[2], Tomo Korošec[5], Nataša Logar[5]

[1]Alpineon d.o.o., Ljubljana, Slovenia
[2]University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia
[3]Jožef Stefan Institute, Ljubljana, Slovenia
[4]University of Ljubljana, Faculty of Arts, Ljubljana, Slovenia
[5]University of Ljubljana, Faculty of Social Sciences, Ljubljana, Slovenia
[6]Amebis d.o.o., Kamnik, Slovenia

Abstract

This paper describes the design phases of the VoiceTRAN Communicator, which integrates speech recognition, machine translation, and text-to-speech synthesis using the Galaxy architecture. The aim of the work was to build a robust multimodal speech-to-speech translation system able to translate simple domain-specific sentences in the language pair Slovenian-English. The work represents a joint collaboration between several Slovenian research organizations that are active in human language technologies.

Govorni komunikator VoiceTRAN

Prispevek opisuje delo na razvoju govornega komunikatorja VoiceTRAN, ki združuje tehnologije prepoznavanja govora, strojnega prevajanja in sinteze govora. Podajamo opis arhitekture sistema ter posameznih sistemskih modulov. Nadalje opisujemo jezikovne vire, ki smo jih uporabili pri izgradnji sistema, ter preskus sistema. Sistem VoiceTRAN omogoča govorno prevajanje za jezikovni par slovenščina-angleščina na omejenem področju uporabe.

1. Introduction

Automatic speech-to-speech (STS) translation systems aim to facilitate communication among people that speak different languages [1, 2, 3]. Their goal is to generate a speech signal in the target language that conveys the linguistic information contained in the speech signal from the source language.

There are, however, major open research issues that challenge the deployment of natural and unconstrained speech-to-speech translation systems, even for very restricted application domains, due to the fact that state-of-the-art automatic speech recognition and machine translation systems are far from perfect.

In addition, in comparison to translating written text, conversational spoken messages are often conveyed with imperfect syntax and casual spontaneous speech.

In practice, when building demonstration systems, STS systems are typically implemented by imposing strong constraints on the application domain and the type and structure of possible utterances; that is, both in the range and in the scope of the user input allowed at any point of the interaction. Consequently, this compromises the flexibility and naturalness of using the system.

The VoiceTRAN Communicator was developed in a Slovenian research project involving 6 partners: Alpineon, the University of Ljubljana (Faculty of Electrical Engineering, Faculty of Arts, and Faculty of Social Studies), the Jožef Stefan Institute, and Amebis as a subcontractor.

The work has been co-funded by the Slovenian Ministry of Defense and the Slovenian Research Agency.

The aim is to build a robust multimodal speech-to-speech translation communicator, similar to Phraselator [4] or Speechalator [5], able to translate simple sentences in the language pair Slovenian-English. It goes beyond the Phraselator device because it is not limited to predefined input sentences.

In the initial phase of the project a system demonstrator was developed. In further phases it will be wrapped into a stand-alone communicator and upgraded to new language pairs. The application domain is limited to common application scenarios that occur in peace-keeping operations on foreign missions when the users of the system have to communicate with the local population. More complex phrases can be entered via keyboard using a graphical user interface.

First an overview of the VoiceTRAN system architecture is given. We continue to describe the individual server modules. We conclude the paper by discussing the speech-to-speech translation evaluation methods and outlining plans for future work.

2. System Architecture

The VoiceTRAN Communicator uses the DARPA Galaxy Communicator architecture [6]. The Galaxy Communicator open source architecture was chosen to provide inter-module communication support because its plug-and-play approach allows interoperability of commercial software and research software components. It was specially designed for development of voice-driven user interfaces in a multimodal platform.

The VoiceTRAN Communicator consists of a Hub and five servers that interact with each other through the Hub as shown in Figure 1.

The Hub is used as a centralized message router through which servers can communicate with one another. Frames containing keys and values are emitted by each server. They are routed by the hub and received by a secondary server based on rules defined in the Hub script.

Audio Server	Receives speech signals from the microphone and sends them to the recognizer. Sends synthesized speech to the speakers.
Graphic User Interface	Receives input text from the keyboard. Displays recognized source language sentences and translated target language sentences. Provides user controls for handling the application.
Speech Recognizer	Takes the signals from audio server and maps audio samples into text strings. Produces an N-best sentence hypothesis list.
Machine Translator	Receives N-best postprocessed sentence hypotheses from the speech recognition server and translates them from a source language into a target language. Produces a scored disambiguated sentence hypothesis list.
Speech Synthesizer	Receives rich and disambiguated word strings from the machine translation server. Converts the input word strings into speech and prepares them for the audio server.

