Statistical modelling of doublets in denominal adjective formation in Russian

Natalia Bobkova

University of Toulouse Jean Jaurès CLLE

natalia.bobkova@univ-tlse2.fr

09/09/2021

Russian denominal adjectives

- Typical case of affixal rivalry [Lindsay and Aronoff, 2013, Aronoff, 2016, Bonami and Thuilier, 2018].
- Russian denominal (relational) adjectives exhibit a wide range of variation in the strategies (affixes) employed;
  - e.g.Townsend [1975] and Švedova [1980] both enumerate at least 25 suffixes / variants.
- The affixes involved have various degrees of productivity;
- They appear with several variants ('extensions') displaying complex interactions with base lexemes;

Adjectival suffixes

- Three main suffixes + their extended variants
  - -N-: *-n-*, *-Ovn-*, *-ičn-*, *-ivn-*, *-on(n)-*, *-en(n)-*, *-(e)stven(n)-*, *-ozn-*, *-al'n-*, *-onal'n-*, *-arn-*, *-in-*;
  - -SK-: *-sk-*, *-esk-*, *-česk-*, *-ičesk-*, *-ističesk-*, *-ijsk-*, *-ansk-*, *-ensk-*, *-insk-*, *-istsk-*, *-Ovsk-*;
  - -OV-: *-Ov-*.

1. -OV-

begemot	→ begemot**ov**(yj)	'hippopotamus'

2. -SK-

universitet	→ universitet**sk**(ij)	'university'
korol'	→ korol**evsk**(ij)	'king'
jazykoved	→ jazykoved**česk**(ij)	'linguist'

3. -N-

kanal	→ kanal**n**(yj)	'canal'
vin(a)	→ vin**ovn**(yj)	'guilt'
bukv(a)	→ bukv**al'n**(yj)	'letter'

Problem statement

Existence of doublets

Base noun	**Adj_1**	**Adj_2**	**Adj_3**	**Suffixes**
ZIMA 'winter'	ZIMOVOJ	ZIMNIJ		*-Ov-/-n-*
MUZEJ 'museum'	MUZEJNYJ	MUZEJSKIJ		*-n-/-sk-*
LONDON 'London'	LONDONOVSKIJ	LONDONSKIJ		*-Ovsk-/sk-*
ANEMIJA 'anemia'	ANEMIČESKIJ	ANEMIČNYJ		*-ičesk-/-ičn-*
DRUID 'druid'	DRUIDIČESKIJ	DRUIDSKIJ		*-ičesk-/-sk-*
LOGIKA 'logic'	LOGIČESKIJ	LOGIČNYJ		*-esk-/-n-*
BOEC 'fighter'	BOJCOVYJ	BOJCOVSKIJ		*-Ov-/-Ovsk-*
KON' 'horse'	KONEVOJ	KONNYJ	KONSKIJ	*-n-/-sk-/-Ov-*

Table: Doublets and triplets in Russian adjectival formation

Goal: use statistical approaches to reveal the main properties of base nouns which may allow the choice of more than one affix

Relevant features

Which factors are the best predictors for the choice of the suffix (variant)?

- Semantics (animacy of the base N);
- Tendency to be more or less relational (vs. qualitative); Švedova [1980]
- Morphology (frequency of combination with nominal affixes);
- Origin of the base noun

No clear specifications about formal properties of the base nouns and about emergence of doublets.

Overview

Russian National Corpus

- Main subcorpus: texts representing standard Russian;
- Media subcorpus: articles from mass media (1990-2000s);
- Multimedia subcorpus: Russian movies (1930-2000);
- Corpus of Spoken Russian: recordings of public and spontaneous spoken Russian;
- Poetry subcorpus: covers the time frame between 1750 and the 1890s, but also includes some poets of the 20th century;
- Dialectal subcorpus: recordings of dialectal speech (presented in loosely standardized orthography) from different regions of Russia;
- Educational subcorpus: small disambiguated corpus adapted for the Russian educational program;
- Parallel text subcorpus: texts in Russian are complemented by their translations into different languages, and vice versa.

