DOI :10.2298/PSI0901107D

STABILITY OF THE SYNTAGMATIC PROBABILITY DISTRIBUTIONS

Strahinja Dimitrijević[1]
Department of Psychology, University of Banja Luka, Bosnia & Herzegovina

Aleksandar Kostić
Department of Psychology, University of Belgrade, Serbia

Petar Milin
Department of Psychology, University of Novi Sad, Serbia

The aim of the present study is to establish criteria for the optimal size of a corpus that can provide stable conditional probabilities of morphological and/or syntagmatic types. The optimality of corpus size is defined in terms of the smallest sample that generates probability distribution equal to distribution derived from the large sample that generates stable probabilities. The latter distribution we refer to as "target distribution". In order to establish the above criteria we varied the sample size, the word sequence size (bigrams and trigrams), sampling procedure (randomly chosen words and continuous text) and position of the target word in a sequence. The obtained distributions of conditional probabilities derived from smaller samples have been correlated with target distributions. Sample size at which probability distribution reaches maximal correlation (r=1) with the target distribution was taken as being optimal. The research was done on Corpus of Serbian language. In case of bigrams the optimal sample size for random word selection is 65.000 words, and 281.000 words for trigrams. In contrast, continuous text sampling requires much larger samples to reach stability: 810.000 words for bigrams and 868.000 words for trigrams. The factors that caused these differences remain unclear and need additional empirical investigation.

Key Words: *corpus linguistics, quantitative linguistics, optimal sample size, conditional probabilities, Serbian language.*

[1] ✉: strahinjad@blic.net

A significant body of linguistic studies are based on estimated probabilities (frequencies) of various language phenomena, which have been derived from language corpora. However, in order to make solid conclusions probabilities of interest should be stable. This requirement faces number of problems because probability distributions of language phenomena change as a function of a corpus size. This, on the other hand, leads to systematic change of relevant statistical parameters, like measures of central tendency and dispersion (Baayen, 2001; Milin & Ilić, 2003). What is the size of a language sample that provides us with stable probabilities or, put differently, what is statistically reliable corpus is a question conspicuously neglected in corpus linguistics. It is generally acknowledged that reliability of probability estimate depends solely on corpus size, where the larger the corpus the more reliable the probability estimate (Church et al., 1991; Biber et al., 1998; 2000; Biber & Conrad, 2001; Yang et al., 2002, etc.). This criterion leaves us with no clear insight about the size of a corpus because index of reliability remains obscured, allowing for an infinite increase of language sample. Other authors specify reliability with respect to the phenomenon under investigation, as the minimal sample size that will allow for stable, unbiased probabilities estimate (cf. Biber, 1990; Atkins et al., 1992).Kostić and his associates define reliability of a text sample as the *minimal* amount of language materials that provides probability distribution equivalent to the distribution derived from a large sample whose probabilities proved to be stable (Kostić, A., 1996; Kostić et al., 2008). The reliability of probability estimate was evaluated for Serbian language at the level of phonemes, case forms of nouns and adjectives, verb person and word types. Due to the fact that systematic increase of the text size brings apparent increase of stability of frequency distributions, it could be postulated that for the very large corpora frequency estimates are close to absolute stability. Consequently, probability distributions derived from such corpora can be treated as *target distributions* for the respective language that have to be matched by distributions derived from smaller samples. The text size on which the given probability distribution perfectly matches the analogue target distribution is considered to be *the optimal* sample size for a given language phenomenon (Kostić et al., 2008). The index of perfect match is the maximal correlation (i.e. correlation coefficient being 1) between target probability distribution and the analogue distribution derived from a given text sample. Maximal correlation implies that further increase of the text sample will not affect probability distribution. Therefore, the distribution could be considered to be stable and, by the same token, probabilities could be treated as being reliable.