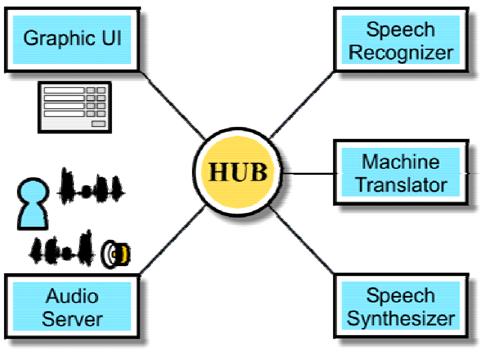

Figure 1. The Galaxy system architecture used in the VoiceTRAN communicator.

2.1. Audio Server

The audio server connects to the microphone input and speaker output terminals on the host computer and performs recoding of user input and playing prompts or synthesized speech.

Input speech captured by the audio server has been automatically recorded to files for posterior system training.

2.2. Speech Recognizer

The speech recognition server receives the input audio stream from the audio server and provides a word graph at its output and a ranked list of candidate sentences; the N-best hypotheses list, which can include part-of-speech information generated by the language model.

The speech recognition server used in VoiceTRAN is based on the Hidden Markov Model Recognizer developed at the Faculty of Electrical Engineering, University of Ljubljana [7]. It has been upgraded to perform medium-size vocabulary (10K words) speaker (in)dependent speech recognition on a wider application domain. A back-off class-based trigram language model is used. Given a limited amount of training data the parameters in the models have been carefully chosen in order to achieve maximum performance.

Because the final goal was a stand-alone speech communicator used by a specific user, the speech recognizer has been additionally trained and adapted to the individual user in order to achieve higher recognition accuracy in at least one language pair direction.

A common speech recognizer output typically has no information on sentence boundaries, punctuation, and capitalization. Therefore, additional postprocessing in terms of punctuation and capitalization has been performed on the N-best hypotheses list before it is passed to the machine translator.

The inclusion of a prosodic module was necessary in order to link the source language to the target language, but also to enhance speech recognition proper. Besides syntactic and semantic information, properties such as dialect, sociolect, sex and attitude etc are signaled by prosody. The degree of linguistic information conveyed by prosody varies between languages, from languages such as English, with a relatively low degree of prosodic disambiguation, via tone-accent languages such as Swedish, to pure tone languages. Prosody information helps to determine proper punctuation and sentence accent information.

2.3. Machine Translator

The machine translator (MT) converts text strings from a source language into text strings in the target language. Its task is difficult since the results of the speech recognizer convey spontaneous speech patterns and are often erroneous or ill-formed.

A postprocessing algorithm inserts basic punctuation and capitalization information before passing the target sentence to the speech synthesizer. The output string can also convey lexical stress information in order reduce disambiguation efforts during text-to-speech synthesis.

A multi-engine based approach was used in the early phase of the project that makes it possible to exploit strengths and weaknesses of different MT technologies and to choose the most appropriate engine or combination of engines for the given task. Four different translation engines have been applied in the system. We combined TM (translation memories), SMT (statistical machine translation), EBMT (example-based machine translation) and RBMT (rule-based machine translation) methods. A simple approach to select the best translation from all the outputs was applied. A bilingual aligned domain-specific corpus was used to build the TM and train the EBMT and the SMT phrase translation models.

The Presis translation system was used as our baseline system [8]. It is a commercial conventional rule-based translation system that is constantly being optimized and upgraded. It was adapted to the application domain by upgrading the lexicon. Based on stored rules, Presis parses each sentence in the source language into grammatical components, such as subject, verb, object and predicate and attributes the relevant semantic categories. Then it uses built-in rules for converting these basic components into the target language, performs regrouping and generates the output sentence in the target language.

We continue the paper by describing the VoiceTRAN SMT experiment.

2.3.1. Statistical Machine Translation Experiment

Some initial machine translation attempts have been reported for the translation from Slovenian into English [8], [9], however, very little has been done for the opposite direction, from English into Slovenian. We have performed experiments in both translation directions, where the latter proved to be an especially complex and demanding task due to the highly inflectional nature of the Slovenian language.

