IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 5, 2019, accepted October 10, 2019, date of publication October 17, 2019, date of current version October 28, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2947898

Dynamic N-Gram System Based on an Online Croatian Spellchecking Service

GORDAN GLEDEC, (Member, IEEE), RENATO ŠOIĆ, (Student Member, IEEE), AND ŠANDOR DEMBITZ, (Member, IEEE)
Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb 10000, Croatia
Corresponding author: Gordan Gledec (gordan.gledec@fer.hr)

ABSTRACT As an infrastructure able to accelerate the development of natural language processing applications, large-scale lexical n-gram databases are at present important data systems. However, deriving such systems for world minority languages as it was done in the Google n-gram project leads to many obstacles. This paper presents an innovative approach to large-scale n-gram system creation applied to the Croatian language. Instead of using the Web as the world's largest text repository, our process of n-gram collection relies on the Croatian online academic spellchecker *Hascheck*, a language service publicly available since 1993 and popular worldwide. Our n-gram filtering is based on dictionary criteria, contrary to the publicly available Google n-gram systems in which cutoff criteria were applied. After 12 years of collecting, the size of the Croatian n-gram system reached the size of the largest Google Version 1 n-gram systems. Due to reliance on a service in constant use, the Croatian n-gram system is a dynamic one. System dynamics allowed modeling of n-gram count behavior through Heaps' law, which led to interesting results. Like many minority languages, the Croatian language suffers from a lack of sophisticated language processing systems in many application areas. The importance of a rich lexical n-gram infrastructure for rapid breakthroughs in new application areas is also exemplified in the paper.

INDEX TERMS Croatian language, Heaps' law, language modeling, lexical n-gram, n-gram system comparison.

I. INTRODUCTION

Lexical n-grams are nowadays an important data infrastructure in many areas of natural language processing (NLP), machine learning, text analytics, and data mining [1]. Many technologies take advantage of large-scale language models based on huge n-gram systems derived from gigantic corpora. ''More words and less linguistic annotation'' is a trend well expressed in [2]. The trend is strictly followed in the research presented here.

Besides English [3], structured big data are the privilege of a dozen languages most advanced in NLP, those treated in the Google n-gram project [4]–[6]. Abundant linguistic data collection is a prerequisite for large-scale language modeling, but in many cases, it is hardly a feasible step in the machine processing of minority languages such as Croatian, which belongs to the subfamily of South Slavic languages and has approximately 4.5 million users, or less than 0.1% of the world's population.

The associate editor coordinating the review of this manuscript and approving it for publication was Zhangbing Zhou.

It is clear that an enormous English or Chinese text corpus cannot be comparable in size with a Croatian one due to differences in the numbers of language users. However, statistical machine translation or speech recognition asks for language models of comparable size in order to produce the desired effectiveness. This means the n-gram system, from which language models are derived, in a minority language must be enriched to approximately the size of n-gram systems for world major languages. It is not feasible to do so by following conventional methods of corpus creation as explained in Section II, based on a Croatian example. Therefore, several months after the appearance of [7] at the *Google AI Blog*, in May 2007, we took advantage of the already operating Croatian online academic spellchecker *Hascheck* and started collecting n-grams for $n = 1, 2, \ldots, 5$. In January 2013, collection was extended to a so-called higher-order n-gram system ($n = 4, 5, 6, 7$) in order to make it comparable with the Japanese n-gram system [5], the only Google system where n goes up to 7.

Furthermore, a convergence of 4- and 5-gram counts in the basic system with an increase in the corpus size intrigued us.

We wanted to see what would happen with them because we expected divergence based on data from [3] and [4]. Our approach proved to be an economic one [8] because it did not require extra manpower or other resources, as opposed to the Web as Corpus approach described in [9].

Regular monthly updates of the n-gram system allowed for Heaps' law modeling of n-gram count growth. It is worth noting that our n-gram system is unique among systems of comparable size because it contains lexical n-grams in the entire range of their frequencies, from *hapax legomena* to extremely frequent token sequences, which is a prerequisite for reliable modeling of this kind [10]. Heaps' law parameters have changed over time, but in the last several years, they have stabilized. This means the obtained Heaps' law functions are a powerful tool for prediction of n-count behavior where Croatian language is concerned. Furthermore, the obtained results may provide guidance, at least for Indo-European languages, for how far the value of n should progress when developing a language model. Jurafsky and Martin, in the very recent draft of the third edition of their famous textbook *Speech and Language Processing* [11] do not mention language models for $n > 5$ and this paper, in Section V, offers an explanation. The obtained n-grams have been used to convert our spellchecker from a conventional to a contextual one, the first of that kind in Slavic languages. Furthermore, they proved to be a crucial data infrastructure for rapid prototyping of Croatian speech technology tools.

The remainder of this paper is organized as follows: section II offers insight into creation of the conventional Croatian text corpora and section III describes our spellchecking service as a tool for unconventional text corpus creation. Section IV explores our n-gram system characteristics, while section V describes the application of Heaps' law to the Croatian n-gram system. Section VI discusses our n-gram system's applications to real-world problems, while section VII concludes the paper.

II. CONVENTIONALLY CREATED CROATIAN CORPORA

The first attempt to create a modern corpus for the Croatian language was made in December 1998 [12]. The outcome of the project was the Croatian National Corpus, a representative corpus of contemporary Croatian standard language written texts published since 1990. The corpus is composed of fiction, nonfiction, and mixed texts and its latest version contains 234 Mtokens, accessible via web interface [13].