Kostić et al. (2008) showed that in Serbian sample of 8.500 nouns provides stable probability distribution of noun cases. Similarly, optimal sample size for case forms of adjectives is 5.250 adjectives, while 900 verbs are sufficient to reach reliable estimate of probability distribution for verb person. When the optimal sample size is weighted with probability of a given language category (e.g. probability of a word to be a noun, adjective, verb etc.), we get *the optimal size of the corpus* to generate reliable probabilities for given language phenomena. Thus, for example, the optimal sample size for probability estimate of case forms of nouns is 38.000 words; for case forms of adjectives is 62.000, and 4.000 words for verb person (Kostić et al., 2008).

Corpora that provide reliable probability estimates proved to be crucial for different tasks in *Natural Language Processing – NLP*. Charniak and his associates proposed approach based on unequal distribution of word probabilities and their forms. The approach relies exclusively on probabilities of tags for given word in text („dumb tagger“) and gives correct Part-of-Speech tagging (automatic detection of word types, usually referred as *POS tagging*) for about 90% of words in an English text (Charniak et al., 1993; Charniak, 1997). Having in mind the simplicity of tagger proposed by Cherniak and his associates, this should be taken as the lower bound of efficiency for any automatic word processing system. However, most of such systems were developed for English where word processing is not a big challenge because „English has very little morphology, and so the need for dealing intelligently with morphology is not acute“(Manning & Schütze, 2000, 132).

The main challenge in an automatic language processing of English or similar analytical languages is *parsing*, i.e. the analysis of syntactic components of a sentence. The first step of parsing is to determine word types. However, English words in isolation are morphologically poor, with reduced inflectional paradigms, and therefore equivocal for word type. Therefore, for further analyses of their morphological attributes it is necessary to rely on contextual information, i.e. information about assonance between two or more morphological types. Current results show that the best parsing systems reach from 95% to 97% of accuracy. Simply stated, syntagmatic contextual information adds 5-7% of accuracy to Charniak's approach, with the simple lawful property that is based on unequal probabilities of words and word forms.

In contrast, satisfactory results in automatic word processing based on information about assonance between two or more morphological types are not observed for languages with rich inflectional morphology. For example, system that uses contextual or syntagmatic information for Part-of-Speech tagging of Slovene achieved 55% of accuracy (Džeroski et al., 2000; Milin, 2005), while in Serbian, with the same source of information success in tagging ranges between 54% and 63% for words that carry more than one possible tag and/or lemma (Dimitrijević et al., 2008). Having in mind the fact that almost 45% of words in Serbian text are homographs (Ilić & Kostić, 2002) this result is far from acceptable. Nevertheless, following the Charniak's idea of „dumb tagger“, by using frequency dictionary along with ranking by frequency, correct lemmatization for Serbian increases up to 94% and 96% (Ilić & Kostić, 2002). It should be emphasized that such high level of correct lemmatization was reached with no use of contextual syntagmatic information whatsoever. With these findings in mind, it is clear that for morphologically rich languages, like Serbian, automatic word processing cannot be successful if we rely exclusively on higher-order information at the level of morphological types, even if we include context such as syntagmatic. Finer grained *knowledge about the world*, inherent to probabilities of individual words and their grammatical forms is required (cf. Milin, 2004; Dimitrijević et al., 2008).

This is not surprising, as word frequency effect is one of the most robust effects in psycholinguistics. Along with the frequency effect for single word, numerous research have also presented the role of two or three words sequences probabilities in production, comprehension and learning of the language. For example, McDonald et al. (2001) showed that conditional probability of bigrams is a good predictor of the gaze duration in reading of the second word in the sequence, while Bod (2000; 2001) found that more frequent sentences of three words (subject-verb-object) are being recognized faster than less frequent sentences. Different studies showed that words in high-frequency or high-probability word couples are reduced in the same way (Gregory et al., 2000; Jurafsky et al., 2001; Bell et al., 2001). In relation to language acquisition, Mintz (2003) underlines that frequent frames, defined as two jointly occurring words with one word intervening, are crucial component for acquisition of grammatical categories (e.g. noun, verbs, etc). Joint element of all these frequency effects is the fact that contextual information are of lexical nature. For time being, research results for frequency effects of larger nonlexical syntactic structures (such as idiom, relative clauses, etc.) on comprehension, are preliminary and relatively inconclusive (cf. Jurafsky, 2003).

