Automatic Symbol Processing for Language Model Building in Slavic Languages

Josef Chaloupka

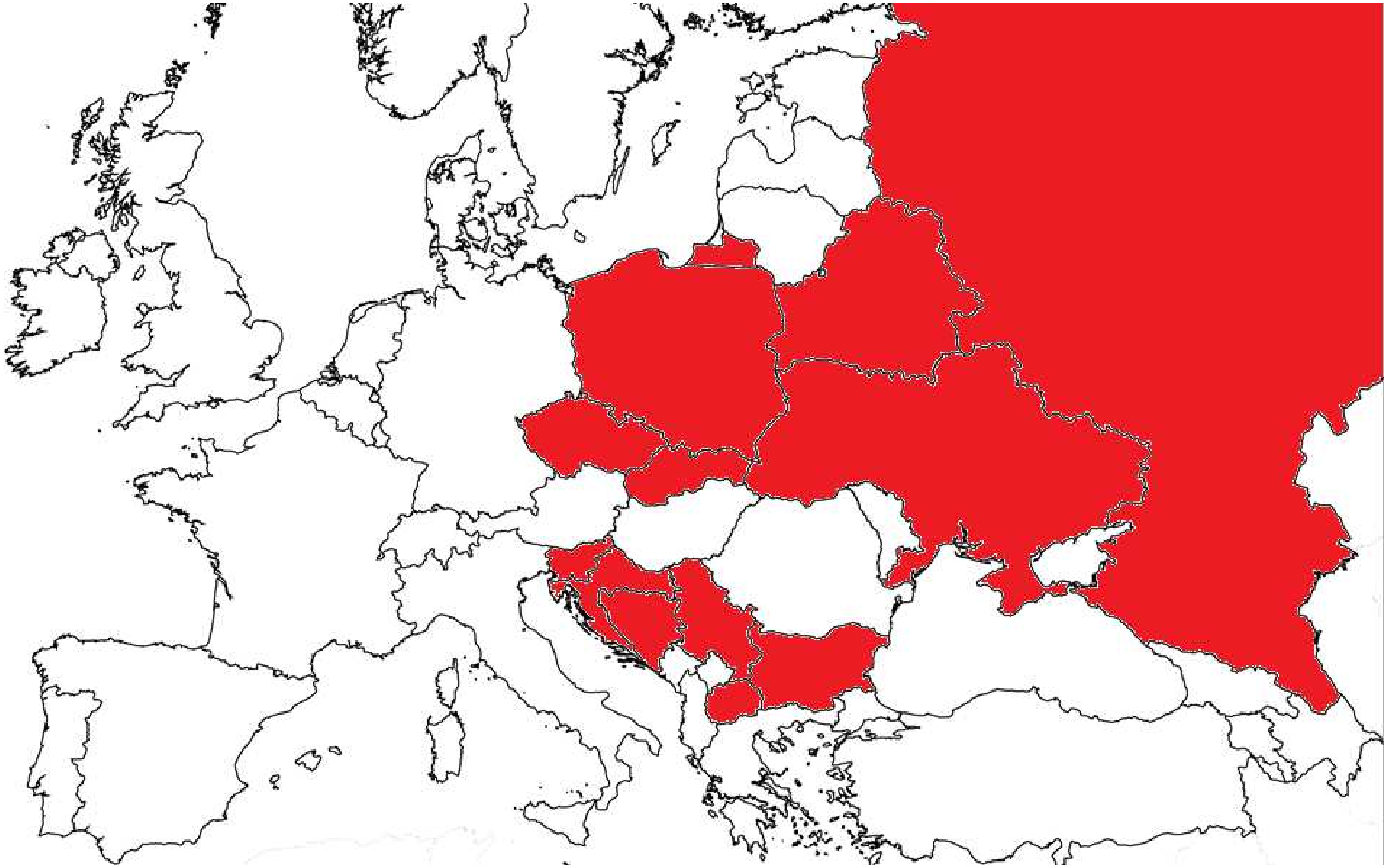
SpeechLab
Technical University of Liberec
Czech Republic

1. Motivation
2. ASR - Automatic Speech Recognition
3. Symbol to Text Translation
4. Cardinal Numbers
5. Ordinal Numbers
6. Decimal Numbers
7. Dates and Years
8. Combination of digits and abbreviation
9. DigitToWord Transcription Tool for Slavic Languages
10. Conclusion and Future Work

Motivation (1)

- Annotation of an existing automatic speech recognition system to a new language - we need a large corpus of texts to create a lexicon, a LM and a database of annotated recordings to train an acoustic model.
- The texts contains not only words but also some other symbols, mainly strings of digits, special characters and some frequent abbreviations of units.
- There is not a straightforward correspondence between their printed form and the spoken one.
- The main goal of this work was to develop efficient tools for automatic translation of symbols or symbolic terms to words for almost all Slavic languages.

