The feasibility of a complete text corpus

Primož Jakopin
Slovenian Corpus Laboratory
Fran Ramovš Institute of Slovenian Language ZRC SAZU
Novi trg 4, 1000 Ljubljana, Slovenia
primoz.jakopin@uni-lj.si

Abstract
In the paper the annual increase in size of a complete text corpus of a single language, Slovenian, is estimated. It comprises the serial publications in Slovenian, monographs and pages, published on Internet. The estimate for the year 2000, based on 21,000 units of serial publications, 675,000 pages from 5,200 units of printed monographs, 377,000 pages from 5,500 units of unpublished monographs (mostly academic theses) and 300,000 pages on Internet is given at less than 1.5 billion words. An extension of the Law of legal deposit, which would also cover electronic versions of printed texts, is proposed. It is suggested that to make the idea of a complete corpus viable, it should be simple and profitable for the publishers to supply web versions of their publications alongside with printed ones.

1. Introduction
Advancement of computer technology in recent years has moved the solution of some old challenges such as machine translation or speech recognition to the more foreseeable, though still elusive future, and has made some more mundane wishes practically possible. One of them would be an establishment of complete national text corpora, consisting of all published works, either in printed form or on Internet. They all have an electronic base, stored on some media during the preparation, but are not nationwide systematically archived as their printed versions and more often than not get lost in time.

The practice of the legal deposit, a legal obligation of all publishers and distributors in a country to send one copy of each of their printed publications to the National library has been first introduced in 1534 in France. It is now operational in virtually all countries of the world and is being extended to other, non-printed publications as well. An important part of those are academic works, which exist just in a few copies, such as degree, postgraduate and doctoral theses; a copy is kept at least by the Department library of the Faculty and descriptions are included in national library indexes. Other publications from this domain are publications available only in electronic form, such as CD ROM encyclopedias, dictionaries, atlases or other educational material. In several European countries, such as UK, the Netherlands, Scandinavian countries, they are also archived by national libraries (Jakac-Bizjak 2001). Archiving is regulated by voluntary agreements between publisher associations and national libraries; the relevant legislation, an extension of the Law of the legal deposit, is expected to be operational in the next few years (in Germany by 2003, for instance).

Works, published on the Internet, are usually freely available, at least for personal use, and their content is stored in website indexes, such as Google, Altavista or NAJDI.SI in Slovenia. They are frequently updated, often daily, and provide an insight into how wonderful things could be if all the relevant information would be available online.

2. Goal of the paper
An estimate of the size of a complete text output for Slovenian language, for a given time span, one year, is the main aim of this paper. Slovenia is a country at the northeastern corner of the Adriatic sea, halfway between Vienna and Florence, 20,000 sq km, 50 km of coastline; Slovenian, most western Slavic language, has about 2 million speakers.

The second aim is to determine which steps would be required and which conditions should be met to make an establishment of such corpus viable.

3. Paper texts
It is reasonable to expect that the vast majority of texts, produced in Slovenian, would see the light of day on paper. The annual report of the National library in Ljubljana (Krstulović & Bračič Fabjančič 2001) would be an obvious point to start. Yet it turned out that the evidence, especially tables concerning the number of serial publications, is not complete. The Internet tool, available to librarians and other users through the web page: http://cobiss.izum.si (COBISS = Co-operative Online Bibliographic System & Services, based in Maribor), has been used to obtain data about both serial publications and monographs. Year 2000 has been selected as the data about all publications from 2001 were, at the time of writing, not available yet.

3.1. Serial publications
In Table 1 on the next page the number of titles of serial publications with various publication frequency is given, followed, in the third column, by the total number of units, printed in 2000.

There are 8 dailies in Slovenian, 7 published in Slovenia: DELO (Ljubljana), the standard, Dnevnik (Ljubljana), Večer (Maribor), Slovenske novice (Ljubljana), a tabloid, Finance (Ljubljana), Ekipa (Ljubljana, sports), Dnevni bilten STA (Ljubljana, Bulletin of the Slovenian Press Agency) and 1 in Italy: Primorski dnevnik (Triest). There are 2 newspapers which appear twice a week: Gorenjski glas (Kranj) and Primorske novice (separate editions for Koper and Nova Gorica).
Number of titles increases sharply from weeklies on (103, 85 independent ones and 18 weekly supplements of the daily newspapers), peaking with 587 monthlies and 1168 serials, published once a year. There is even a serial, published every 3 years, which has been omitted. In all there are 2565 different titles with close to 21,000 published units every year.

