My Recent Work In The NLP Lab: Genre Classification, Discerning Similar Languages, Czech Web Corpus czTenTen

Vít Suchomel

Natural Language Processing Centre,
Faculty of Informatics, Masaryk University

2018-03-22

Genre Classification

- Problem: Classifier applied to enTenTen15 helps but still does perform well enough
- Are the categories well defined? $\Rightarrow$ Experiment using annotations made with the former annotation scheme
  - annotations in 13 classes $\Rightarrow$ low IAA
  - annotations merged into 7 classes $\Rightarrow$ better IAA?

New genre categories

1. Information
2. News
3. Discussion
4. Instructions
5. Fiction
6. Legal
7. Non-text

Old to new class mapping

1. Argumentative → Information + News + Discussion
2. Fictive → Fiction
3. Instructive → Instructions
4. Hardnews → News
5. Legal → Legal
6. Personal → Discussion
7. Commercial → Information
8. Ideology → Information
9. Science → Information
10. Informative → Information
11. Evaluative → Information
12. Apellative → Information
13. Nontext → Non-text

Cohen's kappa pairwise, 13 classes

Class	both YES	both NO	disagreement	kappa
All labels	67	1462	122	0.48
Apellative	1	119	7	0.20
Argumentative	4	114	9	0.44
Commercial	4	113	10	0.40
Evaluative	0	123	4	-0.01
Fictive	5	119	3	0.76
Hardnews	15	96	16	0.58
Ideology	2	108	17	0.15
Informative	9	91	27	0.27
Instructive	9	108	10	0.60
Legal	0	125	2	0.00
Nontext	8	112	7	0.67
Personal	5	114	8	0.53
Science	5	120	2	0.83

Cohen's kappa pairwise, 7 classes

Class	both YES	both NO	disagreement	kappa
All labels	95	699	95	0.60
Discussion	10	102	15	0.50
Fiction	5	119	3	0.76
Information	43	49	35	0.45
Instructions	9	108	10	0.60
Legal	0	125	2	0.00
News	20	84	23	0.52
Non-text	8	112	7	0.67

Jaccard's similarity of positive labels pairwise, 13 classes

Class	both YES	disagreement	Jaccard's similarity	
All labels	67	122	0.35	
Apellative	1	7	0.12	
Argumentative	4	9	0.31	
Commercial	4	10	0.29	
Evaluative	0	4	0.00	
Fictive	5	3	0.62	
Hardnews	15	16	0.48	
Ideology	2	17	0.11	
Informative	9	27	0.25	
Instructive	9	10	0.47	
Legal	0	2	0.00	
Nontext	8	7	0.53	
Personal	5	8	0.38	
Science	5	2	0.71	

Jaccard's similarity of positive labels pairwise, 7 classes

Class	both YES	disagreement	Jaccard's similarity	
All labels	95	95	0.50	
Discussion	10	15	0.40	
Fiction	5	3	0.62	
Information	43	35	0.55	
Instructions	9	10	0.47	
Legal	0	2	0.00	
News	20	23	0.47	
Non-text	8	7	0.53	