Motivation (2)

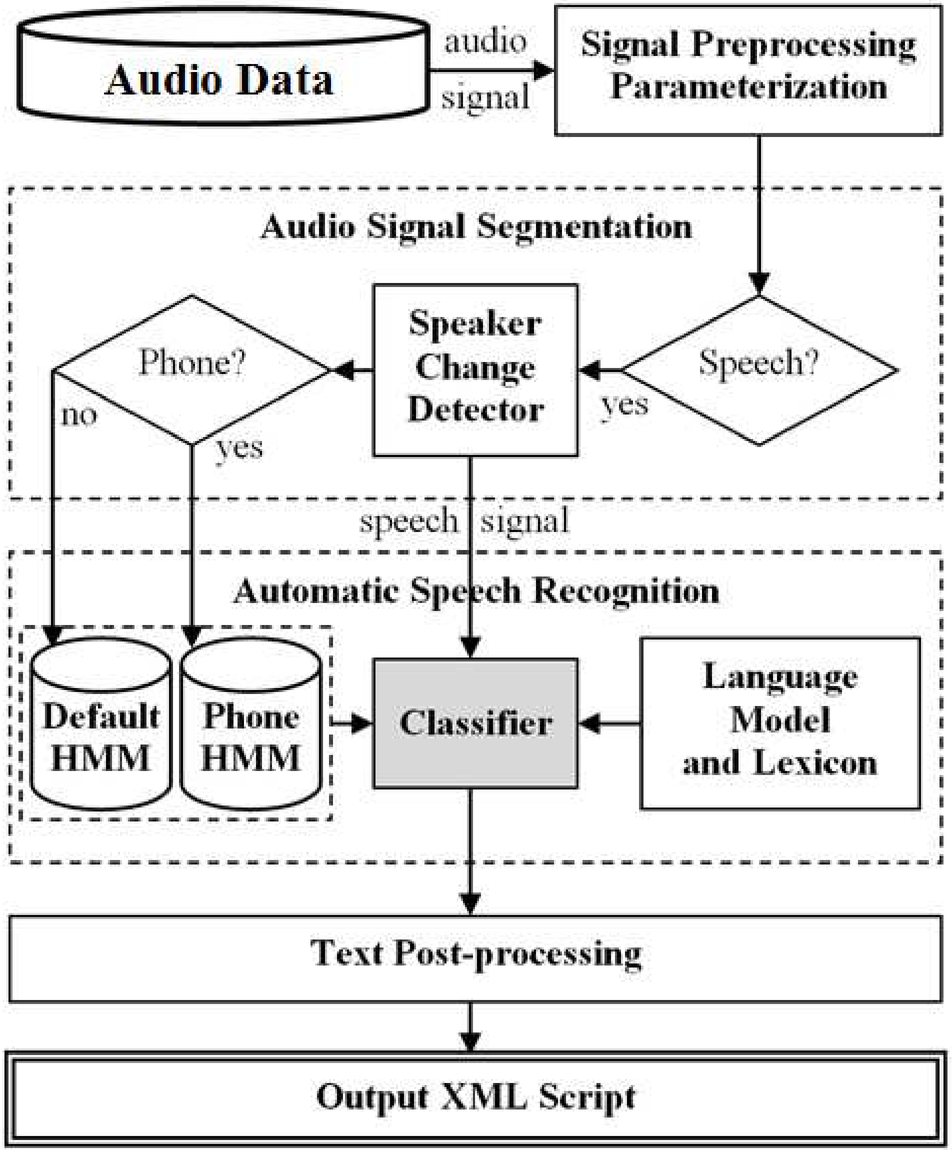


Processed Slavic languages: CZ - Czech, SK - Slovak, PL - Polish, RU - Russian, BY - Belarusian, UA - Ukrainian, HR - Croatian, RS - Serbian, SL - Slovenian, BG - Bulgarian and MK - Macedonian

Motivation (3)

- **Development and design of Tools**:
- To ensure that the words usually represented by their symbols appear in the lexicon.
- To translate symbols and symbolic terms to words (text pre-processing) and when needed also back to symbols (post-processing).
- To enhance the LM by adding translated forms into the corpus. The enhancement can be done also by generating randomly chosen digit and symbol strings using the rules and patterns applicable for each language.
- To enhance acoustic model training by better and more correct annotation of speech data, employing the transcription tools and allowing them to use alternative, minor or even colloquial rules of transcription and pronunciation.