In COBISS there are no data about the size of serial publications, and so an estimate will be required. An overview with number of titles and number of editions per year is given in Table 1.

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Title Count</th>
<th>Editions Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily</td>
<td>8</td>
<td>2480</td>
</tr>
<tr>
<td>Twice a week</td>
<td>3</td>
<td>930</td>
</tr>
<tr>
<td>Weekly</td>
<td>103</td>
<td>5356</td>
</tr>
<tr>
<td>Biweekly</td>
<td>68</td>
<td>1768</td>
</tr>
<tr>
<td>Twice a month</td>
<td>17</td>
<td>408</td>
</tr>
<tr>
<td>Monthly</td>
<td>597</td>
<td>6567</td>
</tr>
<tr>
<td>Bimonthly</td>
<td>132</td>
<td>792</td>
</tr>
<tr>
<td>Quarterly</td>
<td>261</td>
<td>1044</td>
</tr>
<tr>
<td>3 times a year</td>
<td>50</td>
<td>150</td>
</tr>
<tr>
<td>Semiannually</td>
<td>142</td>
<td>284</td>
</tr>
<tr>
<td>Annually</td>
<td>1168</td>
<td>1168</td>
</tr>
<tr>
<td>Biannually</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>2565</td>
<td>20955</td>
</tr>
</tbody>
</table>

Table 1: Serial publications in Slovenian, 2000

In the leading daily newspaper, DELO, around 80,000 words are published in every edition (6 days a week, 2 mill. words per month), other dailies are less than half its size. Monitor, the leading computer monthly, brings around 100,000 words in every number, the majority of other weeklies and monthlies is again less than half that size.

Therefore, it is reasonable to estimate the number of words in serial publications in Slovenian per year at no more than 20,955 times 50,000 words, i.e. at 1,047,750,000 words, or, to round the figure, at no more than 1 billion words.

3.2. Monographs

When using COBISS searching service in command mode it is possible to request just monographs for a given year, given language and, with kind help from administrators of the system, it was possible to override the usual export limit of 50 hits. Month by month, file by file, data on 14,051 documents have been obtained.

<table>
<thead>
<tr>
<th>Month</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>396</td>
</tr>
<tr>
<td>February</td>
<td>677</td>
</tr>
<tr>
<td>March</td>
<td>1082</td>
</tr>
</tbody>
</table>

Table 2: Monographs in Slovenian by month, 2000

In the field physical description (code 215a) 5,217 monographs had a number, followed by str. (pp. in English), and 5,523 a number, followed by f. (foils). The first figure compares favorably with the number of monographs, described in Slovenian Bibliography. Books. (4,805 units, Wagner 2001), and the second can be attributed to academic works, degree theses, postgraduate works and doctoral dissertations. A short overview is given in Table 3.

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Books (pp.)</td>
<td>5,217</td>
</tr>
<tr>
<td>Theses (f.)</td>
<td>5,523</td>
</tr>
<tr>
<td>Total</td>
<td>10,740</td>
</tr>
</tbody>
</table>

Table 3: Monographs, total and avg. number of pages

The longest book in 2000 (1,803 pp.) was European list of existing chemicals, a translation published by Ministry of health, and the longest thesis (811 f.) was a diploma thesis in social psychology: Motivation for altruistic behavior - on example of charity, by M. Žugič. As a standard book page contains about 2,000 characters or 300 words, the total number of words can be estimated at 315,617,700 or, rounded, at 315 mill. words.

4. Online texts

In recent years number of texts, available over Internet, has grown from a very small fraction of printed material to quite a reasonable extent. There are 2 text corpora, one open and one with restricted access (similar situation as with BNC and Bank of English) and several web indexes; 1 has surpassed the competition in 2001.