The second corpus was the Croatian Language Corpus, with approximately 100 Mtokens, a result of the project of the Institute of Croatian Language and Linguistics and funded by the Croatian Ministry of Science, Education, and Sports. It was assembled in 2005 from selected Croatian language texts covering various functional domains and genres. It includes literature and other written sources from the period of the beginning of the final shaping of the standardization of Croatian language, i.e., from the second half of the 19th century. It consists of fundamental Croatian literature (novels, short stories, drama, poetry), nonfiction, scientific publications from various domains, university textbooks, schoolbooks, translated literature from outstanding Croatian translators, online journals and newspapers, and books from the pre-standardization period of Croatian language that are adapted to today's standard Croatian [14]. Research on the project stopped around 2011 according to the copyright claim on the project web site.

Creation of Croatian Web corpora hrWaC started in 2011 and was described in [9] and [15]; in [15], text corpora for three very similar languages were created by crawling the websites under the .hr top level domain (TLD), along with Bosnian (.ba) and Serbian (.rs) TLDs. These are publicly available under the CC-BY-SA license [16].

Creation of hrWaC was subject to criticism in [17] and its shortcomings can be summarized as follows:

1) Much of the content in the Croatian language isn't hosted on .hr TLD, but on generic TLDs such as .com; .org or .com.
2) Crawling the web utilizes network and system resources and involves stripping HTML tags and character set conversion; hrWaC still contains non-textual content and encoding errors.
3) Crawled text requires spellchecking as an additional step to make the corpus clean - one third of hrWaC is user-generated content (e.g., forums, blogs) that is abundant in spelling errors and uses non-diacriticized text.
4) Finally, it is difficult to find the size of the latest version of hrWaC, whose initial size was 1.2 Gtokens [15]. The project's Croatian web page [16] shows information about the version 2.0 size (1.9 Gtokens), while the web page of the Slovenian project partners [18] shows information about the version 2.2 size (1.4 Gtokens) without any explanation for why the newer version has approximately 25% fewer tokens than version 2.0.

Section IV demonstrates that the Croatian n-gram system described in this paper was created from a corpus at least four times larger than any corpus mentioned in this section. No better proof is needed to show that an unconventional approach, well devised and designed, outperforms all the conventional approaches that were developed and tested for major world languages when a minority language is concerned.

III. ABOUT THE SPELLCHECKER

Hascheck is the core engine of our online spellchecking service. It has been operating since 1993, nowadays as *Ispravi.me* (in English *Correct.me*), and is available at *https://ispravi.me/*. *Hascheck* is a "child" of two old Bell Labs spellcheckers, *typo* [19] and *spell* [20]:

1) From *typo*, it borrowed the peculiarity concept expressed through n-graph analysis, where n is extended to $n = 5$ and applied not only to analyzed text, but to the whole known language, which allows a very precise selection of unknown Croatian common word-types.

2) From *spell* it borrowed affix analysis, adapted to Croatian language, which is very inflected and includes many irregularities compared to English. The affix analysis is applied to both unknown word- and name-types.

These two features, integrated as an internet service, made *Hascheck* a learning system. As of June 2019, the service has been accessed from 1.4 million IP addresses belonging to 180 TLDs and by approximately a million browsers (HTTP cookies). Croatian IP-address ranges are dominant in these numbers. Out of all IP addresses allocated to Croatia, 42% are registered in *Hascheck's* log-files. The increase of service popularity is well expressed by the figures in Tab. 1. Croatian TLDs contribute 87% of the traffic; Bosnia and Herzegovina, in which Croats are a constitutive nation, contribute 6%; Germany, Serbia, and the US contribute less than 1% each, and so on.

TABLE 1. Service popularity increase exemplified by traffic in May.

	No. of texts processed	Corpus [tokens]
May 2007	7,361	2,014,783
May 2008	17,320	4,598,905
May 2009	31,906	7,218,386
May 2010	72,623	20,592,369
May 2011	116,338	28,747,996
May 2012	171,637	45,604,059
May 2013	202,005	54,580,366
May 2014	226,166	64,809,133
May 2015	268,866	80,995,990
May 2016	306,206	92,786,865
May 2017	410,587	114,325,867
May 2018	490,853	130,603,095
May 2019	593,391	162,366,903

Hascheck spellchecks the received text in real time. After processing, it logs statistics, including raw, non-filtered n-grams, and performs learning. All user inputs are removed from our file system after the processing is complete.

The outcome of learning is the update of the dictionary, i.e., improvement of the spellchecker's functionality. In order to keep the dictionary as precise as possible, the learning is supervised by humans.

Hascheck's architecture is explained in detail in [21]. Here we give only the size of our dictionary, which has changed significantly since [21] was published. The dictionary is organized into three word-list files that contained as of 1 June 2019:

1) 1,053,791 common Croatian word-types;
2) 963,318 case-sensitive elements (e.g., proper and other names, abbreviations, acronyms, etc.);
3) 70,528 English words, the only file that has not changed significantly in size since the service started.

English word list was included in Croatian spellchecking because English, as the modern *lingua franca*, is often mixed with Croatian in contemporary Croatian writing.

From May 2007, when n-gram collecting began, until June 2019, *Hascheck* has processed 28 million texts, which form a corpus of 7.5 gigatokens (Gtokens).