ASR - Automatic Speech Recognition



Symbol to Text Translation

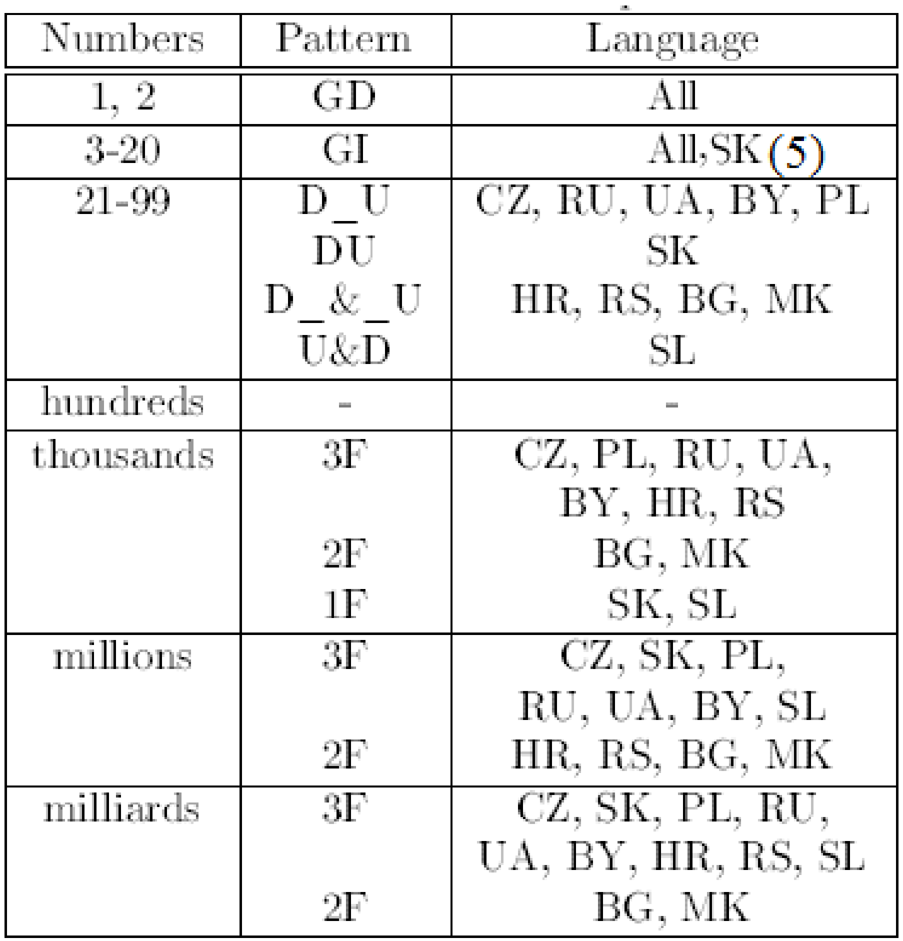
Cardinal numbers	**-256, …, -1, 0, 1, 2 , 3, …,1658, …**
Ordinal numbers	**1., 2., … 25., …**
Decimal numbers	**0,625, …, 128,69875 …**
Dates	**1. 5. 1945, 1. května 1945…**

- Cardinal/decimal number in combination with an abbreviation

10 mm; 4 cm; 125 m; 48 km; 52 km/h, 12 m/s; 12 g; 9 kg; 8 ml; 2 l; 100 %
-12,5 °C; 32 €; 50 $

Cardinal Numbers

- 14713324561412: četrnaest biljunov sedam stotina trinaest milijardi trista dvadeset i četiri milijuna pet stotina šezdeset i jedan tisuća četiri stotine dvanaest