In the present state of corpus linguistics and NLP some trade-offs seem to be mandatory. If word disambiguation cannot be accurately fulfilled with local lexical-probabilistic approach, like in analytic languages, system must rely on context in which a given word appears. This, on the other hand, may not be beneficiary for languages with rich morphology. Research on automatic lemmatization in Serbian showed that relying on contextual grammatical type probabilities leads to marginal increase of lemmatization accuracy, although usability of contextual information at the level of grammatical types was not systematically examined (Milin, 2005).

Statistical models of natural language processing rely on estimated probabilities of word's characteristics. These probabilities are generated on the basis of data derived from corpora, which makes them sensitive to distribution of linguistic categories. Present research starts from the initial assumption that the very first step in NLP must be the determination of the optimal text size for reliable probability estimates.

Kostić (1996) and Kostić et al. (2008) determined optimal sample sizes for isolated morphological categories, but contextual or syntagmatic structures that carry information about assonance between two or more morphological types remained out of their research scope. These syntagmatic structural information can be formalized by means of *conditional probabilities* of the target word type conditioned on its nearest neighbor types that immediately precede or succeed the target (for introductory overview of the Probability theory in language research see Bod, 2003). Present research aims at estimating conditional probabilities at the level of word types in their various syntagmatic contexts, determining the sample size that would be optimal for NLP tasks that make use of those contextual information.

METHOD

Design

The procedure developed by Kostić et al. (2008) was adopted to examine the stability of probability distributions of two and three words sequence (i.e. bigrams and trigrams).[2] The following factors that may influence stability of the probabilities of target words conditioned on syntagmatic context were investigated.

a. Size of the text sample. The conditional probabilities were estimated through systematic increase of text sample.

b. Sampling procedure. Sampling was performed on continuous text and randomly selected individual words;

c. Size of the word sequence. Word sequence consisted of either two (bigram) or three (trigram) words;

d. Position of the target word in a sequence. For bigrams: in front or behind; for trigrams: in front, in the middle, and at the end.

Given factors and their combinations exhaust all possible sampling strategies, hence providing full verification of their eventual effect on stability of syntagmatic probability distributions. Combining the above factors, for each sample size ten types of samples were derived (Table 1).

Table 1: Sampling scheme by type, sequence size and target position

Random word samples					Continues text samples				
Bigrams		Trigrams			Bigrams		Trigrams		
Ox	**xo**	**oox**	**oxo**	**xoo**	**ox**	**xo**	**oox**	**oxo**	**xoo**

Legend: o - context, x – target word.

Sample

There were two distinct procedures of sampling bigram and trigram probability distributions: a. Words were randomly chosen from text until word sample of a given size has been provided, and b. Initial word of a continuous text was randomly chosen and the integral text of a given size was taken as a sample.

The initial sample consisted of 32 words ($n=2^5$). Samples were then increased by exponential progression up to the size of 262.144 words ($n=\{2^5, 2^6,..., 2^{18}\}$). If stability of conditional probabilities was not reached in this range, the sample was increased up to 300.000 words, with additional increase of 100.000 words if necessary. The maximal size of the sample was 900.000 words. In order to get more precise estimate

[2] In this study sequences of two and three words are referred to as „bigrams“ and „trigrams“, even though this terms often refer to the sequences of two or three graphemes.

of the optimal sample size, finer sampling was performed by linear increase of the sample by 1.000 words. This was done for the sample which preceded sample size that perfectly matched target probability distribution (i.e. the first appearance of maximal correlation).

Pearson product-moment correlation coefficient was used as a measure of similarity between the two probability distributions. Thus, for example, if we detected the correlation r=1 on the sample of 800.000 words, the size of the sample was linearly increased by 1000 words, starting from the first preceding sample (700.000), (e.g. 701.000, 702.000, 703.000 words, etc.) until we got four successive maximal correlation coefficients (r=1). For each sample size 100 replications were made (i.e. one hundred independent samplings) to obtain the *average* correlation coefficient. Stability of the obtained mean was additionally evaluated by its corresponding standard deviation. This procedure parallels the procedure used by Kostić at al. (2008).

