Morphology-based vs Unsupervised Word Clustering for Training Language Models for Serbian

Stevan J. Ostrogonac[1], Edvin T. Pakoci[2], Milan S. Sečujski[1], Dragiša M. Mišković[1]

[1]Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia, e-mail: ostrogonac.stevan@uns.ac.rs, secujski@uns.ac.rs, dragisa@uns.ac.rs

[2]AlfaNum – Speech Technologies, Bulevar Vojvode Stepe 40/7, 21000 Novi Sad, Serbia, e-mail: edvin.pakoci@alfanum.co.rs

Abstract: When training language models (especially for highly inflective languages), some applications require word clustering in order to mitigate the problem of insufficient training data or storage space. The goal of word clustering is to group words that can be well represented by a single class in the sense of probabilities of appearances in different contexts. This paper presents comparative results obtained by using different approaches to word clustering when training class N-gram models for Serbian, as well as models based on recurrent neural networks. One approach is unsupervised word clustering based on optimized Brown's algorithm, which relies on bigram statistics. The other approach is based on morphology, and it requires expert knowledge and language resources. Four different types of textual corpora were used in experiments, describing different functional styles. The language models were evaluated by both perplexity and word error rate. The results show notable advantage of introducing expert knowledge into word clustering process.

Keywords: N-gram; language model; word clustering; morphology; inflective languages

1 Introduction

Language models (LMs) are used for solving tasks related to many different fields. They are usually incorporated into system aimed at facilitating different modes and types of cognitive infocommunications, e.g. machine translation [1], automatic speech recognition [2], data compression [3], information retrieval [4], spell checking [5], plagiarism detection [6], diagnostics in medicine [7] etc. One of the most important roles of these models is within systems based on speech technologies and utilized as assistive tools. Assistive technologies, in general,

represent a very popular research topic [8], [9]. Another domain of application of language models is reflilated to the preservation of standards for different styles of communication, given the exponential growth of means through which people conduct their written correspondence. The issue of preserving standards in communication and usage of modern applications and devices has recently gained significant attention [10].

Practical application of language models usually implies some specific tasks for which insufficient training corpora are available. When no training data for a specific purpose are available, general language models may be used, but in such cases, they usually produce inferior results. In case a small training corpus consisting of topic-specific data can be obtained, word clustering can help optimize the resulting language model for the intended task [11]. The model trained by using in-domain data can also be interpolated with a general-purpose language model, by using one of a number of interpolation techniques [12], in order to improve performance.

Statistical *N*-gram language models [13] have been studied for decades and many improvements for specific applications have been developed [14], [15]. The introduction of neural network language models (NNLMs) [16] has brought general improvements over the *N*-gram models (even though NNLMs are more complex), especially when recurrent neural networks were considered as the means to take into account longer contexts (theoretically infinite ones). Recurrent neural network (RNN) language models were later optimized and have shown considerable improvements over many variations of *N*-gram models that they have been compared to [16]. Both statistical *N*-gram and RNN language models have been included in this research in order to obtain detailed information on how expert knowledge can contribute to word clustering, which is the basis for building high-quality class language models.

The corpora used in the experiments are a part of the textual corpus collected for training language models for an automatic speech recognition (ASR) system for Serbian [17]. Four segments have been isolated from the original corpus. Each of the segments represents one of the following functional styles – journalistic, literature, scientific and administrative. It has been shown that the functional style influences morphology-based word clustering since sentence structures differ significantly from one functional style to another [18].