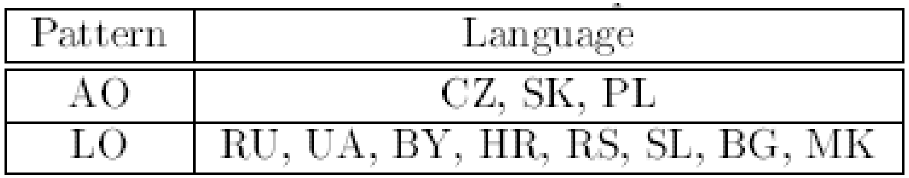


Numbers	Pattern	Language
1, 2	GD	All
3-20	GI	All,SK (5)
21-99	D_U DU D_&_U U&D	CZ, RU, UA, BY, PL SK HR, RS, BG, MK SL
hundreds	-	-
thousands	3F 2F 1F	CZ, PL, RU, UA, BY, HR, RS BG, MK SK, SL
millions	3F 2F	CZ, SK, PL, RU, UA, BY, SL HR, RS, BG, MK
milliards	3F 2F	CZ, SK, PL, RU, UA, BY, HR, RS, SL BG, MK

- GD - gender dependent
- GI - gender independent
- DU: the ten 'D - Decade' comes first, then the Unit 'U'
- D_U: the same as the first one but the space '_' is between the D and U
- D_&_U: the spaces and word 'and - &' (e.g. CZ - a, PL - i, SL - in) are between D and U
- U&D: U comes first, then D, joined together by the word 'and - &,
- Higher Scale Names (HSN - thousands, millions, milliard, …)
- 3F: three different word forms (1 HSN, 2-4 HSN and more than 4 HSN, e.g. CZ - jeden milion (one million), dva miliony (two millions), pět milionů (five millions)).
- 2F: two different word forms (1 HSN, more than 1 HSN)
- 1F: one word form (without declension)

Ordinal Numbers

- 1945. - devatenáctistý čtyřicátý pátý
- AO: All word number forms are Ordinal (e.g. PL - dwudziesty pierwszy (twentieth first))
- LO: only Last member is Ordinal, other words are cardinal numbers (e.g. HR - dvadeset i prvi (twenty and first)))



Pattern	Language
AO	CZ, SK, PL
LO	RU, UA, BY, HR, RS, SL, BG, MK

Decimal Numbers

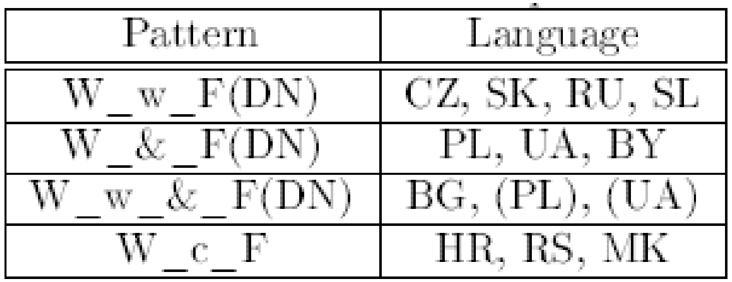
- 12,56987 - дванадцять цілих і п'ятдесят шість тисяч дев'ятсот вісімдесят сім стотисячних

- Decimal marks (separator) are used to separate the Integer part (I) from the Fractional part (F) of a decimal number. The decimal comma is used as decimal mark in all SLang.

- The decimal comma is read as whole (w) (e.g. SL - cela), comma (c) (e.g. HR - zarez) or as and (&) (e.g. PL - i).

- The word - name of the last digit's place value (DN) can be used in decimal number conversion (e.g. tenths, hundredths, thousandths, ten-thousandths, hundred-thousandth, millionth).

- The word whole (e.g. SL - cela) and DN are inflected in SLang.



Pattern	Language
W_w_F(DN)	CZ, SK, RU, SL
W_&_F(DN)	PL, UA, BY
W_w_&_F(DN)	BG, (PL), (UA)
W_c_F	HR, RS, MK

- W_w_F(DN): e.g. CZ - dvě celé šest setin - two whole six hundredths
- W_&_F(DN): e.g. PL - dwa i sześć setnych - two and six hundredths
- W_w_&_F(DN): e.g. BG - два цяло и шест стотни - two and six hundredths
- W_c_F: e.g. MK - два запирка нула еден - two comma zero six

Dates and Years (1)