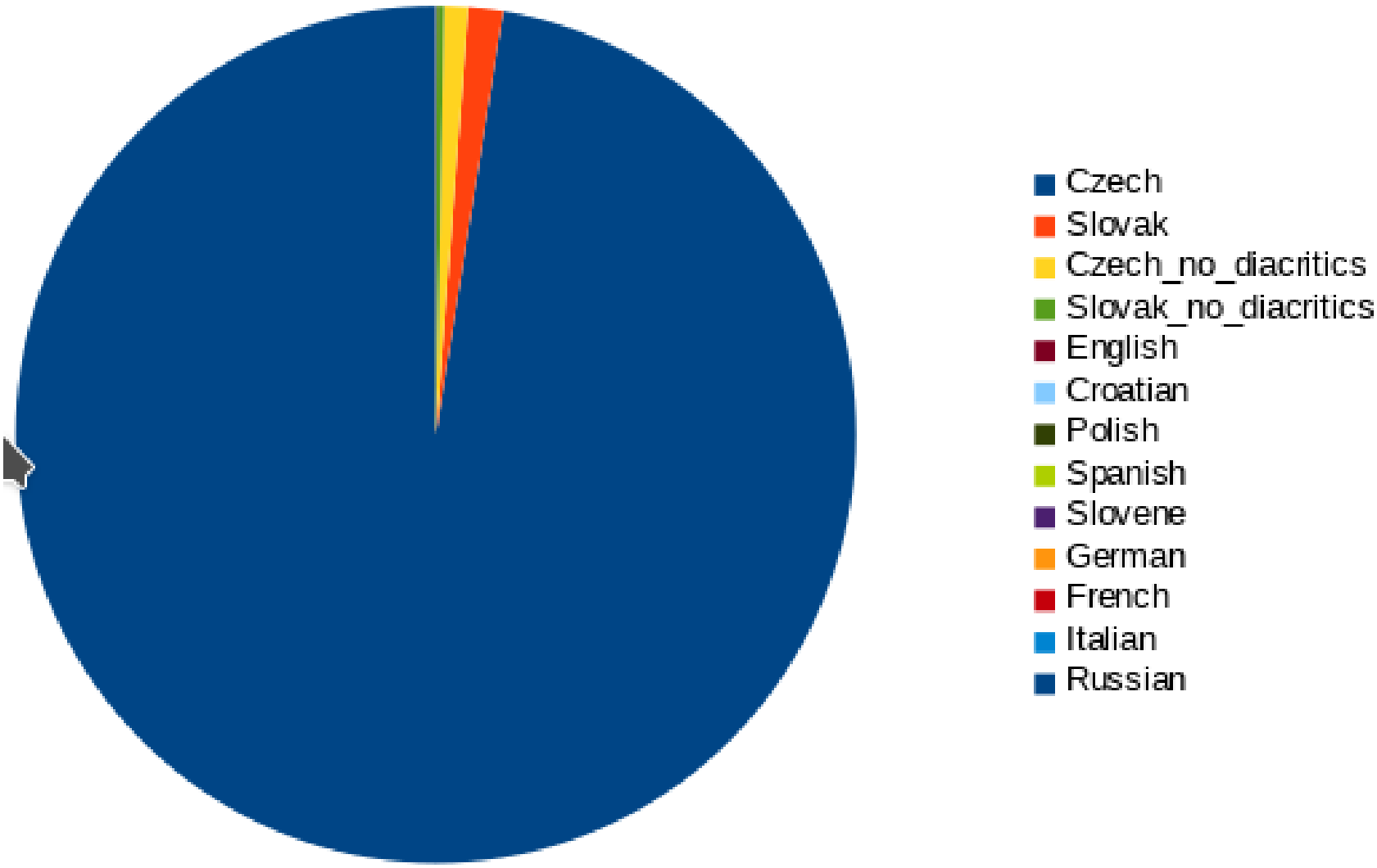
Discerning Similar Languages

- Based on relative frequency of words in a general web corpus
- A simple version of the algorithm submitted to a competition (with Pary, Ondra H., Vítek B.) implemeted as a script easy to use
- Applied to discern 26 languages in web corpora by now
- Will be incorporated in the crawler