In order to implement morphology-based word clustering for Serbian, a part-of-speech (POS) tagging tool [19] and morphologic dictionary [20] for Serbian were used. The clustering was done by assigning each word from the training corpus to a single morphologic class without considering the adjacent words. The number of morphologic classes that were defined within this research is 1117, but not all of them appear in the training corpora. In order to compare morphologic clustering to the unsupervised word clustering method, the number of morphologic classes that appeared in each corpus was set as the input parameter

for the corresponding unsupervised clustering. The unsupervised clustering was conducted by using an optimized version of Brown's algorithm [21]. The original Brown's algorithm was too complex for the experiments to be conducted in reasonable time, and even the optimized version took around 96 hours to complete the clustering on journalistic corpus (for which the vocabulary contained about 300,000 entries) on an Intel Core i5-4570 (3.2 GHz), RAM 16 GB DDR3 (1,333 MHz).

The rest of the paper is organized as follows. Section 2 briefly describes the training corpora used in the experiments. In Section 3, morphologic clustering for Serbian is presented. Section 4 gives a short overview of the unsupervised clustering method. In Section 5, the experiments are described in detail and the corresponding results are presented and discussed. The concluding section of the paper summarizes the main findings and outlines the plans for future research.

2 Training Corpora for Serbian

The training corpora for Serbian consist of many different text documents, which are classified into four groups, as described in the introduction. The journalistic corpus, which is the largest (around 17.4 million tokens), consists mainly of newspaper articles. The literature corpus (around 4 million tokens) consists of a collection of novels and short stories. The scientific corpus (around 865 thousand tokens) includes documents such as scientific papers, master and PhD theses. The administrative corpus is a collection of different legal documents (around 380 thousand tokens). In the experiments, 90% of data for each functional style was used for training LMs, and the remaining 10% was used for evaluation. It should be noted that text preprocessing included the removal of punctuation marks, converting letters to lowercase, and converting numbers to their orthographic transcriptions (POS tagging tool is used to determine the correct orthographic form). In Table 1, detailed information on corpora used for training LMs (90% of the entire textual content for each functional style) is provided.

Table 1
Contents of corpora for training language models for different functional styles

functional style	sentences	total words	vocabulary	morph. classes
administrative	13,399	340,261	17,924	447
scientific	36,621	776,926	59,705	646
literature	272,665	3,557,738	175,523	828
journalistic	662,813	15,645,691	299,472	836

3 Morphologic Clustering for Serbian

Morphology of the Serbian language is very complex and many morphologic features need to be included in the clustering process in order to obtain optimal results. The morphologic dictionary for Serbian contains the most important morphologic features of each entry. Some of these features have been empirically determined to have negligible effect on the quality of morphologic class-based LMs (they appear rarely or never in training corpora). This is, naturally, related to the size and content of training corpora and will most likely change in the future. The features that are currently in use for morphologic clustering, as well as some heuristics, will be presented here for each of the ten word types that exist in the Serbian language:

Nouns. Relevant morphologic information includes case, number, gender and type. Relevant types of nouns are proper (separate classes for names, surnames, names of organizations and toponyms), common, collective, material and abstract.

Pronouns. Morphologic features include case, number, gender, person and type. Not all the features are applicable to all pronoun types. For example, person is only applicable to personal pronouns. Furthermore, some types or groups of pronouns, or even single pronouns have been isolated and represent classes of their own. This is due to empirical knowledge and mostly refers to relative and reflexive pronouns.

Verbs. Features used (if applicable) are related to number, gender, and person, as well as to whether or not a verb is transitive or not and whether it is reflexive or not. Verb form types used to construct particular tenses or moods are, naturally, separated to different classes, although some of them are grouped together. Another relevant detail is related to whether a verb is modal/phase or not. However, as is the case with pronouns, some verbs are treated as separate classes (e.g. for the verb "*nemoj*" (don't) in the imperative mood, forms for each person are treated as separate classes, as is the case with the enclitic form of the verb "*ću*" (will)).

Adjectives. The morphologic features used include degree of comparison, case, number and gender. Invariable adjectives comprise a single class. Only one adjective is treated as a separate class due to its specific behaviour – "*nalik*" (similar to).