- The date occurs frequently in text corpora in the form of strings of digits, e.g. 8. 5. 1945, or in a combination of strings of digits with the name of the month, e.g. PL - 8 maja 1945.
- The main format of date is day-month-year in all SLang.
- Latin-derived names of months are used in SK, RU, RS, SL, BG, MK; a set of older names for the months that differs from the Latin month names is used in CZ, PL, UA, BY, HR.
- There are two possible readings of date strings in SLang, e.g. '1. 1.' - 'first first' or 'first January'. The words for ordinal numbers are inflected by case (N - nominative, G - genitive, …) in the first approach ('first first'). There isn't any inflection by case in BG and MK, therefore the words stay in their basic form (B_B). There are three possible patterns, e.g.:

  G_N: e.g. CZ - prvního (first - genitive) první (first - nominative)

  N_N: e.g. PL - pierwszy (first - nominative) pierwszy (first - nominative)

  G_G: e.g. HR - prvog (first - genitive) prvog (first - genitive)

  B_B: e.g. BG - първи (first) първи (first)

Pattern G_N occurs in CZ, SK, N_N in PL, G_G in HR, RS and B_B in BG and MK.

Dates and Years (2)

- The string of digits is detected as a year in the text if:
- 1) the name of the month precedes
- 2) two short (1 - 31(12)) ordinal numbers precede
- 3) some form of word year (or abbreviation, e.g. BG - г.) precedes or follows the string
- The year is usually cardinal (CZ, SK, SL) or ordinal number (PL, RU, UA, BY, HR, RS, BG, MK).
- There are several exceptions for the transcription of the date and the year in different SLangs therefore our tools use only the main patterns (forms).
- CZ has one specific: years above one thousand and below two thousand are read as multiples of the word one hundred, e.g. 1900 - devatenáct set - nineteen hundred.

Combination of digits and abbreviation

- In our case, the abbreviation were special characters '€', '$' or '%' and abbreviations of physical units 'km', 'l', 'kg', '°C' or 'm/s'.

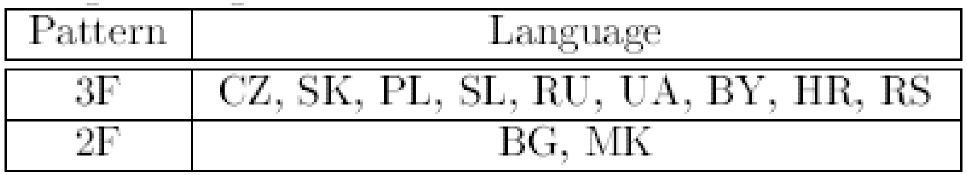
- If the number (a string of digits) before the abbreviation is a cardinal number - there are three (3F) or two (2F) word forms of abbreviation.

- The first word form is in combination with number one, second for numbers from two to four and third for numbers higher than four in 3F, e.g. SL - en kilometer (one kilometer) , dva kilometra (two kilometers), pet kilometrov (five kilometers).

- In 2F, the first word form is for abbreviation in combination with number one (singular) and second word form is for numbers higher than one (plural).