IV. CROATIAN N-GRAM SYSTEM CHARACTERISTICS

A lexical n-gram database system must be as consistent as possible in order to enable creation of an applicable language model. In our n-gram system, this means each n-gram is built only of tokens recognized by *Hascheck* as real words and encountered in Croatian writing. We use the following criteria to select n-grams for our system from a raw n-gram collection:

1) The only acceptable token separator is blank.
2) Non-words and non-numeric tokens containing punctuation marks inside a token are treated as n-gram terminators.
3) Two numeric tokens cannot be successive n-gram constituents.
4) N-gram derivation ends when a punctuation mark followed by a blank space is found as the terminating n-gram character.
5) Semi-compounds are treated as bigrams (e.g., Indo-European is treated as Indo European).
6) Inside and at the end of a numeric token, certain punctuation marks are allowed.
7) No subsequent numeric tokens are allowed in order to exclude table contents from n-gram derivation.

These criteria made selected n-grams for $n \geq 2$ almost purely lexical. Due to reliance on *Hascheck*, we were not forced to apply any cutoff criteria on n-grams, as was done in [3]–[5], and [6]. We have left cutoffs for converting our n-gram system into language models suitable for applications. The update of the n-gram database is performed monthly. Further details about n-gram system creation and maintenance are given in [8].

From the beginning until January 2013, only the basic n-gram system ($n = 1, \ldots, 5$) was maintained, but then we started collecting and filtering n-grams for our higher-order n-gram system ($n = 4, \ldots, 7$). Sizes of our two n-gram systems, reached at the beginning of June 2019, are presented in Tabs. 2 and 3 respectively.

It is worth noting that the changes in n-gram counts presented in Tab. 3 correlate almost perfectly with Pearson's correlation coefficient $R = 0.99$.

A. HAPAX LEGOMENA

Hapax legomenon is a token or n-gram for $n > 1$ that is encountered only once within any context, be it a corpus, a book, or a text. For large corpora, about 40% to 60% of all tokens (unigrams), including misspellings and typos, are *hapax legomena* [22]. As stated in the introduction, the existence of *hapax legomena* is important in Heaps' law modeling. Since our n-gram system does not contain misspellings or typos as n-gram constituents, it is important to show the portion occupied by *hapax legomena* (Tab. 4). The figures are given for comparison with *hapax*, *dis*, etc. *legomenon* shares published for corpora in which misspellings/typos were considered.

Unigrams (1-grams) are not included in Tab. 4 because they deserve special attention. The *Hapax legomenon* share for all

TABLE 2. Croatian basic n-gram system compared with the three largest google systems.

	10 Indo-European lang. WaC 1.3 Ttokens	Chinese WaC 883 Gtokens	English WaC 1.025 Ttokens	Croatian Hascheck's corpus 7.5 Gtokens
1-grams	95,998,281	1,616.150	13,588,391	5,804,402
2-grams	646,439,858	281,107,315	314,843,401	272,926,023
3-grams	1,312,972,925	1,024,642,142	977,069,902	950,463,566
4-grams	1,396.154,236	1,348,990,533	1,313,818,354	1,444,610,384
5-grams	1,149,361,413	1,256,043,325	1,176,470,663	1,524,743,571
Total	4,600,926,713	3,912,399,465	3,795,790,711	4,198,547,946

TABLE 3. Croatian higher-order n-gram system compared to Japanese n-gram system.

	Japanese WaC 255 Ttokens	Croatian *Hascheck's* corpus 6.4 Gtokens
4-grams	707,787,333	1,244,039,116
5-grams	776,378,943	1,302,171,248
6-grams	688,782,933	1,181,230,573
7-grams	570,204,252	1,020,261,479
Total	2,743,153,461	4,747,702,416

TABLE 4. Hapax legomena shares in croatian n-gram systems.

Basic system		Higher-order system	
2-grams	41.8%		
3-grams	50.6%		
4-grams	56.4%	4-grams	56.9%
5-grams	59.8%	5-grams	60.2%
		6-grams	61.9%
		7-grams	62.7%

unigrams is 29.2%, less than what was quoted in [22]. This is a result of absence of misspellings/typos among them. Since numeric tokens dominate among unigrams (52.2% of all Croatian unigrams are numeric), the share of lexical unigrams is much smaller: 10.8%. As previously stated, for $n \geq 2$, our n-grams are almost purely lexical. Among 2-grams only, 2.5% contain numerical tokens, while among n-grams with $n > 2$, this portion is fewer than 2%.

Unigrams are also subject to a learning process. *Hascheck's* word-guessing algorithm, described in [21], suggests to human supervisors which tokens are worth considering as potential new word- or name-types. After checking them in the context they appear in, the supervisor makes a final decision. If a suggestion is accepted, the new word- or name-type is marked as valid and, according to standards of conventional lexicography, its lemma and basic form(s) of irregular inflexions, when these are not present in *Hascheck's* dictionary, are also added to the dictionary. A consequence of this way of learning is that approximately 12% of the dictionary content has no confirmation in unigrams because many Croatian lemmas and basic irregular inflexions have never actually appeared in the processed corpus.