Numbers. Morphologic features include case, number and gender, but different types are treated separately, and there are many exceptions. For example, number one is treated separately and it forms 18 different classes, depending on its morphologic features. Furthermore, classes related to numbers two and three are joined together. Aggregate numbers represent a special group of classes. A class "other" is even formed from very rare cases.

Adverbs, conjunctions, particles. The classes are formed empirically. For frequent conjunctions and particles, most classes contain only one word.

Prepositions. Classification is based on the case of the noun phrase with which the preposition forms a preposition-case construction.

Exclamations. All exclamations form a single class.

As can be concluded, a great effort and expert knowledge are needed to define morphologic classes. When it comes to morphologic clustering for Serbian, it should be noted that the previously mentioned POS tagging tool supports context analysis (based on hand-written rules) and consequent soft clustering of words, which results in higher accuracy of language representation. However, this requires POS tagging in run-time when a language model is used, which is time-consuming, and therefore not suitable for some applications. Furthermore, morphology-based models with soft clustering cannot be compared directly to models based on unsupervised clustering, which is why context analysis was not used in the experiments described within this work.

4 Unsupervised Word Clustering

As opposed to morphologic clustering, automatic clustering that requires no expert knowledge or additional resources, relying only on statistics derived from textual corpus, was considered within the experiments. For unsupervised clustering, Brown's clustering was performed by using the SRILM toolkit [22].

The time complexity of the Brown's algorithm in its original form [21] is $O(V^3)$, where V is the size of the initial vocabulary. The algorithm involves initial assignment of each of the types (distinct words) to a separate class, after which greedy merging is applied until the target number of classes is reached. An optimized version of the Brown's algorithm, also described in [21], which has the time complexity $O(VC^2)$, involves setting a parameter C, which represents the initial number of clusters. The idea is to assign C most frequent types to separate clusters, after which each new type (or cluster) is being merged with one of the existing clusters in an iterative manner. Even though there are some obvious problems with the Brown's algorithm, it has given relatively good results for English [22].

It should be noted that this unsupervised clustering method offers some advantages in the context of semantic information extraction (N-gram statistics often reflect semantic similarity). However, in direct comparison to the morphologic clustering, this is not very noticeable, since the number of target classes is determined by the number of morphologic classes, which is small and results in inevitable merging of groups of words that are not semantically similar.

Another detail that should be mentioned is that the implementation of Brown's clustering within SRILM includes only bigram statistics [22], while morphologic analysis, depending on the case, can take into account much wider context.

5 Experiments and Results

In order to compare unsupervised and morphologic clustering, perplexity (ppl) and word error rate (WER) evaluations were conducted for different types of models. It should be kept in mind that both ppl and WER depend on the data set that is used for evaluation. However, prior to the experiments that will be described within this section, ppl tests were conducted using 10 different test data sets (per functional style) extracted from the corpora, on trigram word-based models. Perplexities obtained on different data sets were very similar for three out of four functional styles, indicating that test data sets are fairly representative. The only style for which ppl varied significantly for different data sets was literature. This was to be expected since the literature corpus contains novels from different time periods that vary in vocabularies, as well as sentence structures. The test data set that was chosen for each of the functional styles was the one for which out-of-vocabulary (OOV) rate, obtained with the model that was trained on the corpus for the corresponding functional style, was the lowest. The OOV rates for administrative, literature, scientific, and journalistic styles are 1.88%, 2.11%, 3.61% and 0.79%, respectively.

5.1 Perplexity Evaluation

Perplexity evaluation was conducted for both statistical *N*-gram and recurrent neural network language models. For training and evaluation, SRILM toolkit was used for *N*-gram models, and RNNLM toolkit [23] for RNN LMs.