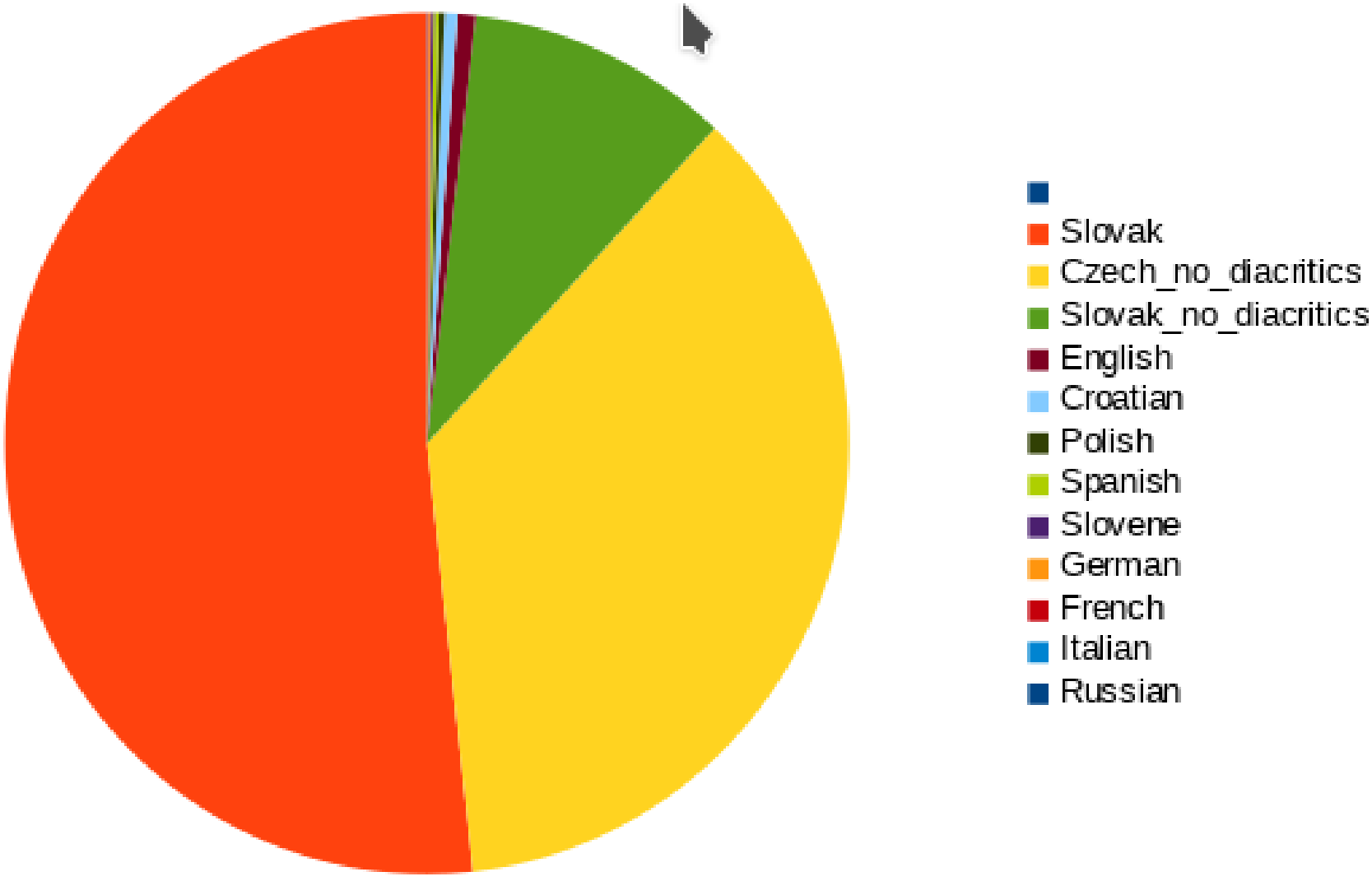
DSL – The case of Czech

- Sub-languages to remove text without diacritics: Czech, Czech without diacritics, Slovak, Slovak without diacritics
- And other languages not caught by a permissively configured trigram model inbuilt in the crawler

DSL – The case of Czech



DSL – The case of Czech



DSL script output sample

```
<p lang="English" lang_scores="Czech: 35.19, Czech_no_diac
   ritics: 31.14, Slovak: 33.90, Slovak_no_diacritics: 31.87,
   English: 36.50, German: 27.55, Polish: 29.62, Slovene: 29.31,
   Croatian: 31.02, Russian: 0.00, French: 29.60, Spanish: 25.20
   Italian: 25.55">
#token    cs    cs-nd     sk sk-nd     en     de     pl     sl     hr
Same      3.06  3.34   4.01  4.23   5.71   3.24   5.14   4.91   4.90   0
day       3.89  4.18   3.95  4.17   5.89   4.22   3.93   3.88   4.15   0
service   3.90  4.19   3.86  4.08   5.74   4.97   3.98   3.65   3.84   0
,         0.00  0.00   0.00  0.00   0.00   0.00   0.00   0.00   0.00   0
servis    4.58  4.86   4.53  4.75   1.69   1.55   3.35   4.51   4.58   0
24/7/365  0.00  0.00   0.00  0.00   0.00   0.00   0.00   0.00   0.00   0
,         0.00  0.00   0.00  0.00   0.00   0.00   0.00   0.00   0.00   0
Hand      3.36  3.65   3.31  3.53   5.30   5.32   3.30   3.43   3.57   0
carry     2.82  3.11   3.14  3.36   4.84   2.93   3.05   2.76   3.07   0
a         7.52  7.81   7.53  7.75   7.32   5.32   6.88   6.17   6.90   0
další     6.06  0.00   3.57  0.00   0.00   0.00   0.00   0.00   0.00   0
</p>
```

csTenTen17

https://ske.fi.muni.cz/bonito/run.cgi/corp_info?corpname=preloaded/cstenten17_0&struct_attr_stats=1&subcorpora=1

Photo: Kathleen & Ryan Rush