B. POSSIBLE IMPLICATIONS OF THE EXTREME FREQUENCY DIFFERENCES TO MACHINE TRANSLATIONS

In highly inflected languages, various word-forms derived from the same lemma may occur with extreme frequency differences: one form is very rare, while another occurs rather frequently. The same can be extended to n-grams, too. In [23] it was shown that rare 2-grams may have higher informativeness than their more frequent counterparts. If statistical machine translation does not take that into account, mistranslations are possible. We conducted a small and rather simple experiment in June 2019 with *Google Translate*, using a Croatian-to-English translation tool to demonstrate this. Here are the mistranslation examples with Google's translation in square brackets:

1) *Renato Šoić je radio s Ivanom Perić.*
   [Renato Šoić worked with Ivan Perić.]
2) *Ivan Perić je radio s Renatom Šoić.*
   [Ivan Perić worked with Renato Šoić.]

In both starting Croatian sentences, the male sentence subjects (Renato Šoić and Ivan Perić) had female collaborators whose names were Ivana Perić and Renata Šoić, respectively. This is not recognizable in instrumental forms of personal names (instrumental is the case in Croatian that follows the preposition *s* = *with*) because both Renato/Renata and Ivan/Ivana become Renatom/Ivanom in instrumental case. This is recognizable by the absence of the suffix –em after the frequent South Slavic surname ending –ić in the original sentences. Perićem/Šoićem would mean the collaborators are male, while Perić/Šoić means the collaborators are female. However, *Google Translate* converted the sex of the collaborating persons to male!

Why? We don't know the details of how *Google Translate* works, but we know it is a statistical machine translation tool and, starting from that, we can make some assumptions based on Google search responses. Google searches with "Ivanom Perićem"/"Ivanom Perić" were giving much better results in the male query case, while searching with "Renatom Šoićem"/"Renatom Šoić" led to no exact matching, only a question, "Do you mean: Renato Šoić?" in the case of the female query.

This gives some hints about why *Google Translate* behaved the way it did in the upper examples. Those translation bugs are easily fixable, so we hope Google experts will do it soon.

V. HEAPS' LAW APPLIED TO CROATIAN N-GRAMS

In a language with W words in its vocabulary, W^n word sequences of length n, n = 1, 2, …, N, are possible. For a given N, the number of all possible sequences is equal to

$$W + W^2 + \ldots + W^N = W \cdot \frac{1 - W^N}{1 - W} \quad (1)$$

This expression (1) is known as a geometric progression. In the case of $W > 1$, it belongs to a family of exponential growth models.

However, natural languages impose many restrictions on sequencing words into meaningful expressions. The main restriction, when n-gram count growth is considered, is limited sentence length. Average sentence length in Google WaCs is as follows:

1) 10 Indo-European languages—8.67 tokens;
2) Chinese—8.65 tokens;
3) English—10.77 tokens;
4) Japanese—12.74 tokens.

These numbers explain why n-gram count reaches its extreme for n = 4 in the first three cases, while in the Japanese case the extreme is reached at n = 5, identical to that of the Croatian case (see Tabs. 2 and 3). Average sentence length in *Hascheck's* corpus is 9.71 tokens. Hence, it is closer to the first three cases than to the Japanese one, so the similarity of n-gram count behavior in Croatian and Japanese may be explained by differences in n-gram selection criteria: dictionary vs. cutoff.

Zipf's law [24], which states that the frequency of tokens in a large corpus of natural language is inversely proportional to the token rank, can be extended to lexical n-grams as well [25]. If a phenomenon obeys Zipf's law, it also obeys Heaps' law [10]. The law connects a ''vocabulary'' size (V), in terms of the number of different n-grams in it, with the size of the corpus (t) in which the n-grams are present:

$$V(t) = \alpha \cdot t^{\beta}. \quad (2)$$

Parameters α and β are free parameters to be determined empirically. The parameter α is strongly language dependent, while β is much less language dependent. In the case of β, a condition $0 < \beta \leq 1$ must be satisfied. Calculating these parameters with sufficient data in hand is a straightforward task using a spreadsheet software program: corpus size and corresponding n-gram count are used to generate a two-dimensional chart on which the trendline calculation with ''power'' option is applied, along with selecting ''displaying equation on chart''; α and β are displayed in the equation.

Since many NLP applications use the noise channel model, which relies on n-gram statistics, it is worth knowing how n-gram count changes as n increases. Furthermore, especially for linguists (in linguistics, the law is called Harden's law), it is interesting to know for which n Heaps' law turns from a convex curve to linear dependence. The turn indicates a change of n-gram properties from the geometrical (statistical) to the algebraic (linguistic) category, and hence from something that is of interest to predominantly engineers to something that is of interest to predominantly linguists.

In [8], we have presented a figure, copied here as Fig. 1, that implies the 4-grams are the richest n-grams in Croatian, and no n-grams, n > 4, can ever overcome them.

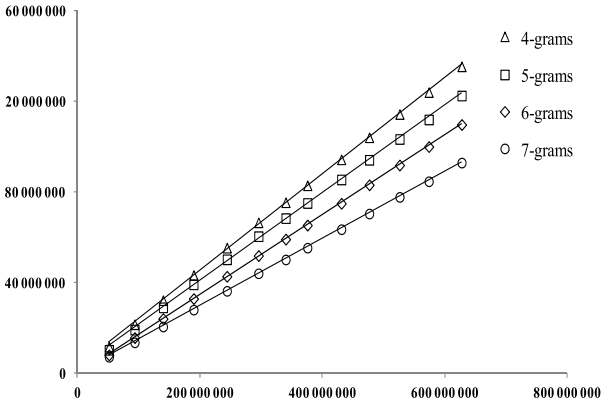


FIGURE 1. Behavior of n-gram counts derived from the higher-order n-gram system in the year 2014 [8]; x-axis: corpus size, y-axis: number of n-grams.