4.1. Text corpora

There are 2 online text corpora of reasonable size, with a search engine and an Internet interface in Slovenia, both operational since 1999. Nova beseda (New word in English, Jakopin 2001a), at the Internet address http://bos.zrc-sazu.si, is freely available, contains 50 mill. words of newspaper text (1998-2000) and fiction (1858-1996); the augmentation to 80 mill. words is in course.

The corpus is supplemented by 3 monolingual dictionaries and a service that returns lemmas and some POS information for a list of up to 2,500 words. It is operated by the author’s home institution, part of the Scientific Research Center of the Slovenian Academy of Sciences and Arts.
The second text corpus, FIDA, at the address http://www.fida.net, contains 100 mill. of mostly newspaper text, is operated by a consortium of 4 partners (2 commercial, 1 educational and 1 research institution); access for users outside the consortium costs around 500 Euros per annum.

Both corpora have broken the ice in the field of Slovenian corpus linguistics and have shown that it is possible to put up a sizeable text corpus with very reasonable means. Nova beseda has logged over 30.000 accesses to its pages in the past 2 years. The expertise gathered could be put to good use in compiling a complete corpus of Slovenian.

4.2. Slovenian web index

NAJDI.SI (najdi could be translated as find) at the address http://www.najdi.si, available since November 2000, by Noviforum Ltd. has established itself as the main search engine for Slovenian Internet with 2.5 million web pages in its index. Noviforum also operates the Hungarian web index with 4 million pages and the Croatian one with 1 million pages.

Its spider scans Slovenian Internet space at least once a week - information on how often particular web pages are being modified is kept and the pages that change frequently, together with pages linked from them, are visited daily (pages longer than 1 MB are ignored). A relevant index, even for pages with news, is assured this way. There were over 3 million web pages at the end of March, 2002, and after removing the duplicates (pages with a different URL yet with the same content), around 2,5 million have remained for indexing. Automatic language identification is performed, an algorithm with n-gram statistics (up to n = 5) for different languages is applied. It has been successful on 1,447,602 pages and the results are shown in Table 4.

Table 4: NAJDI.SI web page languages with frequencies

The share of Slovenian pages is at around 64%, of English around 34% and the remaining pages in 31 other languages take the remaining 2%. The algorithm needs a few lines of text to find a match. The pages without identified language are either too short, frame redirecting pages often have no text at all, or are pages with pictures (882.000 pages, among them 675.000 .JPG and 195.000 .GIF pages).

Not all the words from the pages are included in the index - stop list contains 80 units: 9 numbers, 30 English, 2 German, 36 Slovenian words and 3 abbreviations. From an index, kindly compiled by Samo Login from subcorpus of 968.762 pages, without a stop list, the values for the missing words in the main index have been interpolated, taking into account the share of particular language.

NAJDI.SI word list contains 7,591,414 word and non-word units (also see Jakopin 2001b) with a total frequency of 578,745,747. After the addition of 80 units from the stop list, the values for the missing words in the main index have been interpolated, taking into account the share of particular language.

NAJDI.SI word list contains 7,591,414 word and non-word units (also see Jakopin 2001b) with a total frequency of 578,745,747. After the addition of 80 units from the stop list, the values for the missing words in the main index have been interpolated, taking into account the share of particular language.

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Ivan Cankar} & \text{DELO newspaper} & \text{NAJDI.SI web index} \\
\hline
1. eyes & 3,152 & state & 1,915 & article & 2,107 \\
2. heart & 2,739 & year & 1,833 & page & 1,666 \\
3. face & 2,674 & time & 1,401 & day & 1,526 \\
4. hand & 2,619 & city & 1,315 & year & 1,267 \\
5. man & 2,351 & president & 1,173 & work & 1,252 \\
6. word & 1,645 & law & 1,026 & world & 1,073 \\
7. life & 1,570 & percent & 1,026 & time & 827 \\
8. head & 1,335 & day & 993 & law & 799 \\
9. people & 1,152 & end & 978 & group & 790 \\
10. path & 1,134 & people & 926 & contribution & 776 \\
11. night & 1,122 & tolar & 864 & system & 773 \\
12. mister & 1,113 & party & 799 & city & 717 \\
13. time & 1,080 & million & 793 & connection & 690 \\
14. cheek & 1,054 & group & 777 & data item & 680 \\
15. road & 1,036 & minister & 747 & school & 638 \\
16. window & 1,022 & enterprise & 741 & community & 608 \\
17. voice & 994 & government & 735 & right & 600 \\
18. mother & 958 & case & 698 & use & 559 \\
19. table & 937 & question & 697 & court & 568 \\
20. love & 915 & race & 672 & change & 556 \\
\hline
\end{array}
\]

