Digits to Words Converter for Slavic Languages in Systems of Automatic Speech Recognition

Josef Chaloupka

Technical University of Liberec, Liberec 461 17, Czech Republic,
josef.chaloupka@tul.cz,
WWW home page: https://www.ite.tul.cz/speechlabe/

Abstract. In this paper, a system for digits to words conversion for almost all Slavic languages is proposed. This system was developed for improvement of text corpora which we are using for building of a lexicon or for training of language models and acoustic models in the task of Large Vocabulary Continuous Speech Recognition (LVCSR). Strings of digits, some other special characters (%, €, $, . . .) or abbreviations of physical units (km, m, cm, kg, l, °C, etc.) occur very often in our text corpora. It is in about 5% cases. The strings of digits or special characters are usually omitted if a lexicon is being built or if the language model is being trained. The task of digits to words conversion in non-inflected languages (e.g. English) is solved by relatively simple conversion or lookup table. The problem is more complex in inflected Slavic languages. The string of digits can be converted into several different word combinations. It depends on the context and resulting words are inflected by gender or cases. The main goal of this research was to find the rules (patterns) for conversion of string of digits into words for Slavic languages. The second goal was to unify this patterns over Slavic languages and to integrate them to the universal system for digits to words conversion.

Keywords: digits to words converter, LVCSR, text processing

1 Introduction

Systems of automatic processing, recognition and synthesis of audio speech signal are practically used in many research areas at the present (2017) [1][2]. They are mainly systems for voice dictation to PC, voice controlled PC tools, voice-interactive dialogue systems, automatic broadcast programs transcription systems, text-to-speech synthesis systems, etc. There has been noticeable progress also while recognizing the speech of inflected languages, where the form or ending of some words changes because of nouns' declination, verbs' conjugation and adjectives' or adverbs' escalation. Slavic languages belong among these inflectional languages and they are used approximately by 293 millions people worldwide. Because of inflection much bigger lexicons have to be used during recognition of continuous speech than in uninflected languages (e.g. English). Also the speech models are much more extensive and therefore more difficult as for computational complexity.

In our laboratory of computer speech processing we have developed systems for automatic large vocabulary continuous speech recognition (LVCSR) that work in real time with lexicons containing more than 500 000 words [8]. These systems we originally developed for the Czech language (CZ) [3]. During the last ten years they were however adapted for Slovak (SK) [4], Polish (PL) [5], Russian (RU), Belorussian (BY), Ukrainian (UA), Serbian (RS), Croatian (HR), Slovenian (SL), Bulgarian (BG) and Macedonian languages (MK) [6]. For each language, there was created a Language Model (LM), lexicon and hybrid Acoustic Model (AM) based on triphones. Triphones are presented by combination of Hidden Markov Models (HMM) and Deep Neural Networks (DNN) [7]. A higher Word Recognition Rate (WRR) was achieved with HMM-DNN models during all our Automatic Speech Recognition (ASR) experiments than while using traditional models HMM-GMM (Gaussian Mixture Models). Average WRR in task of Voice Dictation to PC is higher than 98% and it is about 86% in task of broadcast programs transcription for all mentioned Slavic Languages.

To train a LM and create a lexicon, huge text corpora is necessary. We have used mainly internet resources from major newspapers to adapt our LVCSR system to a new Slavic language. However there is a problem with numbers that appear in text as strings of digits and not as strings of words. The second more-less similar problem is if we would like to train or to adapt new AM from audio database. Speech in audio signals is manually or semi-automatically transcribed into text by human annotators. The numbers occur again as strings of digits very often in text part of the audio databases.

A large number of Digits-to-Words converter tools exist e.g. for English at present but almost non for Slavic languages. The main problem is that there isn't any clear way how to create them. The task to create a Digit-to-Words converter tool for Slavic languages isn't unambiguous because Slavic languages are inflected according to case or gender and text forms of numbers can get many different inflected forms. A few studies exist but they aren't mostly described in English literature or it is relatively hard to find them. As the case may be these studies cannot be easily practically realized. Moreover these strings of digits can often be accompanied by abbreviations (mostly of physical units) and their pronunciation (transcription) depends on the previous number or the transcription of digits to words can depend on another word, e.g. the name of a month.

The first goal of this work was to create a tool for digits to words conversion. This tool should be able to translate strings of digits (and possible abbreviations) into text word form. In case that the transcription is not clear, the tool should not transcribe the text. The second goal was to create a generator for word strings and related abbreviations, alternatively also the names of months that exist (with possible alternatives and their probabilities). This generator should be used especially to train the language model – for example randomly generated decimal numbers and randomly generated main or minor patterns (if there are more possibilities). Both tools (systems) should be universally used for any Slavic language.