That assertion is a result of reduction of n-gram count growth, caused by 4- and 5-gram β values very close to 1 (see the values presented in Tab. 4 for the higher-order system), to linear dependence, which should not have occurred for reasons presented in the discussion and calculations at the end of this section.

In [26], an independent study of n-gram count behavior, for n = 1, …6, where n-grams were derived from English and French Wikipedia-based corpora, up to 1 Gtoken each, is presented. According to data presented there, n-gram count increases with n, hence the richest are 6-grams. It is worth noting that n-grams in [26] are defined similarly to those in our work. However, there is a problem.

Tab. 5 was produced by reproducing the first two columns of Tab. 1 in [26]. The data were used to create Fig. 2. in this paper.

TABLE 5. English unigram count from [26].

Corpus size [token]	No. of 1-grams
2,226,162	171,011
4,450,249	275,142
8,955,079	446,746
18,006,731	728,634
35,771,592	1,186,891
72,677,601	1,966.084
140,275,807	3,155,397
245,492,006	4,718,348
490,846,877	7,783,551
981,996,022	12,813,557

With English text corpora, typically α is between 10 and 100, and β is between 0.4 and 0.6 [27]. The same statement, referring to unigrams, may be found in many

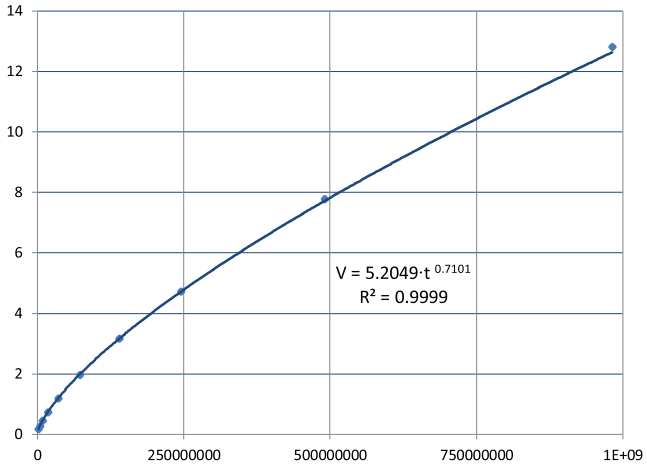


FIGURE 2. Heaps' law derived from Tab. 5; x-axis: corpus size in tokens, y-axis: number of unigrams in millions.

other sources. These conditions are not satisfied in Fig. 2 ($\alpha = 5.2, \beta = 0.71$). Either English Wikipedia is an extremely atypical English corpus, or there are some problems with data in [26]. At the least, it is hardly believable that a Gtoken English corpus can produce a similar number of unigrams as its Ttoken counterpart presented here in Tab. 2. After all, in the Croatian 1 Gtoken corpus there are only 2.23 million unigrams and the Croatian language has many more word-forms for a lemmatized word than does English.

From the data collected so far, we have obtained the Heaps' law parameter values for Croatian n-grams presented in Tab. 6. Correlation between empirical data and the corresponding functions had always $R > 0.9999$. This makes Heaps' law a reliable tool for prediction of n-gram count behavior in the future. Heaps' law functions tell something very important about the past, too. Although Heaps' law parameters for 4- and 5-grams differ significantly in two systems, the corresponding functions intersect almost at the same point, in both cases close to 2 Gtokens. This finding poses a challenge to Croatian WaC approach proponents to test it independently.

TABLE 6. Heaps' law parameters.

	Basic system		Higher-order system	
	α	β	α	β
1-grams	117.2	0.4754		
2-grams	35.86	0.6963		
3-grams	6.237	0.8284		
4-grams	1.235	0.9182	0.572	0.9517
5-grams	0.516	0.9587	0.249	0.9906
6-grams			0.184	1
7-grams			0.158	1

Heaps' law for 6- and 7-grams is reduced to linear functional dependency, which means the n-grams for $n > 5$ are whole sentences, or significant parts of sentences. This is not interesting for statistical language modeling, in which an uncertainty is always needed. The future should tell whether the parameter β will stay equal to 1 in these cases or will change to a "normal" $\beta < 1$. Intersections of "normal" and "abnormal" Heaps' law functions in the higher-order system are also interesting because they offer insight regarding the relation between convex curve and linear Heaps' law dependencies:

1) 4-gram function:
   - intersects 6-gram function at = 15.8 Gtokens;
   - intersects 7-gram function at t = 482 Gtokens;
2) 5-gram function:
   - intersects 6-gram function at t = 95 Ttokens;
   - intersects 7-gram function at t = 4 Ztokens (10^{21} tokens).

The intersection point between 4- and 6-gram functions should be confirmed or refuted in the next few years of our Croatian n-gram collecting process. The intersection between 4- and 7-gram functions is outside the range of any foreseeable Croatian text corpus. The others are outside of any human text corpora range. All books ever published in the world form a corpus of approximately 20 Ttokens [28].

VI. OUR APPLICATIONS BASED ON THE N-GRAM SYSTEM

A. APPLICATION IN THE SPELLCHECKER

Starting with the *Office 2007* release, Microsoft began to offer context-sensitive proofreading for a number of languages. It deals with real-word errors (whether one should write *to* or *too*, for example), undetectable by conventional spellcheckers. Although the application has rather poor context-sensitivity since many real-word errors pass through without warning, one can trust its suggestions. If a Microsoft contextual spellchecker says it is wrong, then it is wrong. Something similar happened to *Hascheck*.