The SMT experiments were performed on a joint corpus, consisting of 3 parallel corpora: the VoiceTRAN application-specific corpus, the SVEZ-IJS [10] and the IJS-ELAN corpus [11], where the words in all three corpora contain automatically assigned context-disambiguated lemmas and morphosyntactic descriptions (MSDs). Sentences longer than 25 words were discarded from the joint corpus.

The freely available GIZA++ tool [12] was used for training the SMT model. The CMU-SLM toolkit [13] was used for generating the language model. The ISI ReWrite Decoder [14] has been applied for the translation of test sentences.

Two different types of test sets were used. The first test set was extracted from the joint corpus. The test sentences were chosen at regular intervals, one out of every 1000 sentences. For the second test we used the sentences from one of the components of the IJS-ELAN corpus, the ORWL file (Orwell's "1984"), which is of a significantly different text type from the rest of the joint corpus.

This set-up enabled us to test the system with sentences, which were highly correlated to the training data, as well as on those that had low correlation to the training set. The test sentences were excluded from the training material for the SMT and language models.

The SMT experiments were performed in two ways. First, we implemented the 'simple' procedure, where the sentences used for training the SMT system were taken directly from the joint corpus, without any prior modifications.

The second or 'combined' procedure was more complex. From the joint corpus we have derived two corpora. In the first corpus, the sentences in both languages have been modified so that the words were replaced by their lemmas, using the lemmatization information provided in the source corpora. In order to derive the second corpus, all original word forms have been replaced by their corresponding morphosyntactic descriptors.

These two corpora were then separately fed to the training system.

The decoding was performed as follows: every test sentence was preprocessed into two sentences, where words had been replaced by lemmas in the first sentence, and by MSDs in the second sentence.

Then we traced how each pair lemma+MSD in the source language changed to the corresponding pair lemma+MSD in the target language. The resulting pair lemma+MSD was ultimately combined to construct the final word in the target language. Our goal was to decrease the data sparsness of the training corpus. That was achieved by translating lemmas instead of original words. By translating MDSs separately, we wanted to preserve the MSD information without affecting the translation of the lemmatized text.

Further improvement was expected by adding a dictionary corpus to the joint corpus in the training phase. We have used only pairs of single words (no multi-word expressions were included at this stage). In cases when one English word had many Slovenian translation equivalents, one entry was added for each of these translations to the dictionary corpus, which ended up by containing approximately 140.000 entries.

We introduced many additional corrections, for which we expected they might improve the translation performance. For example, less important features of the MSDs were replaced by a wild-card character, etc. Further, if the decoding algorithm decided that the MSD gender value of the target word was dual and the dual word form was not found in our word base, plural was used instead, similarly as in [9]. All tokens in the corpus containing numerals were in the initial phase replaced by a unique token, which further reduced the data sparseness. Words marked in English as proper nouns, were handled in a similar way. By tracing word translation, we were able to replace these unique tokens, which appeared in the Slovenian text with the corresponding original token from the English language test sentence. Finally, one additional monolingual annotated text corpus was added to train the Slovenian language modeling tool, the FDV-IJS corpus.

Since the sources of the training data had been automatically tagged with lemmas and MSDs, the resulting imperfections in the training material had negative effects, especially on the combined translation method results. Therefore, we intend to re-tag the source corpora in the continuation of the project.

2.4. Speech Synthesizer

The last part in a speech-to-speech translation task is the conversion of the translated utterance into its spoken equivalent. The input target text sentence is equipped with lexical stress information at possible ambiguous words.

The Proteus unit-selection text-to-speech system is used for this purpose [15]. It performs grapheme-to-phoneme conversion based on rules and a look-up dictionary and rule-based prosody modeling. Domain-specific adaptations include new pronunciation lexica and the construction of a speech corpus of frequently used in-domain phrases.

Special attention was paid to collocations as defined in the bilingual dictionary. They were treated as preferred units in the unit selection algorithm.

We are also exploring how to pass a richer structure from the machine translator to the speech synthesizer. An input structure containing information on POS and lexical stress information resolves many ambiguities and can result in more accurate prosody prediction.

The speech synthesizer produces an audio stream for the utterance. The audio is stream is finally sent to the speakers by the audio server. After the synthesized speech has been transmitted to the user, the audio server is freed up in order to continue listening for the next user utterance.