Materials

Research was performed on the sample of Serbian prose, retrieved from the *Corpus of Serbian Language* /CSL/ (Kostić, Đ., 2001). The sample consists of 994.227 words, and includes over 100 titles of more than 70 authors. It covers publications from different registers (novels, essays, plays) published between 1945 and 1957. With an *a priori* assumption that size of a sample will provide us with stable distributions of conditional probabilities at the level of word types, distributions derived from CSL are treated as target distributions.

Programme for sampling was written in developmental environment *MS VisualC++.net Standard*, while Microsoft Office Excel 2003 has been used for the statistical analyses.

RESULTS

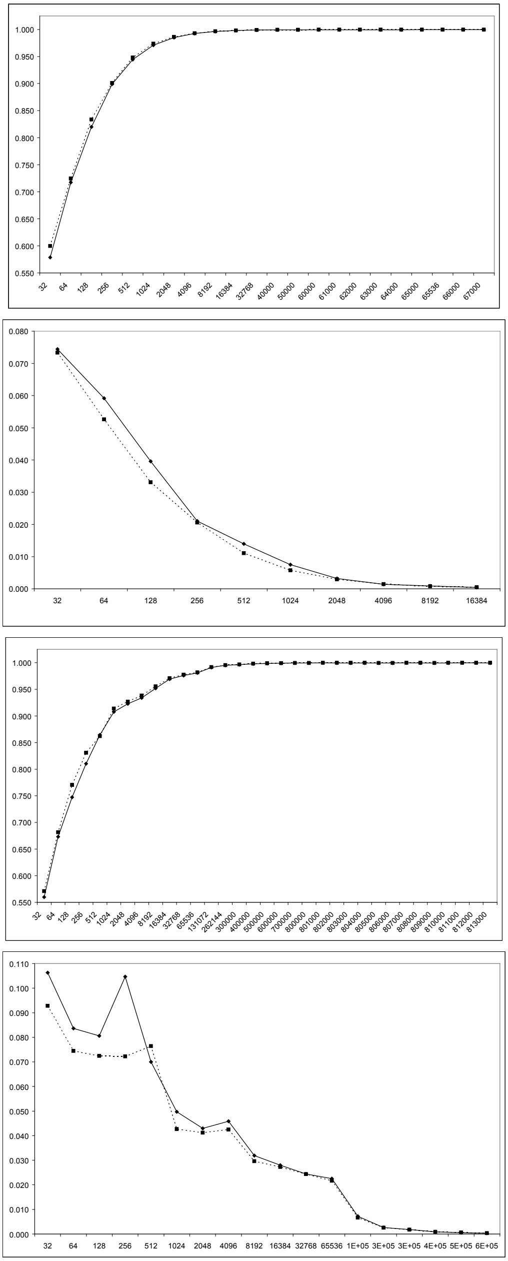
For each sample size total of 100 of Pearson product-moment correlation coefficients were calculated for the two conditional probability distributions (i.e. distribution obtained from a given sample and the target distribution). Separate correlations were calculated on the basis of raw frequencies of word types combinations (e.g. noun, verb, etc.) within bigrams and trigrams, derived from the sample and a large sample that generates stable probabilities. Average correlation coefficient was treated as index of similarity between the two distributions (cf. Kostić et al., 2008). When the average correlation coefficient reached theoretical maximum (r=1), the given sample size was taken as being optimal to obtain reliable probability estimates. Standard deviations were used as an additional indicator of stability. For example, when average correlation reached maximum (r=1), with standard deviation being 0, the similarity between the two distributions was perfect in each of the 100 replications. Distributions of conditional probabilities at the word level were analyzed as a function of: a. sample size, b. sampling procedure (continuous text vs. random words samples), c. size of the word's sequence (bigrams vs. trigrams), and d. position of the target word in a sequence. Results are presented separately for bigram and trigram sequences in Figures 1 and 2, while details have been given in Appendices 1 and 2, respectively.