Based on Croatian n-grams, $n = 2, \ldots, 7$, we have detected so far approximately 700 different grammatical and stylistic (pleonastic) patterns of real-word combinations that are almost surely contextual errors with a relatively high appearance rate (by "pattern" we mean a lemmatized n-gram form, which is in Croatian generally expandable by inflections). Dealing with pleonasms, expressions that use more words or word-parts than is necessary in clear and simple writing, like "mutual cooperation" (is there any unilateral cooperation?), is unusual in spellchecking. Thanks to our colleagues, Croatian linguists, and especially their corpus of the most frequent pleonasms in Croatian [29], we could do this rather quickly. *Hascheck* suggests corrections for grammatical errors, but only warns about pleonasms because these may sometimes be used as a kind of emphasis.

An example of a grammatical error that *Hascheck* corrects today can be drawn from the following data. The Croatian phrase *s obzirom na to da*, meaning "considering that," has 410,827 occurrences in basic 5-grams. Its subphrase, *obzirom na to da*, has 419,644 occurrences in basic 4-grams. Any subphrase appearance without the leading proposition *s* = *with* is wrong. Hence, 8,817 phrase appearances, or 2.1% of all its intended usages, were erroneous. The proportion is sufficiently high to take care of that error.

Many grammatical error cases we successfully dealt with are still not implemented into our online service because the implementation would cause an unacceptable server response time. An example is presented in [30].

B. APPLICATIONS IN SPEECH TECHNOLOGIES

Croatian language belongs to a group of under-resourced European languages in main contemporary NLP application areas: machine translation, speech synthesis, and speech recognition [31]. The under-resourcing is especially emphasized in speech processing, which is monolingual, contrary to machine translation.

The first research and development of speech technologies for the Croatian language was done at Carnegie Mellon University (CMU), Pittsburgh, PA, USA, motivated by the needs of the US Army personnel located in the Balkans at that time [32]. Since the US Army's priorities changed drastically after 11 September 2001, the project ended without delivering an applicable system. This was an additional motivation to us to continue where our CMU colleagues had stopped. Because of its relative simplicity, speech synthesis was addressed first by using Festival and FestVox tools, as was done in [32].

The n-gram system was used to construct a short, 17-minute-long training sample, which covers as great a diversity as possible of Croatian phoneme combinations and acoustic transitions within 270 short sentences. This was possible because of Croatian phonemic orthography and resulted in a publicly available Croatian speech synthesizer called *HascheckVoice* (*https://hascheck.tel.fer.hr/voice/*), which was developed over several months of individual work [33]. Although the system produces intelligible Croatian speech, it still calls for many improvements. Many difficulties arise from the ''simplicity'' of Croatian phonemic orthography, in which accent marks are very seldom used. It makes solving some technological problems simple, as mentioned above, but poses serious challenges for some others. An example is the existence of a great number of homographs-heterophones in Croatian writing whose pronunciation can be disambiguated only from context of appearance. Our progress in facing these challenges is presented in [34] and [35].

After our initial success in speech synthesis, we turned to automatic speech recognition. At the beginning of the first semester of the academic year 2012/2013, students in the graduate course Natural Language Processing at the Faculty of Electrical Engineering and Computing of the University of Zagreb, Croatia, were assigned the task of developing a large vocabulary continuous speech recognition (LVCSR) system for Croatian language from scratch.

In a three-month period, students created an applicable acoustic model by using a sample of 657 sentences in which 4,145 very carefully chosen Croatian word-types appear. The choice was governed by the idea of covering the greatest possible diversity of Croatian phoneme combinations and acoustic transitions within a small sentence sample. The sentences were recorded by 15 non-professional speakers, 4 female and 11 male students attending the course, in a moderately noisy environment. This produced a 16-hour-long training set for acoustic modeling. After adding the 15,000 most frequently used Croatian word-types to the words represented in the training set, a vocabulary of 15,396 words, which covers over 75% of Croatian word usage, was created. The vocabulary served for derivation of a language model from Croatian n-grams. Before that, a pronunciation dictionary in the Sphinx-4 dictionary format was compiled. Finally, they incorporated the acoustic and language model into Sphinx-4 and obtained an applicable Croatian LVCSR system that recognizes freely chosen utterances reasonably well [36], expressed by a relatively small word-error-rate, WER = 15%. Professional systems of the same dictionary size are treated as well designed if their WER $\leq$ 10%.

VII. CONCLUSION

The paper demonstrates how it is possible to produce a maintainable and upgradable linguistic data infrastructure for serious language modeling in a minority language used by less than 0.1% of the world population. Instead of resorting to World Wide Web crawling for purposes of corpus creation, we relied on an existing language service, the Croatian online spellchecker *Hascheck*, for data collecting. It proved to be an economical and reliable method for obtaining a large-scale lexical n-gram system, comparable in size with the largest publicly available Google n-gram-systems. The only disadvantage of our approach to n-gram system creation was the long period needed to collect data. However, according to our experience, researchers dealing with NLP technologies for under-resourced languages can compensate for lack of many thing missing only by time.