2.5. Graphical User Interface

In addition to the speech user interface, the VoiceTRAN Communicator provides a simple interactive user-friendly graphical user interface, as shown in Fig. 2. Input text in the source language can also be entered via keyboard or selected by pen input.

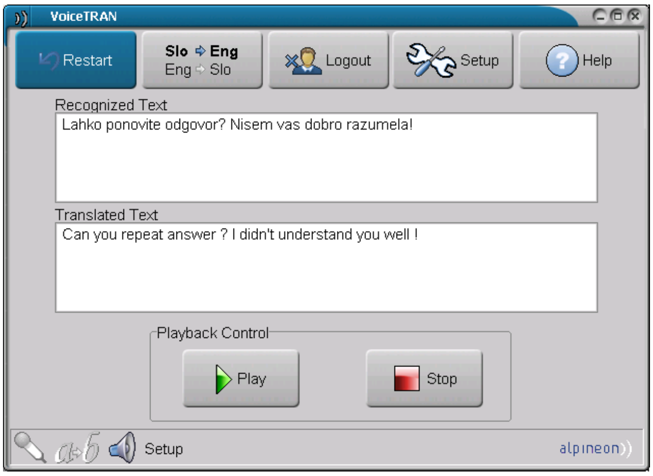

Figure 2. Screenshot of the graphical user interface in the VoiceTRAN communicator application. The source language text provided by the speech recognition module and the translated text in the target language are displayed.

Recognized sentences in the source language along with their translated counterparts in the target language are displayed.

A push-to-talk button is provided to signal an input voice activity, and a replay button serves to start a replay of the synthesized translated utterance. The translation direction can be changed by pressing the translation direction button.

The setup menu enables the user to customize the application according to his needs. It also provides the possibility to choose between different text-to-speech engines.

3. Language Resources

Some of the multilingual language resources needed to set up STTS systems and include Slovenian are presented in [16]. For building the speech components of the VoiceTRAN system, existing speech corpora have been used [17]. Since the speech corpora have been collected from different sources, adaptations have been carried out. The language model has been trained on a domain-specific text corpus that was collected and annotated within the project.

The Proteus pronunciation lexicon [15] has been used for both speech recognition and text-to-speech synthesis. Speech synthesis is based on the Proteus speech corpus. It has been expanded by the most frequent in-domain utterances.

As mentioned in section 2.3., for developing the machine translation component, a dictionary of military terminology [18], and various existing aligned parallel corpora were used [10], [11]. We have syntactically annotated an in-domain large Slovenian monolingual text corpus, the FDV-IJS that was collected at the Faculty of Social Studies, University of Ljubljana. This corpus contains over 5.5 million words and has been used for training the language model in the speech recognizer, as well as for inducing relevant multiword units (collocations, phrases, and terms) for the domain.

An aligned bi-lingual in-domain corpus with 300,000 words – the VoiceTRAN corpus – has been collected within the project. The compilation of the corpus involved selecting the digital original of the bi-texts, re-coding to XML TEI P4, sentence alignment, word-level syntactic tagging, and lemmatization [19]. The corpus has been used to induce bi-lingual single word and phrase lexica for the MT component, and as direct input for SMT and EBMT systems. It was also used for training of the speech recognizer language model.

4. Evaluation

The evaluation tests of a speech-to-speech translation system serve two purposes:

1. to evaluate whether we have improved the system by introducing improvement of individual components of the system;
2. to test the system acceptance by the end users in field tests.

We have performed individual component tests in order to select the most appropriate methods for each application server. Speech recognition was evaluated by computing standard word error rates, which were below 10%. The TTS system was evaluated using ITU-T recommendations for subjective performance tests. The results are reported in [15].