In bigram sequences average correlation coefficient proved to be non-linear positively decelerating function of sample size, reaching its asymptote in r=1. Random word samples reach stability faster than continuous text samples (62.500 words vs. 805.500 words). Within each of the two sampling procedures sample sizes are in the same order of magnitude, irrespective of the position of the target word. Stable conditional probabilities on random words samples are derived from samples of 60.000 words for preceding targets, and 65.000 words for succeeding targets. On the other hand, conditional probabilities for samples of continues text reach stability with 801.000 words when target preceded context, and 810.000 words when context preceded target.

Difference between the two sampling procedures is also mirrored in standard deviations. For random word sampling standard deviation becomes stable at 16.384 words, with the average correlation coefficient being 0,998, while for the continuous text stability occurs at 600.000 words, with the average correlation coefficient of 0,999 (Figure 1). In both sampling procedures standard deviations are negatively decelerating functions of the sample size. Note that in random word samplings standard deviations declines continuously, while in continuous text sampling somewhat irregular decrease was observed in the range between 32 and 65.536 words (Figure 1). For samples of continues text function that captures variation of standard deviation does not resemble any of the known regular functions. Additional analyses of the sampling software confirmed that this outcome is not a side-effect of some systematic or uncontrolled error in sampling procedure.

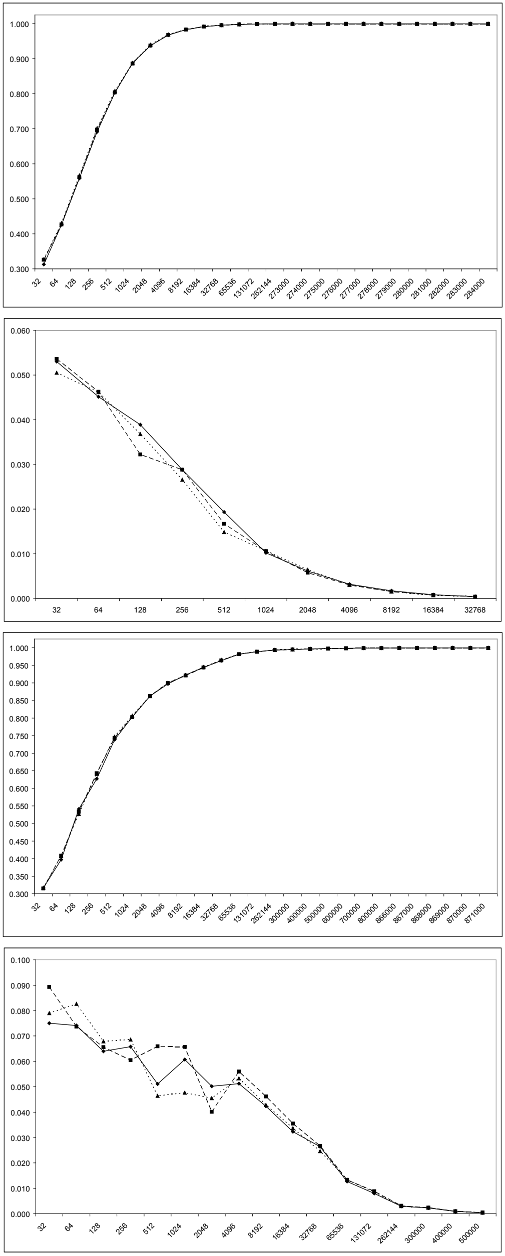
Sampling procedure affects distribution of conditional probabilities of trigram sequences as well (Figure 2). Again, random word sampling seems to be superior in reaching full stability, regardless of the position of the target-word (277.667 vs. 866.667). Within the sampling procedure, the position of the target word does not affect the optimum for sample size. Random word sampling reaches stable conditional probabilities with 273.000 words when target is the leftmost, 279.000 words when target is in the middle, and 281.000 words when target is the rightmost word in a trigram sequence. For samples of continues text stable conditional probabilities are reached with 866.000 words when the target-word is far left or far right in a trigram, and with 868.000 words when the target-word is in the middle of the trigram sequence. Regardless of the position of the target, standard deviations become stable with 500.000 words, with the average correlation coefficient being 0,998.