Statistical *N*-gram models of different orders were included in the experiments in order to compare how the length of the context that is taken into account influences the quality of LMs depending on the manner in which word classes are derived. As mentioned before, four different functional styles were analyzed. For each morphology-based LM (hereinafter referred to as M model), a corresponding model with the same number of word classes derived by using optimized Brown's algorithm was created (hereinafter referred to as U model). Since the number of classes is small for all the models (class "vocabulary", hereinafter referred to as *C*, contains between 443 and 836, depending on functional style), there was no need for pruning LMs after training.

The experiments included models of orders from 2 to 5. Since the difference between the results obtained for 4-gram and 5-gram models was insignificant,

only the results for bigram, trigram and 4-gram models will be presented. Table 2 shows the obtained perplexity values.

Table 2
Evaluation results for *N*-gram language models of different order, that are based on different word clustering methods (U – unsupervised, M – morphology-based) and different functional styles (*C* – class "vocabulary" size)

functional style	clustering type	2-gram ppl	3-gram ppl	4-gram ppl
administrative (C = 443)	U	1,052.64	816.64	762.74
	M	1,250.72	912.68	834.86
literature (C = 828)	U	8,089.15	6,974.93	6,896.57
	M	3,629.93	2,949.15	2,877.67
scientific (C = 646)	U	6,596.25	5,868.64	5,795.55
	M	3,268.83	2,727.51	2,679.21
journalistic (C = 836)	U	9,235.24	6,450.65	5,631.81
	M	7,744.16	5,753.14	5,057.89

The perplexity values for class *N*-gram models are calculated by using word *N*-gram probabilities estimated according to Equation 1 (w represents words, c represents classes):

$$P(w_n \mid w_1 \ldots w_{t-1}) = P(w_n \mid c_n)\, P(c_n \mid c_1 \ldots c_{n-1}). \quad (1)$$

The values presented in Table 2 seem to be large in general, when compared to some results that were obtained in previous research for Serbian, on standard models [24]. This indicates that increasing the number of classes would help improve the quality of the models, since the number of morphologic classes is rather small, and is appropriate for either situations when some domain-specific, very small corpora are available for training, or when class models are interpolated in some way with standard models, in order to resolve issues with words that appear rarely but avoid over smoothing at the same time. There are also some applications that require language models to be small due to some hardware restrictions, in which cases word clustering, even to a very small number of classes, is the appropriate approach. However, the aim of this research was to compare morphologic clustering and clustering based on Brown's algorithm. It can be concluded that morphologic clustering is better for initial clustering, but increasing the number of classes and finding the optimal number for a specific application should be performed. Increasing the number of classes that are initially created by using morphologic information could be performed by a number of criteria, even by applying Brown's algorithm for further clustering within each of the morphologic classes. As additional information related to the comparison of the clustering methods, class-level perplexity values for the models presented in Table 2 are given in Table 3, illustrating that the M models predict classes more successfully than the U models.

Table 3
Class-level perplexity values for *N*-gram language models of different order, that are based on different word clustering methods (U – unsupervised, M – morphology-based) and different functional styles (*C* – class "vocabulary" size)

functional style	clustering type	2-gram ppl	3-gram ppl	4-gram ppl
administrative (*C* = 443)	U	55.49	41.5	38.76
	M	31.05	22.55	20.68
literature (*C* = 828)	U	125.94	108.6	107.38
	M	64.52	52.14	50.93
scientific (*C* = 646)	U	124.32	110.61	109.23
	M	43.74	36.23	35.68
journalistic (*C* = 836)	U	77.72	54.29	47.4
	M	43.8	32.31	28.35

The RNN language models were trained using parameter values that were within recommended ranges [23] for average-size tasks – hidden layer contained 500 units (-hidden 500), a class layer of size 400 was used in order to decrease complexity (-class 400), and the training (backpropagation through time – BPTT) algorithm ran for 10 steps in block mode (-bptt-block 10). Since these models consist of a much larger set of parameters, and the training parameters were not optimized within this research, they can not be compared to *N*-gram models directly (and there is no need for that since the goal is to compare different types of clustering), but the general conclusion related to M and U clustering methods can be drawn from the same evaluation procedure. The results are given in Table 4.