Nobody creates n-gram systems for purely *l'art pour l'art* motivations. We started collecting n-grams twelve years ago, looking ahead. Now we can tell that our expectations have even been exceeded. *Hascheck* became a contextual spellchecker with functionality comparable to those of the most recent contextual spellcheckers. Contextual spellchecking is a privilege of just a handful of languages, among which Croatian, measured by number of users, is the smallest. With a mature n-gram system at hand, rapid breakthroughs in Croatian's ''virgin'' speech processing capabilities became easily feasible, as has been demonstrated in this paper.

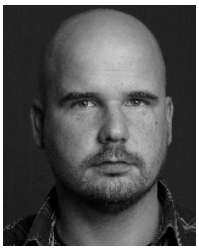
Looking further ahead in time, our 25-year-old decision to make *Hascheck* an online spellchecker has proven beneficial for the future. Although people still think online spellchecking is an auxiliary form of the service, the immense amount of data needed to become able to offer serious computational proofreading, comparable to human proofreading, points to cloud computing as a very probable future form of the service. Something similar should happen with other data-driven NLP applications, including machine translation, especially in the case of speech-to-speech translation.

REFERENCES

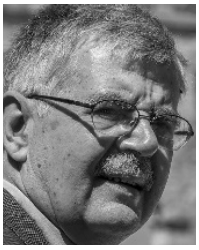
[1] J. Pueyo and J. A. Quiles-Follana, ''Trends in natural language processing and text mining,'' *Upgrade*, vol. 11, no. 3, pp. 33–39, Jun. 2010.

[2] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intell. Syst.*, vol. 24, no. 2, pp. 8–12, Mar./Apr. 2009.
[3] T. Brants and A. Franz, (Sep. 19, 2006). *Web 1T 5-Gram Version 1*. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2006T13
[4] T. Brants and A. Franz. (Oct. 20, 2009). *Web 1T 5-Gram, 10 European Languages, Version 1*. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2009T25
[5] T. Kudo and H. Kazawa. (Apr. 16, 2009). *Japanese Web N-Gram Version 1*. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2009T08
[6] F. Liu, M. Yang, and D. Lin. (Apr. 19, 2010). *Chinese Web 5-Gram Version 1*. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2010T06.
[7] T. Brants and A. Franz. (Aug. 3, 2006). All Our N-Gram are Belong to You. Google Research Blog. Accessed: Aug. 2019. [Online]. Available: https://ai.googleblog.com/2006/08/all-our-n-gram-are-belong-to-you.html
[8] Š. Dembitz, G. Gledec, and M. Sokele, "An economic approach to big data in a minority language," *Procedia Comput. Sci.*, vol. 35, pp. 427–436, Sep. 2014. doi: 10.1016/j.procs.2014.08.123.
[9] N. Ljubešić and T. Erjavec, "hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene," in *Proc. 14th Int. Conf. Text, Speech Dialogue*. Berlin, Germany: Springer, Aug. 2011, pp. 395–402. doi: 10.1007/978-3-642-23538-2_50.
[10] D. C. van Leijenhorst and T. P. van der Weide, "A formal derivation of Heaps' Law," *Inf. Sci.*, vol. 170, nos. 2–4, pp. 263–272, Feb. 2005. doi: 10.1016/j.ins.2004.03.006.
[11] D. Jurafsky and J. H. Martin, "N-gram language models," in *Speech Language Processing*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, Sep. 2018. Accessed: Aug. 2019. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf
[12] M. Tadić, "Building the croatian national corpus," in *Proc. 3rd Int. Conf. Lang. Resour. Eval. (LREC)*, Las Palmas, Spain, May 2002, pp. 441–446.
[13] University of Zagreb, Faculty of Humanities and Social Sciences. *Croatian National Corpus v3.0*. Accessed: Jul. 2019. [Online]. Available: http://filip.ffzg.hr
[14] Institute of Croatian Language and Linguistics. *Croatian Language Repository*. Accessed: Jul. 2019. [Online]. Available: http://riznica.ihjj.hr
[15] N. Ljubešić and F. Klubička, "{bs, hr, sr} WaC—Web corpora of Bosnian, Croatian and Serbian," *Proc. 9th Web as Corpus Workshop (WaC)*, Gothenburg, Sweden, Apr. 2014, pp. 29–35. doi: 10.3115/v1/W14-0405.
[16] Faculty of Humanities and Social Sciences, University of Zagreb. *hrWaC—Croatian Web Corpora*. Accessed: Jul. 2019. [Online]. Available: http://nlp.ffzg.hr/resources/corpora/
[17] J. Šnajder, S. Padó, and Ž. Agić, "Building and evaluating a distributional memory for croatian," in *Proc. 51st Ann. Meeting Assoc. Comput. Linguistics*, Sofia, Bulgaria, vol. 2, Aug. 2013, pp. 784–789.
[18] *Common Language Resources and Technology Infrastructure (CLARIN), Slovenia*. Accessed: Jul. 2019. [Online]. Available: https://www.clarin.si/noske/run.cgi/corp_info?corpname=hrwac
[19] R. Morris and L. L. Cherry, "Computer detection of typographical errors," *IEEE Trans. Prof. Commun.*, vol. PC-18, no. 1, pp. 54–64, Mar. 1975. doi: 10.1109/TPC.1975.6593963.
[20] M. D. McIlroy, "Development of a spelling list," *IEEE Trans. Commun.*, vol. COM-30, no. 1, pp. 91–99, Jan. 1982. doi: 10.1109/TCOM.1982.1095395.
[21] Š. Dembitz, M. Randić, and G. Gledec, "Advantages of online spellchecking: A croatian example," *Softw.-Pract. Exper.*, vol. 41, pp. 1203–1231, 2010. doi: 10.1002/spe.1037.
[22] A. Kornai, *Mathematical Linguistics*. London, U.K.: Springer-Verlag, 2008, p. 73. doi: 10.1007/978-1-84628-986-6.
[23] D. Jurić, M. Banek, and Š. Dembitz, "Informativeness of inflective noun bigrams in croatian," in *Proc. 6th KES Int. Conf., KES-AMSTA*, Dubrovnik, Croatia, Jun. 2012, pp. 114–123. doi: 10.1007/978-3-642-30947-2_15.
[24] G. K. Zipf, *Human Behavior and the Principle of Least Effort*. Cambridge, MA, USA: Addison-Wesley, 1949.
[25] L. Q. Ha, E. I. Sicilia-Garcia, J. Ming, and F. J. Smith, "Extension of Zipf's law to words and phrases," in *Proc. 19th Int. Conf. Comput. Linguistics*, Taipei, Taiwan, vol. 1, Sep. 2002, pp. 1–6. doi: 10.3115/1072228.1072345.
[26] J. F. Silva, C. Goncalves, and J. C. Cunha, "A theoretical model for N-gram distribution in big data corpora," in *Proc. IEEE Int. Conf. Big Data*, Washington, DC, USA, Dec. 2016, pp. 134–141. doi: 10.1109/BigData.2016.7840598.
[27] N. Boccara, *Modeling Complex Systems*, 2nd ed. New York, NY, USA: Springer, 2010, p. 376. doi: 10.1007/978-1-4419-6562-2.
[28] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden, "Quantitative analysis of culture using millions of digitized books," *Sci.*, vol. 331, no. 6014, pp. 176–182, Jan. 2011. doi: 10.1126/science.1199644.
[29] L. Hudeček, K. Lewis, and M. Mihaljević, "Pleonasms in the Croatian standard language," (in Croatian), *Rasprave Instituta hrvatski jezik jezikoslovlje*, vol. 37, no. 1, pp. 41–72, Dec. 2011.
[30] I. Srdić and G. Gledec, "Contextual spellchecking based on N-grams," in *Proc. 28th CECIIS, FOI*, Varaždin, Croatia, Sep. 2017, pp. 29–33.
[31] META-NET White Paper Series. (2013). *Key Results and Cross-Language Comparison*. [Online]. Available: http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison
[32] A. W. Black, R. D. Brown, R. Frederking, R. Singh, J. Moody, and F. Steinbrecher, "TONGUES: Rapid development of a speech-to-speech translation system," in *Proc. 2nd Int. Conf. Hum. Lang. Technol. Res. (HLT)*, San Francisco, CA, USA: Morgan Kaufmann, 2002, pp. 183–186.
[33] R. Šoić, "Speech synthesis for Croatian language using Festival," (in Croatian), M.S. thesis, Fac. Elect. Eng. Comput., Univ. Zagreb, Zagreb, Croatia, 2010.
[34] R. Šoić, M. Vuković, and Ž. Car, "Enabling text-to-speech functionality for websites and applications using a content-derived model," *EAI Endorsed Trans. Ambient Syst.*, vol. 17, no. 13, p. e3, May 2017. doi: 10.4108/eai.17-5-2017.152547.
[35] R. Šoić, P. Skocir, and G. Jezic, "Agent-based system for context-aware human-computer interaction," in *Proc. 12th Int. Conf. Agents Multi-Agent Syst., Technol. Appl. (KES-AMSTA)*, Gold Coast, QLD, Australia, Jun. 2018, pp. 34–43. doi: 10.1007/978-3-319-92031-3_4.
[36] D. Bajo, D. Turković, and Š. Dembitz, "Rapid prototyping of a croatian large vocabulary continuous speech recognition system," in *Proc. IARIA*, Red Hook, NY, USA: Curran Associates, Nov. 2014, pp. 13–18.

