The Slovak Categorized News Corpus

Daniel Hládek, Ján Staš, Jozef Juhár

daniel.hladek@tuke.sk, jan.stas@tuke.sk, jozef.juhar@tuke.sk
Department of Electronics and Multimedia Communications
Technical University of Košice
Slovakia

Abstract
The presented corpus aims to be the first attempt to create a representative sample of the contemporary Slovak language from various domains with easy searching and automated processing.

Corpus Gathering
A specialized web agent is used to explore the newspaper web site Content of each web-page is analyzed and saved in a database. Collected items are parsed and given to the appropriate form. The data processing step should preserve as much information as it is possible.

Corpus Contents
All articles are written by a professional reporters and should fulfill a characteristic style of the particular newspaper. The texts should cover the ”correct” Slovak grammar with proper expressions and vocabulary. On the other hand, the newspaper articles in this corpus does not cover the Slovak language in general. Colloquial forms of the language are present only as parts of interviews with people. Improper, vulgar or expressive language is omitted at all. Artistic form of the language is not present.
The size of the corpus has been chosen to be as small as it is possible. The reason is that it should be possible to check the text manually.

Label	# tokens	sentences	documents
Politics	472 305	32 258	1 655
Sport	339 791	24 699	1 042
Culture	13 373	757	22
Economy	376 036	22 893	1 022
Health	231 275	13 067	560
World	149 042	8 388	481
Together	1 581 822	102 062	4 782

Morphological Analysis
The most important part of the annotation process is the morphological annotator Dagger.
This classifier uses second-order hidden Markov model and Viterbi algorithm and can utilize grammatical features for smoothing of the observation and transition matrix for improvement classification accuracy.
The model has been trained on trigram counts from the Slovak National Corpus and uses their tag set containing 3500 distinct tags. The search space of the classifier is restricted by a lexicon that contains a list of possible tags for each known word. Observation probabilities are smoothed using custom algorithm that takes morphological features of words into account. The classifier is 86% correct.

Lemmatization
The classification system is almost the same as in the case of morphological tags: second-order HMM model smoothed using additional information about word suffices.. Each token in the corpus has assigned the most probable lemma. In the case when it was not possible to assign a lemma, a not-available sign is used (%).

Freely available at http://nlp.web.tuke.sk

the Corpus, papers, processing tools, web search, web language processing demo, other resources

Index File with Annotations

corpus/Správy/037261txt Svet varuje Kóreu pred jadrovým testom 04.10.2006 09:08 TASR
corpus/Správy/037271txt Eurosocialisti prijali poslancov Smeru 24.10.2006 15:56 ČTK
corpus/Správy/037274txt Británia zníži počet vojakov v Iraku 22.08.2006 16:16 TASR
corpus/Správy/037301txt Pri Atlantise sa objavil záhadný objekt 19.09.2006 20:01 pravda
corpus/Správy/037302txt Izrael a Libanon sa nedohodli na rezolúcii 09.08.2006 11:46 TASR
corpus/Správy/037325txt Premiér Topolánek vraj chystá rozvod 21.08.2006 12:51 TASR
corpus/Správy/037379txt Ruského expremiéra Gajdara asi otrávili 29.11.2006 14:16 TASR
corpus/Správy/037383txt Pod troskami metra sú ešte stále ľudia 17.10.2006 10:48 ČTK

Keyword-based search

Using any text processing tool, such as text editor, Perl or Python

Annotated Document

Ani|%|W|% množstvo|%|SSns4|% ťažko|%|Dx|% zranených|%|Gtmp2x|% osôb|%|SSfp2|% sa|<STOP>|R|% ešte|<STOP>|T|% nepodarilo|%|VLdscn-|% záchranárom|%|SSms7|% spresniť|%|VId+|% .|<PUN>|Z|bodka
pôvodne|%|Dx|% hovorili|%|VLepcm+|% o|<STOP>|Eu6|% desiatich|%|NUip6|% až|<STOP>|T|% pätnástich|%|NUip6|% prípadoch|%|SSip6|% ,|<PUN>|Z|čiarka teraz|%|Dx|% naopak|%|Dx|% o|<STOP>|Eu6|% piatich|%|NUfp6|% vážnych|%|AAfp6x|% zranených|%|Gtfp6x|% .|<PUN>|Z|bodka
predchádzajúce|%|Gkfp1x|% správy|%|SSfp1|% uvádzali|%|VLepco+|% ,|<PUN>|Z|čiarka že|%|O|% okrem|%|Eu2|% tridsaťročnej|%|AAfs2x|% ženy|%|SSfs2|% prišiel|%|VLdscm+|% o|<STOP>|Eu4|% život|%|SSis4|% vodič|%|SSis4|% jednej|%|NFfs2|% súpravy|%|SSfs2|% ,|<PUN>|Z|čiarka potom|<STOP>|Dx|% však|<STOP>|O|% záchranári|%|SSmp1|% uviedli|%|VLdpcm+|% ,|<PUN>|Z|čiarka že|%|O|% sa|<STOP>|R|% ho|<STOP>|PFms4|% podarilo|%|VLdscn+|% vytiahnuť|%|VId+|% z|<STOP>|Eu2|% trosiek|%|SSfp2|% vlaku|%|SSis2|% živého|%|AAns2x|% .|<PUN>|Z|bodka
rozdielne|%|AAip1x|% informácie|%|SSfp1|% pripisujú|%|VKepc+|% agentúry|%|SSfp1|% zjavnému|%|AAis3x|% zmätku|%|SSis3|% ,|<PUN>|Z|čiarka ktorý|%|PAis1|% na|%|Eu6|% mieste|%|SSns6|% nehody|%|SSfs2|% na|%|Eu6|% začiatku|%|SSis6|% vládol|%|VLesci+|% .|<PUN>|Z|bodka