Table 4
Evaluation results for RNN language models based on different word clustering methods (U – unsupervised, M – morphology-based, *C* – class "vocabulary" size) and different types of training data

functional style	M ppl	U ppl
administrative (*C* = 443)	1,389.87	1,636.44
literature (*C* = 828)	4,065.68	10,500.93
scientific (*C* = 646)	3,994.52	10,412.45
journalistic (*C* = 836)	6,273.07	11,543.21

The results presented in Table 4 refer to the same training corpora that were used in the experiments for which the results are given in Table 2 (except that the test data set was split to validation and test data sets of equal sizes), the symbols for clustering methods have the same meaning and the sizes of class vocabularies are the same as well. The advantage of morphologic over unsupervised clustering is evident with RNN LM for all functional styles. Furthermore, it seems that the difference between the compared techniques is more emphasized with RNN LMs. This is probably due to long context that RNNs take into account. Theoretically, longer contexts can be modelled with higher order *N*-grams as well. However, in

practice, the back-off procedure introduces inaccuracies in the probabilities estimation process, which prevail over the benefits of introducing some information on longer contexts. RNNs model longer contexts more successfully, and therefore make better use of the contextual information contained within morphologic class models. This explains why here the results for M models are better than the results for U models for administrative style as well.

5.2 Word Error Rate Evaluation

Perplexity values calculated on test data do not always correlate with a language model's contribution when it is tested within a real system [16]. A common way of evaluating a language model within a practical application is conducting a word error rate test. The goal of a WER test is to determine the contribution of a language model to the accuracy of an automatic speech recognition system.

In order to perform word error rate comparisons between results using morphologic and automatic word clustering respectively, several tests were run, using AlfaNum speech recognition system [17]. All tests were based on a Serbian corpus of around 18 hours of speech material, including 26 different male and female speakers, divided into 13,000 utterances consisting of almost 160,000 tokens (words) and around 27,000 types (distinct words) [25]. This speech corpus is the most comprehensive corpus that currently exists for the Serbian language, and it has two quite different parts, one consisting of utterances from studio quality professionally read audio books, in which, naturally, the literature functional style dominates, and the other one, made of mobile phone recordings of commands, queries, questions and similar utterances expected in human-to-phone interaction via voice assistant type applications. This needs to be kept in mind when analysing WER results for different functional styles. All audio recordings were sampled at 16 kHz, 16 bits per sample, mono PCM [26].

As an acoustic model, a purely sequence trained time delay deep neural network (TDNN) for Serbian was used [17]. These so-called "chain" models are trained using connectionist temporal classification (CTC) in the context of discriminative maximum mutual information (MMI) sequence training with several specifics and simplifications, most notably frame subsampling rate of 3. It was trained on the training part of the above-mentioned speech corpus, which has almost 200 hours of material (140 hours of which were audio books). Neural network parameters were optimized on a range of different values until the best combination was decided on. This setting included the usage of three additional pitch features alongside standard MFCCs and energy, and separate models for differently accented vowels, which produced the best WER using the original 3-gram language model trained with SRILM on the described training corpus transcriptions, with the addition of a section of the journalistic corpus for better probability estimation.

In WER experiments within this research, class language models were used, along with corresponding class expansions files (in the form required by SRILM). Furthermore, *N*-grams including words missing from the particular language model training corpus were excluded from the final language model. This was done for all 4 functional styles, for both morphologic and unsupervised word clustering methods. As the testing was done for bigram, trigram and 4-gram models, there were 24 tests performed in total. In this way, many out-of-vocabulary words were created, but the same number of them existed for all experiments for the given functional style, so WERs can be compared to each other.