2 Patterns for Digits to Words Converter

Study of rules how to convert digits to words had to be made in the beginning of this research work. String of digits can be translated as cardinal, ordinal or decimal number. These strings can be part of a date. The converter was designed also for translation of abbreviations of physical units or special characters which are pre-connected with strings of digits. The translation of abbreviations or special characters depends on previous number. The study was performed for all examined Slavic languages. Universal patterns for conversion were searched in this study.

2.1 Cardinal Number Patterns

A string of digits is converted as a cardinal number if it does not contain character dot and if name of month, special character or abbreviations does not follow this string. The words for numbers 1 and 2 (or 3, 4 in SK) are inflected by gender in almost all Slavic languages. The converter does not translate these numbers because it is a very difficult task and we cannot solve this task at present. Numbers from 3 (5) to 20 are translated by XML–translation table (XMLtab). It isn't a good idea to generate numbers from 11 to 19 as units (1-9) plus 'teen' because there exist several exceptions in different Slavic languages.

Numbers from 21 to 99 are generated from connection words for Units (U) and Decades (D). There exist several patterns how to make it in Slavic languages, see table 1. There are only main patterns, other (minor) patterns are used in several Slavic languages, e.g. U_&_D in CZ. Words for decades (20, 30, ...) and hundreds (100, 200, 300, ...) are again saved in XMLtab. There exist again several exceptions here therefore decades and hundreds are not generated.

Table 1. Cardinal number patterns for numbers 21-99

Language	**Example**	**Pattern**
CZ	dvacet pět	D_U
SK	dvadsaťpäť	DU
PL	dwadzieścia pięć	D_U
RU	двадцать пять	D_U
BY	дваццаць пяць	D_U
UA	двадцять п'ять	D_U
HR	dvadeset i pet	D_&_U
RS	двадесет и пет	D_&_U
SL	petindvajset	U&D
BG	двайсет и пет	D_&_U
MK	дваесет и пет	D_&_U

The conversion of numbers higher than 999 is once again specific. Being gendered, all the higher scale names (thousands, millions, milliards, ...) follow the declension rules in different Slavic languages. They are most often 3 word forms of higher scale names (HSN). First is for one HSN, second from two to four HSN and third for more than four HSN, see example in table 2. BG and MK have only two word forms (as in English) – one HSN and more than one HSN. The reason is that BG and MK have relatively simple declension rules and they don't have cases. Word forms are saved in XMLtab. The same word is saved for second and third word form in case of BG and MK.

Table 2. Example of three (CZ) and two word forms (BG)

Digits	Words - CZ	Words - BG
1000000	jeden milion	един милион
2000000	dva miliony	два милиона
3000000	tři miliony	три милиона
4000000	čtyři miliony	четири милиона
5000000	pět milionů	пет милиона

The interesting thing is that two different large-number naming systems are used for different Slavic languages. They are long and short scale systems. Every new word-term higher than million is one million times larger than the previous term in long scale system and every new word-term higher than million is one thousand times larger than the previous term in short scale system. The long scale system is used in CZ, SK, PL, HR, RS and the short scale system in RU, UA, BY, BG and MK. There is one exception in the Slavic short scale system: The word for billion is replaced by word milliard.

2.2 Decimal Number Patterns

The string of digits is converted as decimal number if it contains character dot or comma (decimal separator) inside a string. The string is separated into two parts (Integer (I) and Fractional (F)) according to decimal separator. These two strings are translated as cardinal numbers and they are connected with a word depending on Slavic language, see table 4. The words comma (c), and (&) or 'whole' (w) are used in different Slavic languages.

Table 3. Word 'whole' in decimal numbers - SK

Digits	Words
0,5	nula celých päť desátin
1,5	jedna celá päť desátin
2,5 - 4,5	dve, tri, štyri celé pět desetin
5,5	päť celých päť desátin

The word 'whole' is inflected according to previous integer part, see table 3. There exist several exceptions in different Slavic languages therefore different word forms are saved in XMLtab for I—numbers (0, 1, 2, 3, 4 and more than 4).

The name of the last digit's place value (DN) can be used in decimal number conversion in several Slavic languages (e.g. tenths, hundredths, thousandths, ten-thousandths, hundred-thousandth, millionth). DN can be again inflected depending on resulting decimal number and different word forms are in XMLtab. Reading of decimal number, where integer part is zero, is specific in some Slavic languages, see example in table 5. Third pattern is the most common in the CZ.