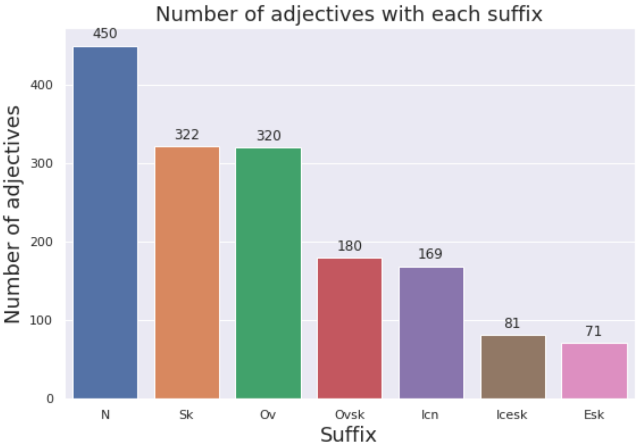
Data set constitution

- Source : http://www.ruscorpora.ru/
  $\sim$ 360M words
- Automatic extraction of denominal adjectives (**75k** adjectives)
- Automatic grouping of adjectives derived from the same noun base
- Automatic reconstitution of nominal bases
- **1968** raw base nouns with at least two adjectives
- Manual cleaning:
  - correct assignment of adjectives
  - exact shape of base noun

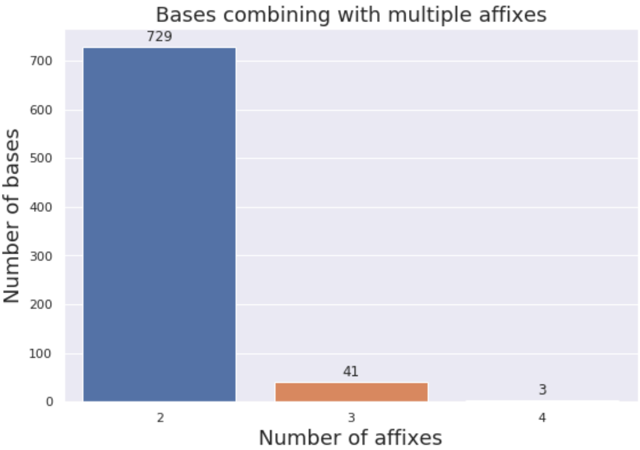
Data set constitution

- Manual cleaning: false positives
    - proper nouns (-*Ovsk*-, -*Ov* and -*sk*- suffixes): STANISLAVSKIJ, MENDELEEV, AJVAZOVSKIJ;
    - forms of genitive plural (-*ov* inflection): dvor$_{\text{NOM}}$ 'yard' - dvorov$_{\text{GEN}}$;
    - possessive adjectives (-*ov*- suffix): DED 'grandpa' - DEDOV.
- **773** base nouns **1593** derived adjectives

Data set exploration



Data set exploration



Data set exploration

Suffixes	N of bases
-Ov-/-n-	226
-n-/-sk-	100
-Ovsk-/sk-	75
-ičesk-/-ičn-	71
-ičesk-/-sk-	63
-esk-/-n-	51
-Ov-/-Ovsk-	41

Table: Distribution of doublets

Semantic properties

Property	Type of values	Values
±prop	binary	0,1
±conc	binary	0,1
±count	binary	0,1
±anim	binary	0,1
±human	binary	0,1
animacy	non binary	A,B,C,D,E

- AnimA: proper human (PIFAGOR'Pithagoras');
- AnimB: common human/animate (SOBAKA 'dog');
- AnimC: common concrete (DOM 'house');
- AnimD: proper non-human (AL'PY 'Alps');
- AnimE: common abstract (SOJUZ 'alliance').