The tests were performed using the open source Kaldi speech recognition toolkit [27], which utilizes weighted finite state transducers (WFSTs) and the token passing decoding algorithm for calculation of the best path through the generated lattice. All the tests were run automatically using a shell script that invoked particular helper scripts and Kaldi programs on several server machines. After initial high-resolution feature extraction (40 MFCCs, as in most typical similar setups) and per-speaker *i*-vector calculation (in an "online" manner), for each language model the decoding graph was created using information from the language model, pronunciation dictionary, desired context dependency and acoustic model topology (transitions), and finally the decoding procedure and best possible WER calculation was initiated. A range of language model weight values were tried (in comparison to a fixed acoustic weight), as well as several word insertion penalties.

The results of the tests are given in Table 5. It should be noted that OOV rates for administrative, literature, scientific and journalistic models on the transcription of the speech database that was used in these tests were quite high (especially for the administrative style, for which the corpus is very small, and contains very specific content): 36.37%, 3.93%, 19.3% and 4.62%, for the above-mentioned functional styles, respectively, which may explain generally high WER.

For scientific and journalistic style, morphologic clustering showed significantly better results. For administrative style, M models were only slightly more successful, while for literature style, U models were slightly more adequate. As expected, perplexity results were not correlated to WER results for all tests. However, WER results depend on acoustic models, as well as other parameters. Still, a general impression related to the content of Table 5 is that M models are more suitable for an ASR task. An interesting detail is related to relative WER between functional styles. The models related to different functional styles are not of the same size and cannot be compared directly. However, it can be observed that the best WER result (by far) was obtained for the model that was trained on literature style, even though journalistic training corpus is much larger, for example. This confirms the importance of functional style adaptation when training language models since the corpus that was used for WER tests consisted mainly of textual content written in literature style. Another interesting

observation is that ASR does not seem to benefit from trigram and 4-gram entries. This might be related to the quality of modelling longer contexts with *N*-gram models (effects of the backoff procedure). Unfortunately, RNN models that could provide more information on this phenomenon were not included in WER tests within this study, since the implementation of evaluation framework for these types of models is not yet finished.

Table 5

Evaluation results in terms of WER [%] for language models based on different word clustering methods (U – unsupervised, M – morphology-based, *C* – class "vocabulary" size), different types of training data, and different *N*-gram order

functional style	clustering type	2-gram WER	3-gram WER	4-gram WER
administrative (C = 443)	U	58.14	58.31	58.32
	M	57.45	57.61	57.65
	M	31.15	31.30	32.07
scientific (C = 646)	U	45.64	45.58	45.55
	M	42.81	43.14	43.44
journalistic (C = 836)	U	40.59	41.09	41.29
	M	35.66	36.29	36.82

One significant advantage of morphologic clustering is the fact that the models can lean on information from the morphologic dictionary for Serbian, that was mentioned earlier. Namely, for all the words that are contained within the dictionary (around 1,500,000 orthographically distinct surface forms) morphologic classes can be determined from the corresponding morphologic information, by applying the same procedure as with training corpora. In this way, a new word-class map is generated. If every word w, that belongs to a class c, is then assigned a probability $P(w|c)$, these words can be used to deal with the OOV word problem. In order to explore the benefits of using the information from the morphologic dictionary, another set of WER experiments was conducted. The added words were assigned values of $P(w_i|c_i)$ that were basically the averaged values of corresponding probabilities of all the words that originally belonged to classes c_i. The results of the experiments are presented in Table 6.

Table 6

Evaluation results in terms of WER [%] for language models based on morphologic word clustering, different types of training data, and different *N*-gram order, when additional information is obtained from the morphologic dictionary for Serbian

functional style	2-gram WER	3-gram WER	4-gram WER
administrative	24.66	25.13	25.24
literature	27.51	27.78	28.58
scientific	22.44	23.00	23.30
journalistic	31.37	32.06	32.57

A drastic improvement in terms of WER can be observed for all functional styles. Furthermore, in order to optimize models, the added words to class expansions were implemented as separate maps that are used only when a word cannot be found in the initial class expansion file. In other words, the addition of dictionary information does not significantly increase a model's complexity since the new map is only used when an OOV word is encountered.