Available Annotations

Token boundary identification
Sentence boundary identification
Stop-Words annotation
Morphological Analysis
Named Entity Recognition
Named Entity Transcription
Lemma

Sample Document

Ani množstvo ťažko zranených osôb sa ešte nepodarilo záchranárom spresniť .
pôvodne hovorili o desiatich až pätnástich prípadoch , teraz naopak o piatich vážnych zranených .
predchádzajúce správy uvádzali , že okrem tridsaťročnej ženy prišiel o život vodič jednej súpravy , potom však záchranári uviedli , že sa ho podarilo vytiahnuť z trosiek vlaku živého .
rozdielne informácie pripisujú agentúry zjavnému zmätku , ktorý na mieste nehody na začiatku vládol .

Úrady už ale vyvrátili špekulácie , ktoré naznačovali , že za nehodou by mohol byť teroristický útok .
" nič tomu nenasvedčuje , " citovala agentúra Reuters nemenovaný úradný zdroj .
obavy z teroristického útoku mal aj rímsky starosta Walter Veltroni , ktoré svoje obavy vyjadril pred novinármi na mieste nehody .

Prevádzka na linke metra A , ktorá tvorí základnú spojnicu rímskej hromadnej dopravy , bola na nespresnenú dobu prerušená .
" snažíme sa vyslobodiť ďalších cestujúcich , ktorí sú ešte uväznení v súprave .

Corpus Utilization

Automatic excerpt extraction,
language model evaluation,
document categorization,
language model adaptation,
linguistic research.

Word and Sentence Boundary Detection
The main goal is to distinguish between types of tokens that are interesting for further processing by adding and removing spaces and unnecessary characters as it is required.

The following types of tokens are recognized:
• words and acronyms,
• abbreviations,
• various number representations,
• URLs and e-mails,
• punctuation.

1. List of recognized tokens is searched. The longest matching token is selected.
2. If recognized token is a dot, colon, empty line, exclamation mark or question mark, the end of sentence is found.
3. If no token is found, the first character is discarded and the search process continues.
4. If some other token is found, it is added to the sentence, characters are discarded from the input and the search process continues. If the token is the first in the sentence and it is not in the list of exceptions then it is lowercased.
5. If there are no more characters in the input string, the search process finishes.

Named Entity Recognition
The named entity recognition uses a very simple rule-based approach. Each recognized named entity is recognized using a certain rule that can be a regular expression or a dictionary item.
The Slovak language has very similar features to the Czech and the future research in the named entity recognition will be focused on a statistical approach.

entity	tag
Integer numbers	<INT>
Floating point numbers	<FLOAT>
Names of months	<MONTH>
Male Names	<MM>
Male Surnames	<MP>
Female Names	<ZM>
Female Surnames	<ZP>
Slovak cities and villages	<OBEC>
Slovak street names	<ULICA>
Organization names	<ORG>
Names of countries	<COUNTRY>
Other geographical locations	<LOCATION>
Stop-words	<STOP>

Named Entity Transcription
If it is applicable, a transcription of a named entity to the spoken form is provided. Again, it is performed using a rule-based system that is capable of utilizing morphological information to transcribe token with a correct grammatical form. Transcription to a verbal form is helpful for normalization of the token meaning and for training a language model.

Európska únia
Európsky fond regionálneho rozvoja
Operačný program VÝSKUM a VÝVOJ
Agentúra Ministerstva školstva, vedy, výskumu a športu SR pre štrukturálne fondy EÚ

We support research activities in Slovakia / This project is being co-financed by the European Union

Acknowledgements
The research presented in this paper was supported by Research and Development Operational Program funded by the ERDF under the project numbers ITMS-26220220182 (50%) and ITMS-26220220141 (50%).

LREC 2014: The 9th edition of the Language Resources and Evaluation Conference, 26-31 May, Reykjavik, Iceland