Morphological properties

Property	Type of values	Values
`InflCl`	non binary	1, 2, 3
`VowAlt`	binary	0, 1
`ConsM`	binary	0, 1

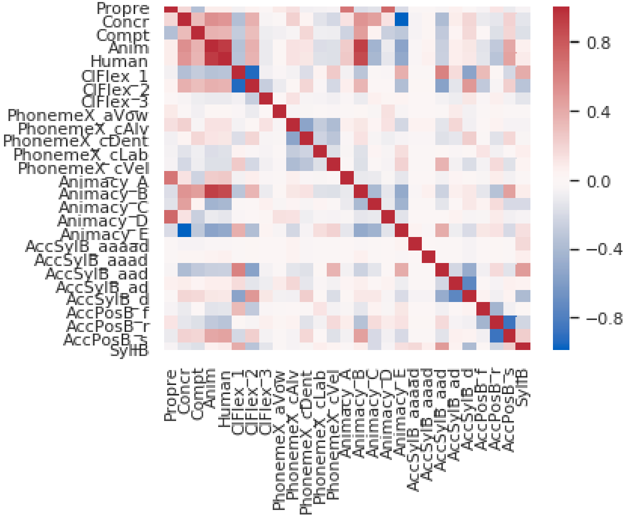
- `InflCl`: canonical distinction between inflectional classes
  PAPA$_{\mathrm{I.M}}$ 'dad', PESNJA$_{\mathrm{I.F}}$ 'song';
  STOL$_{\mathrm{II.M}}$ 'table', DELO$_{\mathrm{II.N}}$ 'business';
  TEN'$_{\mathrm{III.F}}$ 'shadow'
- `VowAlt`: vowel / Ø alternation
  DVOREC 'palace' - DVORCOVYJ
- `ConsM`: consonant mutation
  (TVOROG 'cottage cheese' - TVOROŽNYJ)

Phonological properties

Property	Type of values	Values
`LastP`	non binary	`Lab`, `Den`, `Alv`, `Vel`, `Vow`
`SyllB`	continuous	`1-7`
`AccSyl`	non binary	`d`, `ad`, `aad`
`AccPos`	non binary	`r`, `s`, `f`

- `AccSyl`: which syllable is stressed
  \zim'a \'winter', \'viʃnʲa \'cherry', \'raduga \'rainbow');
- `AccPos`: where the stressed is positioned
  \'son \'dream', \mark'sizm \'marxism', \galav'a \'head').

Correlation of features



Correlation of features

	Propre	Concr	Compt	Anim	Human
AnimA	**0.66**	0.13	-0.24	0.24	0.25
AnimB	-0.17	**0.51**	**0.46**	**0.92**	**0.87**
AnimC	-0.16	**0.46**	0.05	**-0.44**	**-0.41**
AnimD	**0.73**	0.13	-0.26	-0.11	-0.13
AnimE	-0.19	**-0.99**	**-0.31**	**-0.54**	**-0.51**
AccSylAad	-0.05	**-0.38**	**-0.30**	-0.26	-0.25
AccPosR	0.12	-0.13	-0.17	**-0.30**	**-0.33**
AccPosS	-0.10	0.21	0.19	**0.41**	**0.44**
InflCl1	-0.07	**-0.36**	-0.28	**-0.35**	**-0.33**
InflCl2	0.05	**0.37**	**0.31**	**0.37**	**0.35**

Table: Correlation coefficient for predictors, where $|\rho| \geq 0.3$

Correlation to suffixes (1)

-Ov-/-n-	-n-/-sk-	-Ovsk-/-sk-	-ičesk-/-ičn-
Anim (0.59)	**Anim (0.24)**	**Anim (0.51)**	SyllB (0.36)
SyllB (0.35)	ConsM (0.24)	**LastP (0.16)**	**LastP (0.28)**
AccPos (0.26)	AccSyl (0.19)	InflCl (0.15)	**Anim (0.27)**
VowAlt (0.23)	LastP (0.16)	AccPos (0.14)	AccSyl (0.25)
InflCl (0.17)	AccPos (0.15)	SyllB (0.13)	InflCl (0.23)
AccSyl (0.16)	SyllB (0.14)	ConsM (0.07)	ConsM (0.21)
LastP (0.006)	InflCl (0.12)	AccSyl (0.05)	AccPos (0.20)
ConsM (0.04)	VowAlt (0.08)	VowAlt (0.04)	VowAlt (0.06)

Table: Cramer's V for each suffix, from ρ max to ρ min

Correlation to suffixes (2)