Conclusions

The experiments described within this paper have shown that morphologic word clustering for Serbian, in comparison to the unsupervised clustering method based on Brown's algorithm, generally results in considerably more adequate language models, regardless of the language modelling concept (RNN or *N*-gram) or of the type of textual data (functional style). Morphologic clustering with the restriction of assigning each surface form to only one class has shown fairly good results, which is important for practical applications, since it only requires a simple look-up table for run-time word classification. Naturally, context analysis in the process of morphologic clustering can introduce further improvements (with inevitable rise in complexity).

The WER results for LMs based on morphologic classes, while promising, are not sufficiently good for many applications. In some applications, where there is no limit on memory storage or computational cost, these models can be interpolated with word-based LMs, in order to obtain better results. However, if only small class LMs are acceptable, it is an imperative to store as much linguistic information as possible in a small number of word classes. The aim of further research will be to explore other approaches to improving the word clustering process. The main idea is to increase the number of classes by starting with morphologic classes described within this research and perform further division of classes based on some other criteria. These models would still be much smaller than word-based models, but the number of classes would be adjustable in order to obtain optimal results for a specific application. Furthermore, word clustering based on semantics is another challenge and an object of further research for Serbian. It will, however, require deeper knowledge of how language is learned by a human brain, which is a topic that is also gaining popularity [28].

Acknowledgement

The presented study was sponsored by the Ministry of Education, Science and Technological Development of the Republic of Serbia, under the grant TR32035. Speech and language resources were provided by AlfaNum – Speech Technologies from Novi Sad, Serbia.

References

[1] Brants T., Popat A., Xu P., Och F., Dean J.: Large Language Models in Machine Translation. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational

Natural Language Learning (EMNLP-CoNLL), Association for Computational Linguistics, Prague, Czech Republic, pp. 858-867 (2007)

[2] Mengusoglu E., Deroo O.: Turkish LVCSR: Database Preparation and Language Modeling for an Agglutinative Language. Acoustics, speech and signal processing, student forum, Salt Lake City, Utah, USA (2001)

[3] El Daher A., Connor J.: Compression Through Language Modeling. NLP courses at Stanford, URL: http://nlp.stanford.edu/courses/cs224n/2006/fp/aeldaher-jconnor-1-report.pdf (accessed on April 21st, 2016)

[4] Song F., Bruce Croft W.: A General Language Model for Information Retrieval. In Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 279-280 (1999)

[5] Verberne S.: Context-Sensitive Spell Checking Based on Word Trigram Probabilities. Master's thesis, University of Nijmegen, Netherlands (2002)

[6] Miranda-Jiménez S., Stamatatos E.: Automatic Generation of Summary Obfuscation Corpus for Plagiarism Detection. Acta Polytechnica Hungarica, Special Issue on Computational Intelligence, Vol. 14, No. 3, pp. 99-112 (2017)

[7] Rentoumi V., Paliouras G., Danasi E.: Automatic detection of linguistic indicators as a means of early detection of Alzheimer's disease and of related dementias: A computational linguistics analysis. Proceedings of the 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Debrecen, Hungary, pp. 33-38 (2017)

[8] Izsó L.: The significance of cognitive infocommunications in developing assistive technologies for people with non-standard cognitive characteristics: CogInfoCom for people with nonstandard cognitive characteristics. 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Győr, Hungary (2015)

[9] M. Macik, I. Maly, J.Balata, Z. Mikovec: How can ICT help the visually impaired older adults in residential care institutions: The everyday needs survey. 8th IEEE International Conference on Cognitive Infocommunications, Debrecen, Hungary (2017)

[10] Toth A., Tovolgyi S.: The Introduction of Gamification - A review paper about the applied gamification in the smartphone applications. 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Wroclaw, Poland (2016)