Table 5: Top nouns in 3 corpora (per mill. running words)
If the NAJDI.SI index is taken as a good representation for the texts on Slovenian Internet, the amount of new texts coming from this source in one year can be estimated at no more than 150 million words, one third of its index that covers Slovenian language.

5. Size estimate

<table>
<thead>
<tr>
<th>SERIAL PUBLICATIONS</th>
<th>1,000,000,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monographs</td>
<td>315,000,000</td>
</tr>
<tr>
<td>Internet pages</td>
<td>150,000,000</td>
</tr>
<tr>
<td>Total</td>
<td>1,465,000,000</td>
</tr>
</tbody>
</table>

Table 6: Slovenian written output, words per year

An estimate, most likely in form of an upper bound, of the total Slovenian written output in one year, in words is shown in Table 6.

The figure is less than 1.5 billion words or, speaking in storage space, around 10.5 GB. It would fit onto 17 CD ROMs, onto 1 larger capacity DVD or onto a typical web server.

6. Feasibility

To get such a corpus rolling there are obviously no technology-related obstacles. It in itself would be a major achievement, a universal source of knowledge, a wonderful tool for all, for study, for research, for fun.

From institutional point of view, it could best be done by a united effort of three partners: National and university library as a keeper of that part of the national heritage by law ([http://www.nuk.uni-lj.si](http://www.nuk.uni-lj.si)), Noviforum Ltd. ([http://www.noviforum.si](http://www.noviforum.si)) as the search engine provider and the Institute of Slovenian language ZRC SAZU as the main lexicographic and research body of the language ([http://bos.zrc-sazu.si](http://bos.zrc-sazu.si)) and provider of technology for POS tagging and lemmatisation of text corpus.

To accomplish it, legislative steps, mentioned in the introduction, would however not be enough. Even a recommendation by the Ministry of Education, Science and Sport, which would force the publishers of state-sponsored publications to contribute their electronic versions would be difficult to implement. Change of climate in the publishing business would be required.

One of the two essential steps would be the success of a project such as the Open eBook ([http://www.openebook.org](http://www.openebook.org)) that would bring a universal tool, a widely adopted desktop publishing software that would produce both the printed version of the book and its web version in one go, without any additional effort.

The second step would be adoption of electronic books by the readers, by the users. If it would be profitable to put books on the web, that the publisher would obtain the sum comparable to that now paid to borrow a book in the public library, it would be done promptly and without hesitation.

To collect the publications, already on the web, in one large index, where only close context (up to 3 sentences) of an observed linguistic phenomenon or, for that matter, searched unknown fact, could be displayed, with a link to the publisher’s server, where the entire text could be read for a small sum, would make perfect sense and would make the complete text corpus of a language a realistic and unproblematic task.

7. Conclusion

In the paper the size of complete written text output of Slovenian language in one year has been estimated. It has been shown that building such a corpus would be, with slight extra advancement of computer technology into our daily lives, a straightforward job.

Let us hope that the time of electronic books will come soon, the day of joy for the corpus boy.

8. Acknowledgments

The author would like to thank Noviforum Ltd., Samo Login in particular, for providing the data about NAJDI.SI search engine and words in its index, Tjaša Pavletić-Lacko from the Slovenian National and university library for help in compiling the statistics about Slovenian serial publications and Matjaž Rebolj from the Library of Department of history, Faculty of Arts in Ljubljana, for a hand with data about Slovenian monographs.

9. References