Table 4. Decimal number patterns - example for 8.25

Language	Example	Pattern
CZ	osm celých dvacet pět setin	I_w_F(DN)
SK	osem celých dvadsaťpäť stotín	I_w_F(DN)
PL	osiem i dwadzieścia pięć setnych	I_&_F(DN)
RU	восемь целых двадцать пять сотых	I_w_F(DN)
BY	восем і дваццаць пяць сотых	I_&_F(DN)
UA	вiсiм цiлих i двадцять п'ять сотих	I_w_&_F(DN)
HR	osam zarez dvadeset i pet	I_c_F
RS	осам зарез двадесет и пет	I_c_F
SL	osem celih petindvajset stotite	I_w_F(DN)
BG	осем цяло и двайсет и пет стотни	I_w_&_F(DN)
MK	осум запирка дваесет и пет	I_c_F

Table 5. Different patterns for 0.25 - CZ

No.	Pattern	Words	English translation
1.	I_w_F(DN)	nula celá dvacet pět setin	zero point twenty five hundredths
2.	I_w_F	nula celá dvacet pět	zero point twenty five
3.	F(DN)	dvacet pět setin	twenty five hundredths

2.3 Ordinal Number Patterns

The string of digits can be an ordinal number if the last character is dot or if it precedes or follows other ordinal number (it can be date expression) or it precedes name of a month. There exist two patterns for string of digits to word conversion for different Slavic languages.

All words are Ordinal (AO) numbers or only Last word is Ordinal number (LO), see table 6. Combination of digit(s) and abbreviation is used in some Slavic languages, e.g. in RU - 1-й (первый - first). This isn't solved in our converter but it is solved by a look-up table in our other text pre-processing tool.

Table 6. Ordinal number patterns - example for 48

Language	**Example**	**Pattern**
CZ	čtyřicátý osmý	AO
SK	štyridsiatyôsmy	AO
PL	czterdzieści ósmy	AO
RU	сорок восьмой	LO
BY	сорак восьмы	LO
UA	всорок восьмий	LO
HR	četrdeset i osmi	LO
RS	четрдесет и осми	LO
SL	osmiinštirideset	LO
BG	четиридесет и осми	LO
MK	четириесет и осми	LO

2.4 Date Expression

The date can occur in the text in form of strings of digits (e.g. 14. 4. 2017) or as combination of string of digits with the name of month (e.g. CZ: 14. dubna 2017 or RU: 1 апреля 2017 года). We solve separately day together with month and year. Latin-derived names (in SK, RU, RS, SL, BG and MK) or older Slavic names (in CZ, PL, UA, BY and HR) are used as names of months in Slavic languages.

String of digits, which represented day or month, are always ordinal numbers in all Slavic languages and the ordinal numbers are inflected by case. Day is an ordinal number in genitive and month is ordinal number in nominative in CZ and SK. Both day and month are ordinals number in nominative in PL and both are genitive in HR.

Name of month instead of string of digits is more common in date expression in all other Slavic languages. The string of digits is detected as year if:

1. The name of the month precedes the string of digits.
2. Two short strings of digits precede the string of digits.
3. Word 'year' or its abbreviation precedes or follows the string of digits. The year expression is cardinal number in CZ, SK and SL or it is an ordinal number in all other mentioned Slavic languages.

2.5 Abbreviation to Words Conversion

Some special characters or mainly abbreviations of physical units which follow cardinal or decimal number are translated in our conversion system. There are integrated following special characters: ’%’ percent, ’ €’ euro, ’$’ dollar, and abbreviations of physical units: ’mm’ millimeter, ’cm’ centimeter, ’m’ meter, ’dm’ decimeter, ’km’ kilometer, ’km/h’ kilometer per hour, ’m/h’ meter per hour, ’km/s’ kilometer per second, ’km/h’ kilometer per hour, ’m/s’ meter per second, ’m/h’ meter per hour, ’g’ gram, ’dg(dkg)’ decagram, ’kg’ kilogram, ’ml’ milliliter, ’cl’ centiliter, ’l’ liter, ’°C’ degree Celsius.