		Relative changes in the average values of the MT metrics with a tested system configuration in comparison to the baseline system configuration			
		Δ WER [%]	Δ GTM [%]	Δ NIST [%]	Δ BLEU [%]
1)	Tested configuration: combined MT method Baseline system: simple MT method Test sentences: from the joint corpus	+3 to +5	-20 to -8	-10 to -5	-25 to -10
2)	Tested configuration: combined MT method Baseline system: simple MT method Test sentences: ORWL corpus	-3 to -2	-5 to +6 [1] -8 to -1 [2]	-2 to 0	+20 to +80
3)	Tested configurations: various additional corrections from 2.3.1 Baseline system: simple/combined MT method without additional corrections Test sentences: joint corpus, ORWL	-2 to 0	+5 to +10	+5 to +10	+25 to 200

Table 1: Evaluation results for the SMT system. Relative changes in the average values of the MT metrics with a tested system configuration in comparison to the baseline system configuration are given. In experiments 1) and 2), the range of percentages reflects the following variations in system configuration: with or without the dictionary, with or without the additional corrections from 2.3.1. In experiment 3), the range of percentages reflects the following variations in system configuration: with or without dictionary, combined method/simple translation method, different test sets.

For the machine translation component, initial objective evaluation tests were performed, which we describe in the following subsection.

4.1. SMT Evaluation Results

To measure the 'closeness' between the SMT-generated hypothesis and human reference translations, standard objective MT metrics were used: Word Error Rate (WER), General Text Matcher (GTM) [20], NIST and BLEU [21].

The SMT evaluation efforts were centered on three system variation impacts:

1) the impact of the choice of the translation method, i.e. simple or combined,
2) the impact of the addition of a dictionary, and
3) the impact of the combination of the additional corrections, described by the end of chapter 2.3.1.

In Table 1, relative changes in evaluation scores (WER, GTM, NIST and BLEU) of the tested MT system and training set configuration versus the baseline system are given. The obtained values for the BLEU score were so small that the obtained results have not been considered as reliable.

In comparison to the simple translation method, the combined translation method did not perform well for test sentences extracted from the unprocessed joint corpus.

The combined method performed better when ORWL test sentences were used, proving its potential for translation of out-of-domain sentences.

Surprisingly, in all cases, the NIST score was slightly better for the simple translation method.

The simple translation method apparently adapted well to inflected Slovenian words, some of which were frequent enough in the training material to allow for sufficient training of the statistical model. As a consequence, when testing on test sentences from the joint corpus, which were well correlated to the training corpus, the test set translations were translated rather well. As expected, the combined translation method performed better when translating texts, which were very different from the training sentence set, as was the case with the ORWL test corpus.

For every test configuration we found that the addition of a dictionary had a minor and more or less random influence on the translation quality. The dictionary contained many entries, which translated one English word to more than one Slovenian word candidates, which proved to be an obstacle in the training process. One of these words usually dominated and the system too often picked it as a result.

From the other corrections, introduced in 2.3.1, only the special treatment of numeric tokens and proper nouns has yielded a better performance, whereas the addition of the IJS-FDV corpus to the language model has not.

The scores for translation quality using the standard metrics were generally low. We would like to stress that we found that these evaluation methods are not suitable for evaluating translations into Slovenian. These tools are all based on an exact comparison of entire words, which works well for English. Due to the rich inflectional paradigms in Slovenian, words, which are semantically correctly translated, but their ending is wrong, have the calculated score of zero. A method, which attributes score points for finding a correct word stem would provide a much better translation quality estimation. Nevertheless, the used evaluation methods were suitable for the purposes of our research since we were only interested in an indicator for improvement or deterioration when using various MT system and training set configurations.

---

[1] without dictionary
[2] with dictionary

5. Conclusion

The implementation concept of the VoiceTRAN communicator demonstrator has been discussed in the paper. It is able to translate simple domain-specific sentences in the language pair Slovenian-English.

The chosen system architecture makes it possible to test a variety of server modules. The end-to-end prototype was evaluated in and is ready for end-user field trials.

6. Acknowledgements

The work presented in this paper was supported by the Slovenian Ministry of Defense and the Slovenian Research Agency under contract no. M2-0019.

7. References

[1] A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavaldà, T. Zeppenfeld, and P. Zhan, "Janus-III: Speech-to-Speech Translation in Multiple Languages," Proceedings of the ICASSP, Munich, Germany, 1997, pp. 99–102.

[2] W. Wahlster, Verbmobil: Foundation of Speech-to-Speech translation, Springer Verlag, 2000.

[3] A. Lavie, F. Metze, R. Cattoni, E. Costantin, S. Burger, D. Gates, C. Langley, K. Laskowski, L. Levin, K. Peterson, T. Schultz, A. Waibel, D. Wallace, J. McDonough, H. Soltau, G. Lazzari, N. Mana, F. Pianesi, E. Pianta, L. Besacier, H. Blanchon, D. Vaufreydaz, and L. Taddei, "A Multi-Perspective Evaluation of the NESPOLE! Speech-to-Speech Translation System," Proceedings of the ACL 2002 Workshop on Speech-to-Speech Translation: Algorithms and Systems, Philadelphia, PA, 2002.