Figure 1: The average correlation coefficients (Mr) and their standard deviations (SD_{Mr}) as a function of sample size in bigram sequences



Legend: Upper: Mr (left) and SD_{Mr} (right) for random word samples. Lower: Mr (left) and SD_{Mr} (right) for continues text sample sizes. Solid lines represent sequences when context precedes the target, and dashed lines when context follows the target.

Figure 2: The average correlation coefficients (Mr) and their standard deviations (SD_{Mr}) as a function of sample size in trigram sequences



Legend: Upper: Mr (left) and SD_{Mr} (right) for random word samples. Lower: Mr (left) and SD_{Mr} (right) for continues text sample sizes. Solid lines represent sequences when context precedes the target, dotted lines when context follows the target, and dashed lines when target is in the middle of the sequence.

As with bigram sequences, the same general shape of functions is observed with trigrams as well. There is a non-linear decelerating growth of correlation coefficients to the asymptote of r=1, and decelerating decline of the standard deviations as a function of the sample size. Again, continuous text is characterized by unsystematic, irregular decline in the range between 32 and 4.096 words (Figure 2). As in the case of bigrams, additional examination confirmed the reliability of this finding.

DISCUSSION

Following the seminal work of Kostić (1996) and Kostić et al. (2008) this study focused on reliability of conditional probabilities at the level of word types in Serbian. Its main goal was to determine the optimal sample size for sequences of two and three words. The effects of the sampling procedure, sequence size and the target position in bigrams and trigrams were investigated as well.

The outcome of this research indicates that size of the sample at which conditional probabilities become stable depends on size of the syntagmatic sequence (bigram or trigram) and sampling procedure. For random word samples stability is reached at 65.000 words for bigrams and 281.000 words for trigrams, while for continues text samples the optimal sample size is 810.000 words for bigrams and 868.000 words for trigrams. Differences due to the position of the target-word in a sequence are negligible.

The observed difference in the optimal sample size between bigrams and trigrams should be attributed to the fact that the number of possible combinations grows exponentially as the word sequence expands. Number of possible combinations with bigrams is 120 and with trigrams it is 1.440.[3] In other words, the number of possible combinations with trigrams is in the order of magnitude larger. Consequently, for the same sample size we get sparse frequency distribution and unstable probability values.

Combinatorial explosion seems to be obvious explanation for the effect of the sequence size. However, differences caused by sampling procedure (random words vs. continues text) are not as easy to account for. Results show that it takes significantly larger amount of continues text to get stable probability distributions. We may hypothesize that the reason for this finding hinges to the structure of continues text, which is built according to syntactic and semantic rules that govern non-random occurrence of words (cf. Baayen, 2001). However, if the text structure makes sample non-random question remains why this non-randomness is not reducing the uncertainty which will allow for faster reaching stability. The reason for such an outcome remains unclear. There is yet another finding that needs additional examination. Namely,

[3] While target-word could appear in 10 modalities (word types), context could appear in 12 (10 word types + punctuation marks and non-coded context). For example, when target-word in bigram is behind context number of combination is 12x10=120. When target-word is far to the right in trigram, number of combinations is 12x12x10=1440. Note that punctuation marks have been used only as a context and that, rarely, some contexts have not been tagged.

irregular decline of standard deviations in samples of continues text was observed in the range between 32 and 65.566 words for bigrams, and 32 and 4.096 words for trigrams. This irregularity appears to be stable, thus excluding any explanation that will rely on calculation or procedural flaws.

Continuous text samples show one advantage over random word samples: the difference between bigram and trigram optimal sample size for random words is much larger than the difference observed with continuous text (216.000 vs. 58.000 words). The cause of this difference could be the trade-off between the combinatorial explosion from bigrams to trigrams and the optimal size of random vs. continuous sample. Since continuous text samples become stable much later, almost all noise (or variability) has been already taken into account, while random word samples are more sensitive to factors such as the size of the syntagmatic sequence.