GORDAN GLEDEC was born in Zagreb, Croatia, in 1973. He received the Ph.D. degree from the University of Zagreb, in 2004. He was a Researcher on more than ten national and international scientific projects and projects with industry partners. He published more than 40 conference and journal articles. He is a coauthor of the chapter Spellchecker in *Wiley Encyclopedia of Computer Science and Engineering*. His research interests include the Internet technologies, web usability, human–computer interaction, and natural language processing.



RENATO ŠOIĆ was born in Zagreb, Croatia, in 1985. He received the master's degree from the University of Zagreb, in 2010. He participated in many industrial projects from different domains. He is currently a Research Assistant with the Department of Telecommunications, Faculty of Electrical Engineering and Computing, University of Zagreb. He has coauthored four conference articles and two journal articles. His research interests include speech technologies and human–computer interaction in smart environments.



ŠANDOR DEMBITZ was born in Split, Croatia, in 1951. He received the Ph.D. degree in electrical engineering from the University of Zagreb, in 1993. He spent his professional career as a Teacher with the Faculty of Electrical Engineering and Computing, University of Zagreb. He is the author of Hascheck and initiator of Croatian n-gram system collecting. He is also the author of more than 100 conference papers and journal articles, as well as several book chapters. His research interest includes natural language processing. He is a member of IEEE Croatia Section. In 2014, he was a recipient of the IEEE Croatia Section Award for an outstanding engineering achievement in spellchecking.