Table 7. Example of three (CZ) and two word forms (BG)

Digits + Abbrev.	**Words - CZ**	**Words - BG**
1 km. (км)	jeden kilometr	един километър
2 km. (км)	dva kilometry	два километра
3 km. (км)	tři kilometry	три километра
4 km. (км)	čtyři kilometry	четири километра
5 km. (км)	pět kilometrů	пет километра

There exist three or two word forms for transcription of abbreviations or special characters into words, see table 7. The word form depends on previous number. First word form is for previous cardinal number one, second for previous cardinal numbers from two to four and third form for previous cardinal numbers higher than four or if the previous number is decimal.

3 Digits to Words Converter - System Overview

Designed system for digits to words conversion is relatively complex, see fig. 1. The input to the system is string of text, conversion patterns (described above) for selected Slavic language and XML script where translation table for cardinal, ordinal or decimal numbers, date and abbreviations is saved. The input text string is tokenized to short strings (words, strings of digits, abbreviations, etc.). In the first step, it is investigated if single short string (ShS) contains character dot (’.’). The system decides that ShS is decimal number if ShS contains dot, all other characters are digits and last character isn’t dot. The string of digits is separated to fractional and integer part according to dot. These two parts are translated as cardinal numbers and they are connected with word(s) which expresses separator – decimal mark (’whole’, ’comma’, etc.). A name of last digit place is added to the end of resulting word string if it is usual in the particular Slavic language. The subsystem for conversion of abbreviation is used if abbreviations follow the decimal number. Third word form of abbreviation is used always in such case, see chapter 2.5.

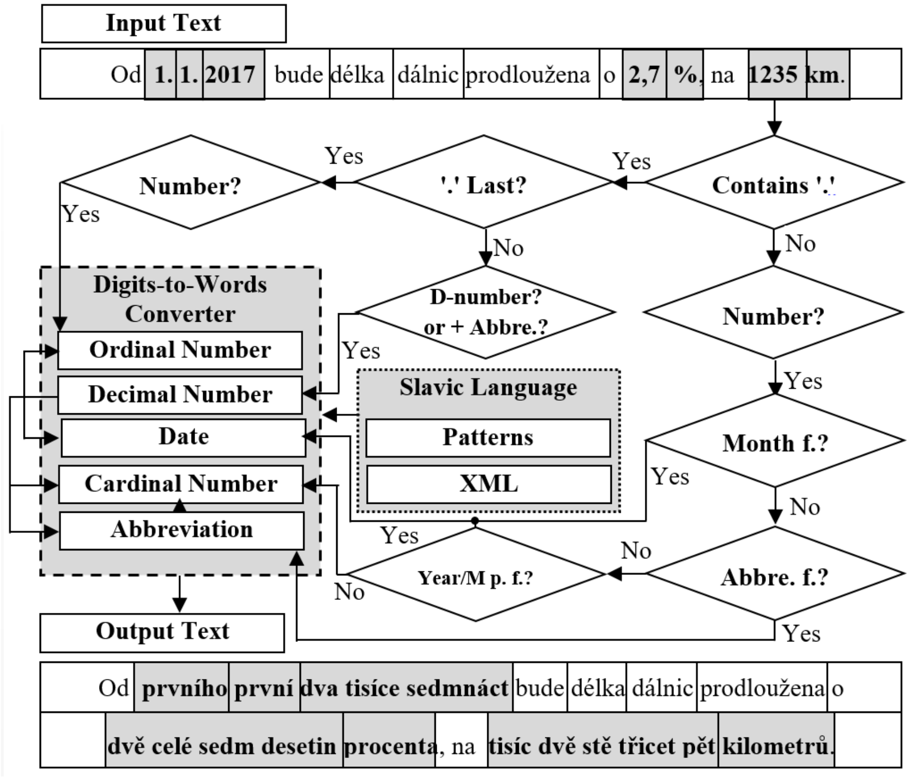

Fig. 1. The principle of digits to words converter.

The ShS is ordinal number if it contains dot as last character and all other characters are digits. This step is valid only for Slavic languages where ordinal numbers are written as strings of digits with dot in the end. It is specified in the set of patterns. The ordinal number is determined as part of date if name of month follows the ordinal number or if some other ordinal number follows the first ordinal number.

The ShS is assessed afterwards if any dot character doesn't exist in it and all characters are digits. A problem is that string of digits can be translated into several different word combinations. It depends on the context:

1. It could be part of date if the name of month follows.
2. It is verbal expression of year if word 'year' or its abbreviation precedes or follows the string of digits or if name of month precedes the string of digits. The string of digits is converted as cardinal or decimal number depending on Slavic language and selected (year) pattern.
3. The subsystem for abbreviation conversion is used if abbreviation of physical unit or some special character follows the string of digits. The string of digits

is converted as cardinal number and abbreviation (or special character) is converted to word form according to (abbreviation) pattern.