Along with several open questions which require further empirical examination, this research established the basis for empirical specification of the optimal sample size for building corpora that will provide us with stable conditional probabilistic at the level of bigrams and trigrams. In addition to practical implications, sample size and stability of conditional probabilities open number of questions related to the very nature of language and its cognitive aspects.

REFERENCES

Atkins, S., Clear, J., & Ostler, J. (1992). Corpus design criteria. *Literary and Linguistics Computing*, *7*, 1-16.

Baayen, R. H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America*, *113*(2), 1001-1024.

Biber, D., Conrad, S., & Reppen, L. (1998). *Corpus Linguistics*. Cambridge: University Press.

Biber, D., & Conrad, S. (2001). Quantitative corpus-based research: Much more than bean counting. *TESOL Quarterly*, *35*, 331-6.

Bod, R. (2000). The storage vs. computation of three-word sentences. *AMLAP 2000*. Leiden, The Netherlands.

Bod, R. (2001). Sentence memory: The storage vs. computation of frequent sentences. *14th Annual CUNY Conference on Human Sentence Processing*. Philadelphia, Pennsylvania, USA.

Bod, R. (2003). Introduction to Elementary Probability Theory and Formal Stochastic Language Theory. In R. Bod, J. Hay & S. Jannedy (Eds.), *Probabilistic Linguistics* (pp. 11-38). Cambridge, MA: MIT Press.

Charniak, E. (1997). Statistical Techniques for Natural Language Parsing. *AI Magazine*, *18*, 33-44.

Charniak, E., Hendrickson, C., Jacobson, N., & Perkowitz, M. (1993). Equations for part-of-speech tagging. In *Proceedings of the Eleventh National Conference on Artificial Intelligence* (pp. 784-789). Menlo Park, CA.

Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). Using Statistics in Lexical Analysis. In U. Zernik (Ed.), *Lexical Acquisition: Using On-line Resources to Build a Lexicon* (pp. 115-164). Hillsdale, NJ: Lawrence Erlbaum.

Dimitrijević, S., Milin, P., & Kostić, A. (2008). Primjena kontekstualnih jezičkih informacija na nivou vrsta riječi u zadatku automatske lematizacije. *Empirijska istraživanja u psihologiji – XIV,* Beograd.

Džeroski, S., Erjavec, T., & Zavrel, J. (2000). Morphosyntactic Tagging of Slovene: Evaluating Taggers and Tagsets. In *Proceedings of the Second Conference on Language Resources and Evaluation* (pp. 1099-1104).

Gregory, M., Raymond, W., Fosler-Lussier, E., & Jurafsky, D. (2000). The effects of collocational strength and contextual predictability in lexical production. *Chicago Linguistic Society*, *35*, 151-166.

Ilić, N., & Kostić, A. (2002). Problem homografije pri automatskoj lematizaciji. *Empirijska istraživanja u psihologiji – VIII*, Beograd.

Jurafsky, D. (2003). Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production. In R. Bod, J. Hay & S. Jannedy (Eds.), *Probabilistic Linguistics* (pp. 11-38). Cambridge, MA: MIT Press.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. L. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 229-254). Amsterdam: John Benjamins.

Kostić, A. (1991). Informational approach to processing inflected morphology: Standard data reconsidered. *Psychological Resea*rch, *53*, 62-70.

Kostić, A. (1995). Informational load constrains on processing inflected morphology. In L. B. Feldman (Ed.), *Morphological Aspects of Language Processing* (pp. 317-345). Hillsdale, NJ.: Erlbaum.

Kostić, A. (1996). Reprezentativnost jezičkog korpusa i mentalni leksikon. *LEP Saopštenje*, 39.

Kostić, A., Ilić, S., & Milin, P. (2008). Aproksimacija verovatnoća i optimalna veličina jezičkog uzorka. *Psihologija*, *41*(1), 35-51.

Kostić, Đ. (1999). *Frekvencijski rečnik savremenog srpskog jezika*. Beograd: Institut za eksperimentalnu fonetiku i patologiju govora i Laboratorija za eksperimentalnu psihologiju.