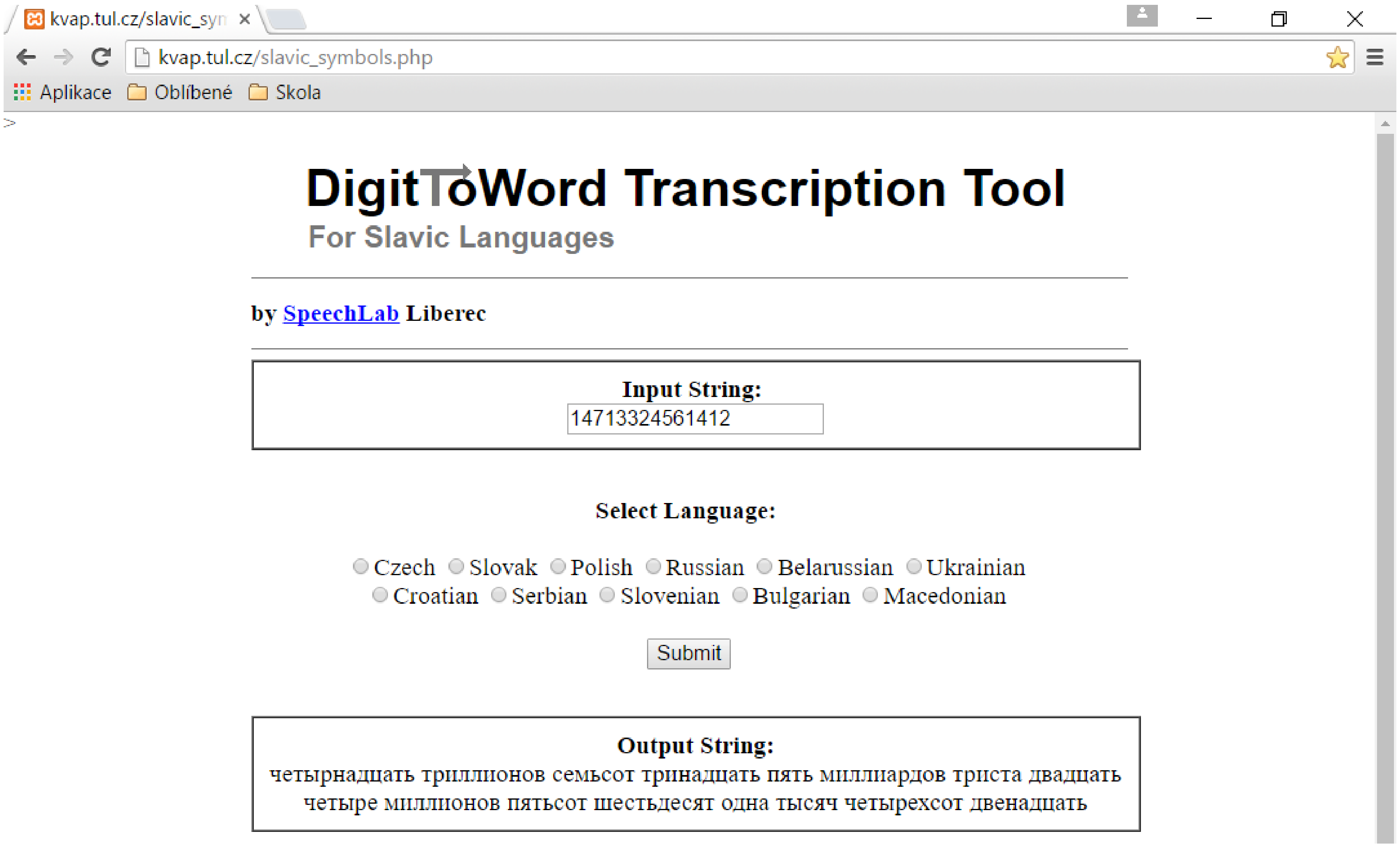
- Pattern 2F is the same as in English.

- There are several exceptions for the inflection of some abbreviations in pattern 3F or 2F in different SLang. For example, the word euro ('€') isn't inflected in PL, RU, UA, BY, BG, MK and pattern 2F (not 3F) is used in HR and RS.



Pattern	Language
3F	CZ, SK, PL, SL, RU, UA, BY, HR, RS
2F	BG, MK

DigitToWord Transcription Tool for Slavic Languages



DigitToWord Transcription Tool

For Slavic Languages

by SpeechLab Liberec

Input String:
14713324561412

Select Language:

○ Czech ○ Slovak ○ Polish ○ Russian ○ Belarussian ○ Ukrainian
○ Croatian ○ Serbian ○ Slovenian ○ Bulgarian ○ Macedonian

Output String:
четырнадцать триллионов семьсот тринадцать пять миллиардов триста двадцать четыре миллионов пятьсот шестьдесят одна тысяч четырехсот двенадцать

http://kvap.tul.cz/slavic_symbols.php

Conclusion and Future Work

- We have defined several patterns for the translation of any digit string in texts of almost all Slavic languages.

- The digit strings are a cardinal, ordinal, decimal number or date or it is a number in combination with abbreviation.

- The rules are relatively complex but we have focused primarily on the main patterns because we need it for building systems for the automatic transcription of broadcast programs.

- The main application area for these tools is the enhancement of language models or improvement of speech data annotation for training the acoustic model. The tools have been designed and implemented in the same way for all Slavic languages.

- We would like to find the probability of alternative or minor patterns in our audio recordings in the near future. These alternative patterns will be used for random generation of words from symbols in the process of language model re-training

Automatic Symbol Processing for Language Model Building in Slavic Languages

Josef Chaloupka

SpeechLab
Technical University of Liberec
Czech Republic

Thank you for your attention!