4. The string of digits is translated as cardinal number if all previously mentioned cases don't occurr.

Typical adjustment of main patterns for digits to words conversion system in e.g. CZ is:

D_U, GD 2, AO, I_w_F, DN Yes, ZERO Yes, Year 11 CN

where D_U: pattern for cardinal numbers – first Decades, second Units, connected by space. GD 2: digits 1 and 2 are not transcribed. AO: pattern for ordinal numbers – All Ordinal. I_w_F: pattern for decimal numbers – Integer part connected with Fractional part by Czech word 'whole'. DN Yes: parameter for decimal numbers – the name of the last digit's place value is used. ZERO Yes: parameter for decimal numbers – first word is zero if Integer part is zero. Year 11 CN: year is cardinal number (CN) and 11 indicates that years above 1000 and below 2000 are read as multiples of the word one hundred.

Simplified digits to words conversion system is presented on web pages: `http://kvap.tul.cz/slavic_symbols.php`. It is possible to convert only short strings in this simplified system and only main patterns are used here but the functionality of the system is maintained here.

4 Conclusion and Future Work

The complex system for digits to words conversion has been designed and created in this research work. This system is usable for almost all Slavic languages. The system is developed as a universal tool where only XML-like conversion table and pattern parameters are set for specific Slavic language. It is easy to change pattern parameters and set main or minor patterns which occurs in selected Slavic language. It is possible to enhance text corpora for training of the language model and to enhance text annotation of speech data for acoustic model training in LVCSR systems by adding translated forms from digits to words converter which is described above. The system for digits to words conversion is still being developed and improved with the help of native speakers.

Second function of the converter is to generate words connections from randomly chosen string of digits. This function can enhance and extend text corpora but it is necessary to find a probability of frequency occurrence of main and minor patterns firstly. We plan to investigate this probability for every single Slavic language from our audio databases in the near feature.

Acknowledgments. The research was supported by the Technology Agency of the Czech Republic in project no. TA04010199.

References

1. Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Laurent, C., Bengio, Y., Courville, A.: Towards end-to-end speech recognition with deep convolutional neural networks. In Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2016, pp. 410-414, ISSN: 2308-457X, 2016.
2. Deng, L., Li, J., Huang, J.-T., Yao, K., Yu, D., Seide, F., Seltzer, M., Acero, A.: Recent advances in deep learning for speech research at Microsoft. In IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP 2013, pp. 8604-8608, ISBN: 978-147990356-6, 2013.
3. Nouza, J., Zdansky, J., David, P., Cerva, P., Kolorenc, J., Nejedlova, D.: Fully Automated System for Czech Spoken Broadcast Transcription with Very Large (300K+) Lexicon. In Interspeech 2005, Lisboa, Portugal, pp. 1681-1684, ISSN 1018-4074, 2005.
4. Nouza, J., Silovsky, J., Zdansky, J., Cerva, P., Kroul, M., Chaloupka, J.: Czech-to-Slovak Adapted Broadcast News Transcription System. In Proceedings of the 9th Annual Conference of the International Speech Communication Association, (Interspeech 2008), pp. 2683-2686, ISSN: 1990-9772, September 22-26, Brisbane, Australia, 2008.
5. Nouza, J., Cerva, P., Safarik, R.: Cross-Lingual Adaptation of Broadcast Transcription System to Polish Language Using Public Data Sources, In: 7th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poland, pp. 181-185, ISBN 978-83-932640-8-7, 2015.
6. Nouza, J., Safarik, R., Cerva, P.: ASR for South Slavic Languages Developed in Almost Automated Way, In: Proc of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016), San Francisco, USA, pp. 3868 - 3872, DOI: 10.21437/Interspeech.2016-747, Scopus EID: 2-s2.0-84994385032, ISSN 2308-457X, 2016.
7. Dahl, G.E., Sainath, T.N., Hinton, G.E.: Improving deep neural networks for LVCSR using rectified linear units and dropout, In IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, ICASSP 2013, pp. 8609-8613, ISBN: 978-147990356-6, 2013.
8. Nouza, J., Blavka, K., Zdansky, J., Cerva, P., Silovsky, J., Bohac, M., Chaloupka, J., Kucharova, M., Seps, L.: Large-scale processing, indexing and search system for Czech audio-visual cultural heritage archives. In 2012 IEEE 14th International Workshop on Multimedia Signal Processing, MMSP 2012, pp. 337-342, ISBN 978-146734572-9, 2012.