[11] Whittaker E. W. D., Woodland P. C.: Efficient Class-Based Language Modeling for very Large Vocabularies. Acoustics, Speech and Signal Processing, Vol. 1, pp. 545-548, Salt Lake City, Utah, USA (2001)

[12] Broman S., Kurrimo M.: Methods for Combining Language Models in Speech Recognition. Proceedings of 9th European Conference on Speech Communication and Technology, pp. 1317-1320 (2005)

[13] Mikolov T., Deoras A., Kombrink S., Burget L., Černocký J.: Empirical Evaluation and Combination of Advanced Language Modelling Techniques. Proceedings of Interspeech, Florence, Italy, Vol. 2011, pp. 605-608 (2011)

[14] Majdoubi J., Tmar M., Gargouri F.: Language Modeling for Medical Article Indexing. Chapter, Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, Vol. 295 of the series Studies in Computational Intelligence, pp. 151-161 (2010)

[15] Kuhn R., De Mori R.: A Cache-Based Natural Language Model for Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 12, No. 6 (1990)

[16] Mikolov T.: Statistical Language Models Based on Neural Networks. PhD Thesis, Brno University of Technology, Czech Republic (2012)

[17] Pakoci E., Popović B., Pekar D.: Fast Sequence-Trained Deep Neural Network Models for Serbian Speech Recognition. Proceedings of DOGS, Novi Sad, Serbia, pp. 25-28 (2017)

[18] Ostrogonac S., Mišković D., Sečujski M., Pekar D., Delić V.: A Language Model for Highly Inflective Non-Agglutinative Languages. SISY – International Symposium on Intelligent systems and Informatics, Subotica, Serbia, pp. 177-181, ISBN 978-1-4673-4749-5 (2012)

[19] Ostrogonac S.: Automatic Detection and Correction of Semantic Errors in Texts in Serbian. Primenjena lingvistika, ISSN: 1451-7124, accepted for publication (2016)

[20] Sečujski M.: Accentuation Dictionary for Serbian Intended for Text-to-Speech Technology. Proceedings of DOGS, pp. 17-20, Novi Sad, Serbia (2002)

[21] Brown P., De Souza P., Mercer R., Della Pietra V., Lai J.: Class-based *N*-gram Models of Natural Language. Computational Linguistics, Vol. 18, No. 4, pp. 467-479 (1992)

[22] Stolcke A.: SRILM – An Extensible Language Modeling Toolkit. Proc. Intl. Conf. on Spoken Language Processing, Vol. 2, pp. 901-904 (2002)

[23] Mikolov T., Kombrink S., Deoras A., Burget L., Černocký J.: RNNLM – Recurrent Neural Network Language Modeling Toolkit, In: ASRU 2011 Demo Session (2011)

[24] Ostrogonac S., Sečujski M., Mišković D.: Impact of training corpus size on the quality of different types of language models for Serbian, 20. Telecommunications forum TELFOR, Belgrade, 20-22 November (2012)

[25] Pakoci E., Popović B., Pekar D.: Language Model Optimization for a Deep Neural Network Based Speech Recognition System for Serbian. Proceedings of SPECOM, Hatfield, United Kingdom, LNAI, Vol. 10458, pp. 483-492 (2017)

[26] Suzić S., Ostrogonac S., Pakoci E., Bojanić M.: Building a Speech Repository for a Serbian LVCSR System. Telfor Jornal, Vol. 6, No. 2, pp. 109-114 (2014)

[27] Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlicek P., Qian Y., Schwarz P., Silovsky J., Stemmer G., Vesely K.: The Kaldi Speech Recognition Toolkit. In: ASRU 2011 (2011)

[28] Katona J., Kovari A.: Examining the Learning Efficiency by a Brain-Computer Interface System. In Acta Polytechnica Hungarica, Vol. 15, No. 3 (2018)