Kostić, Đ. (2001). *Korpus srpskog jezika*. Beograd: Institut za eksperimentalnu fonetiku i patologiju govora i Laboratorija za eksperimentalnu psihologiju,

Manning, C. D., & Schütze, H. (2000). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

McDonald, S., Shillcock, R., & Brew, C. (2001). Low-level predictive inference in reading: Using distributional statistics to predict eye movements. *AMLAP 2001*. Saarbrucken, Germany.

Milin, P. (2004). *Probabilistički pristup određivanju gramatičkog statusa reči i kognitivne strategije u obradi jezika*. Beograd: Filozofski fakultet. Doktorska disertacija.

Milin, P. (2005). Istraživanja jezičkih fenomena pomoću računarskih simulacija obrade prirodnog jezika. *Empirijska istraživanja u psihologiji – XI*, Beograd.

Milin P., & Ilić, N. (2003). Text as Binary Sequence: A Case of Characteristics Constant of Text. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora; 10th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 47-53).

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91–117.

Yang, D. H., Lee, I. H., & Cantos, P. (2002). On the Corpus Size Needed for Compiling a Comprehensive Computational Lexicon by Automatic Lexical Acquisition. *Computers and the Humanities*, *36*(2), 171-190.

REZIME

STABILNOST SINTAGMATSKIH DISTRIBUCIJA VJEROVATNOĆA

Strahinja Dimitrijević
Odsjek za psihologiju, Univerzitet u Banjoj Luci

Aleksandar Kostić
Odeljenje za psihologiju, Univerzitet u Beogradu
Laboratorija za eksperimentalnu psihologiju, Univerzitet u Beogradu

Petar Milin
Odsek za psihologiju, Univerzitet u Novom Sadu
Laboratorija za eksperimentalnu psihologiju, Univerzitet u Beogradu

Ovo istraživanje je imalo za cilj da uspostavi kriterijume za određivanje optimalne veličine korpusa, koja može da obezbijedi stabilne zavisne vjerovatnoće morfoloških i/ili sintagmatskih tipova. Optimalnost veličine korpusa definisana je u odnosu na najmanji uzorak koji proizvodi distribuciju vjerovatnoća koja je jednaka distribuciji dobijenoj na kvazi-populaciji, velikom uzorku koji generiše stabilne vjerovatnoće. Ova posljednja distribucija označena je kao ciljna distribucija.

Da bi se uspostavili gore pomenuti kriteriji, varirana je: a. veličina uzorka (eksponencijalnim povećavanjem od veličine 2^5 *riječi), b. broj riječi u nizu (dvije riječi i tri riječi), c. način uzorkovanja (slučajno odabrane riječi ili kontinuirani tekst) i d. položaj riječi-mete u nizu (za niz od dvije riječi: ispred ili iza konteksta; za tri riječi: ispred, u sredini ili na kraju niza). Distribucije zavisnih vjerovatnoća dobijene na manjim uzorcima korelirane su sa ciljnim distribucijama, pri čemu je koeficijent korelacije tumačen kao indeks sličnosti, a ne indeks povezanosti dvije distribucije. Veličina uzorka na kojoj distribucija vjerovatnoća dostiže maksimalnu korelaciju (r=1) s ciljnom distribucijom, smatrana je optimalnom. Istraživanje je izvršeno na Korpusu srpskog jezika.*

Za bigrame, optimalna veličina uzorka za slučajno odabrane riječi je 65.000 riječi, i 281.000 riječi za trigrame. Nasuprot tome, potrebni su mnogo veći uzorci kontinuiranog teksta da bi se postigla stabilnost: 810.000 riječi za bigrame i 868.000 riječi za trigrame. Faktori koji su prouzrokovali ove razlike ostaju nejasni i traže dodatnu empirijsku provjeru.

Ključne riječi: *korpusna lingvistika, kvantitativna lingvistika, optimalna veličina uzorka, zavisne vjerovatnoće, srpski jezik.*

RAD PRIMLJEN: 12.12.2008.