[4] A. Sarich, "Phraselator, one-way speech translation system," available at http://www.sarich.com/translator/.

[5] A. Waibel, A. Badran, A.W. Black, R. Frederking, D. Gates, A. Lavie, L. Levin, K. Lenzo, L. Mayfield, L. Tomokyo, J. Reichert, T. Schultz, D. Wallace, M. Woscsyna, and J. Zhang, "Speechalator: Two-Way Speech-to-Speech Translation on a Consumer PDA," Proceedings of the Eurospeech'03. Geneva, Switzerland, 2003, pp. 369–372.

[6] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, P. and V. Zue, "Galaxy-II: A Reference Architecture for Conversational System Development," Proceedings of the ICSLP'98, Sydney, Australia, pp. 931–934, available at http://communicator.sourceforge.net/, 1998.

[7] S. Dobrišek, "Analysis and Recognition of Phrases in Speech Signals," PhD Dissertation, University of Ljubljana, Slovenia, 2001.

[8] M. Romih and P. Holozan, "A Slovenian-English Translation System," Proceedings of the 3rd Language Technologies Conference, Ljubljana, Slovenia, 2002, p. 167.

[9] J. Vičič and T. Erjavec. "Vsak začetek je težak : avtomatsko učenje prevajanja slovenščine v angleščino," Proceedings of the conference Jezikovne tehnologije, Ljubljana, Slovenia, 2002, pp. 20-27.

[10] T. Erjavec, C. Ignat, P. Pouliquen, and R. Steinberger, "Massive Multi-lingual Corpus Compilation: Acquis Communautaire and Totale," Proceedings of the 2nd Language and Technology Conference, Poznań, Poland, 2005.

[11] T. Erjavec, "The IJS-ELAN Slovene-English Parallel Corpus," International Journal on Corpus Linguistics, Vol. 7, 2002, pp. 1–20.

[12] F. J. Och and H. Ney. "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, volume 29, number 1, pp. 19-51 March 2003, available at http://www.fjoch.com/GIZA++.html.

[13] R. Rosenfeld. "The CMU Statistical Language Modeling Toolkit, and its use in the 1994 ARPA CSR Evaluation," Proceedings of the ARPA SLT Workshop, available at http://www.speech.cs.cmu.edu/SLM/toolkit.html.

[14] U. Germann. "Greedy Decoding for Statistical Machine Translation in Almost Linear Time," Proceedings of the HLT-NAACL-2003, available at http://www.isi.edu/licensed-sw/rewrite-decoder/.

[15] J. Žganec Gros, "Text-to-speech synthesis for embedded speech user interfaces," WSEAS trans. commun., Mar. 2006, vol. 5, iss. 3, pp. 543-548.

[16] D. Verdonik, M. Rojc. Jezikovni viri projekta LC-STAR. Proceedings B of the 7th International Multi-Conference Information Society IS 2004: Jezikovne tehnologije, October 2004, Ljubljana, Slovenia, pp. 24-47.

[17] F. Mihelič, J. Žganec Gros, S. Dobrišek, J. Žibert, and N. Pavešić, "Spoken Language Resources at LUKS of the University of Ljubljana," Int. Journal on Speech Technologies, Vol. 6., No. 3, 2003, pp. 221–232.

[18] T. Korošec, "Opravljeno je bilo pomembno slovarsko delo o vojaškem jeziku," Slovenska vojska, Vol. 10, No. 10, 2002, pp. 12–13.

[19] T. Erjavec and S. Džeroski, "Machine Learning of Language Structure: Lemmatising Unknown Slovene Words," Applied Artificial Intelligence, Vol. 18, No. 1, 2004, pp. 17–41.

[20] Joseph P. Turian, Luke Shen, and I. Dan Melamed, "Proteus technical report #03-005: Evaluation of Machine Translation and its Evaluation," available at http://nlp.cs.nyu.edu/eval/.

[21] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. "Bleu: a Method for Automatic Evaluation of Machine Translation," 2001. RC 22176(W0109-022), IBM Research.