Evaluating Russian Adj-Noun Word Sketches against Dictionaries: a Case Study

Maria Khokhlova

St Petersburg State University

RASLAN 2020

Outline

- Introduction
- Methods
- Russian Word Sketch Grammar
- Results
- Conclusion
- Acknowledgement

Introduction

- The purpose of our research is to compare word sketches with verified lexicographic data, i.e. trace the intersection between data collected by experts and automatically extracted from an up-to-date corpus.
- A profound evaluation of word sketches was presented by Adam Kilgarriff et al. [2010] for four languages (Dutch, English, Japanese and Slovene).
- 'user evaluation' perspective

Methods

- Gold Standard of Russian collocations;
- 6 different Russian dictionaries:
  - two explanatory dictionaries, i.e. the Dictionary of the Russian Language; the Large Explanatory Dictionary of the Russian Language;
  - three collocation dictionaries [Borisova 1995; Oubine 1987; Reginina, Tjurina, Shirokova 1980];
  - an online dictionary based on the Russian National corpus [Kustova 2008].
- adjective / participle + noun, e.g., *prakticheskoye znacheniye* 'practical meaning', *zhiznennyy uspekh* 'life success', *oslepitel'naya krasota* 'dazzling beauty', etc;
- Russian Word sketch grammar described in [Khokhlova 2009].
- ruTenTen corpus.

Methods

20 headwords with the largest number of collocates (given in parentheses):

sila 'force' (97), *uspekh* 'success' (59), *bor'ba* 'fight' (55), *toska* 'boredom' (54), *lyubov'* 'love' (49), *interes* 'interest' (46), *delo* 'case' (43), *bolezn'* 'illness' (42), *radost'* 'joy' (40), *pamyat'* 'memory' (40), *krasota* 'beauty' (38), *znacheniye* 'meaning' (37), *chuvstvo* 'sense' (36), *sistema* 'system' (36), *nenavist'* 'hate' (36), *um* 'intellect' (35), *strast'* 'passion' (34), *rol'* 'role' (34), *kholod* 'cold' (33), *usiliye* 'effort' (32).

Methods

- top-50 examples produced by word sketches;
- logDice and joint frequency;
- precision: proportion of collocations (identified either in the gold standard or by expert evaluation) in the output;
- recall: proportion of collocations from the gold standard that were correctly extracted from the corpus.

Russian Word Sketch Grammar

1. adj-noun (e.g. *temperaturnyy rezhim* 'temperature regime');
2. adj-adj-noun (e.g. *vysokochastotnyy elektricheskiy tok* 'high-frequency electrical current');
3. adj-adj-adj-noun (e.g. *global'naya sputnikovaya navigatsionnaya sistema* 'global navigation satellite system');
4. adj-adj-adj-adj-noun (e.g. *kitayskiy zelenyy baykhovyy krupnolistovoy chay* 'Chinese green loose large leaf tea');
5. adj-conj-adj-noun (e.g. *mobil'nyy ili domashniy telefon* 'mobile or home phone number');
6. adj-,-adj-conj-adj-noun (e.g. *tekhnicheskaya, informatsionnaya i reklamnaya podderzhka* 'technical, information and advertising support');

Russian Word Sketch Grammar

7. adj-,-adj-,-adj-conj-adj-noun (e.g. *administrativnoye, pensionnoye, sotsial'noye i trudovoye zakonodatel'stvo* 'administrative, pension, social and labour law');
8. adj-,-adj-noun (e.g. *federal'nyy, regional'nyy uroven'* 'federal, regional level');
9. adj-,-adj-,-adj-noun (e.g. *neftyanaya, khimicheskaya, pischevaya promyshlennost'* 'oil, chemical, food industry');
10. adj-,-adj-,-adj-,-adj-noun (e.g. *doshkol'noye, obscheye, dopolnitel'noye, vyssheye obrazovaniye* 'preschool, general, supplementary, higher education').

Results

- morphological errors;
- representation of participles as verb forms (infinitives);
- number of such cases: 7.4% for logDice and 4% for joint frequency respectively;
- logDice tends to extract more peculiar collocations; *tsepkaya pamyat'* 'tenacious memory' and *fotograficheskaya pamyat'* 'photographic memory', i.e. these collocations can be listed in entries of dictionaries for Russian language learners.
- precision against gold standard is 0.33 and 0.34 for logDice and joint frequency respectively;
- precision against expert evaluation is 0.51 and 0.44 respectively.

Results: Precision and Recall

evaluation: d. - dictionary, ex. - expert; ID - logDice; f - frequency.

Headword	Prec. (d., ID)	Prec. (ex., ID)	Prec. (d., f)	Prec. (ex., f)	R. (d., ID)	R. (d., f)
bolezn'	0.48	0.88	0.52	0.84	0.57	0.62
bor'ba	0.38	0.54	0.42	0.54	0.35	0.38
chuvstvo	0.16	0.24	0.28	0.30	0.22	0.39
delo	0.18	0.36	0.22	0.36	0.21	0.26
interes	0.28	0.46	0.24	0.36	0.30	0.26
kholod	0.42	0.50	0.36	0.42	0.64	0.55
krasota	0.36	0.56	0.36	0.44	0.47	0.47
lyubov'	0.32	0.66	0.30	0.56	0.33	0.30
nenavist'	0.32	0.38	0.38	0.44	0.44	0.53
pamyat'	0.20	0.58	0.18	0.42	0.25	0.23

Results: Precision and Recall

evaluation: d. - dictionary, ex. - expert; ID - logDice; f - frequency.

Headword	Prec. (d., ID)	Prec. (ex., ID)	Prec. (d., f)	Prec. (ex., f)	R. (d., ID)	R. (d., f)
radost'	0.38	0.46	0.36	0.38	0.48	0.45
rol'	0.48	0.56	0.44	0.50	0.71	0.65
sila	0.46	0.84	0.44	0.70	0.24	0.23
sistema	0.08	0.40	0.08	0.36	0.11	0.11
strast'	0.32	0.42	0.36	0.42	0.47	0.53
toska	0.40	0.44	0.50	0.54	0.37	0.46
um	0.38	0.52	0.34	0.38	0.51	0.49
usiliye	0.18	0.34	0.24	0.28	0.28	0.38
uspekh	0.56	0.64	0.40	0.52	0.47	0.34
znacheniye	0.28	0.40	0.36	0.02	0.38	0.49

Conclusion

- examined word sketches for 20 nouns with the largest number of collocates according to six Russian dictionaries;
- precision of the word sketches output is a bit low with regard to the data from the gold standard and more promising assessed by expert evaluation;
- at least half of the produced word sketches can be called "true collocations" and can be included into dictionaries (that do not list them yet);
- logDice measure turns out to be much more successful for extracting and ranking word sketches according to the expert assessment;
- plan to evaluate other models described in the word sketch grammar and analyse more headwords.

Acknowledgement

This work was supported by the grant of the Russian Science Foundation (Project No. 19-78-00091).