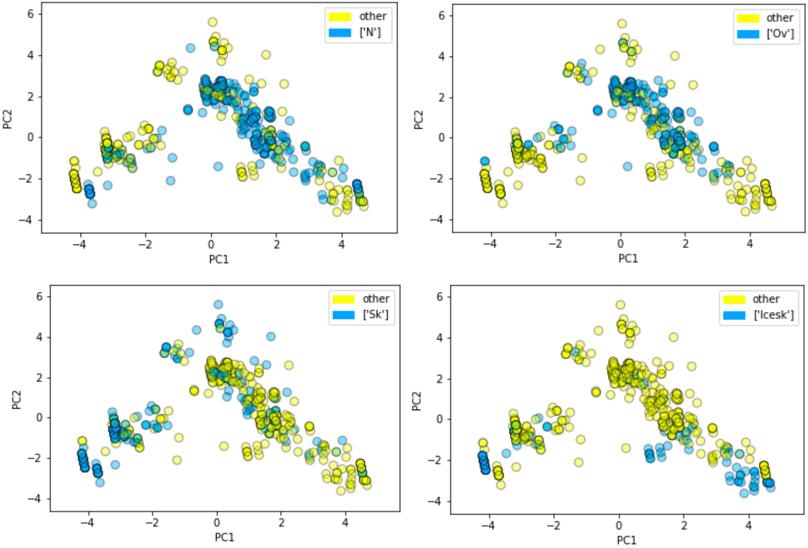
-ičesk-/-sk-	*-esk-/-n-*	*-Ov-/-Ovsk-*
AccPos (0.41)	AccSyl (0.49)	SyllB (0.19)
Anim (0.38)	**LastP (0.49)**	InflCl (0.13)
LastP (0.25)	InflCl (0.35)	LastP (0.10)
AccSyl (0.14)	SyllB (0.24)	Anim (0.09)
InflCl (0.13)	Anim (0.22)	ConsM (0.08)
SyllB (0.12)	AccPos (0.11)	AccSyl (0.08)
ConsM (0.10)	ConsM (0.11)	AccPos (0.06)
VowAlt (0.05)	VowAlt (0.00)	VowAlt (0.00)

Table: Cramer's V for each suffix, from ρ max to ρ min

Logistic regression

- Interpretability
- Visualization
- Weights
- Two types of classification
  - One-vs-Rest
  - Two-vs-Rest

Multilabel classification: One-vs-Rest approach



Results

	Accuracy
-esk-	90
-ičn-	86
-Ovsk-	82
-Ov-	81
-sk-	75
-ičn	70
-n-	53

Table: Results of logistic regression model (Multilabel approach)

Binary classification: Two-vs-Rest approach

Input	*OvN*	*NSk*	*OvskSk*	*IceskIcn*	*IceskSk*	*EskN*	*OvOvsk*	*IceskN*
...	1	0	0	0	0	0	0	0
...	1	0	0	0	0	0	0	0
...	0	1	0	0	0	0	0	0
...	0	0	0	0	1	0	0	0
...	0	0	0	1	0	0	0	0
...	0	0	0	1	0	0	0	0
...	0	0	0	0	0	0	1	0

Table: Two-vs-Rest approach

Results

	Accuracy
-esk-/-n-	95.79
-ičesk-/-sk-	94.52
-Ovsk-/-sk-	93.84
-Ov-/-Ovsk-	92.47
-ičesk-/-ičn-	91.78
-n-/-sk-	87.67
-Ov-/-n-	81.50

Table: Results of logistic regression model (Binary classification)

Relevant features

-*Ov*- and -*n*-:

- nouns designating common concrete entities, $p<0.001$
  OGUREC 'cucumber' - OGURCOVYJ / OGUREČNYJ
- nouns with stressed radical, $p<0.027$,
  /smet'ana/ 'cream' - SMETANNYJ / SMETANOVYJ
- nouns affected by vowel alternation, $p<0.049$
  PEPEL 'ashes' - PEPLOVYJ / PEPEL'NYJ.

-*ičesk*- and -*ičn*-:

- inflectional class II, $p<0.034$
  SINONIM 'synonym' - SINONIMIČNYJ / SINONIMIČESKIJ
- stress on derivational affix of the base noun, $p<0.050$
  -*izm* or -*ist*, as in CINIZM 'cynicism' - CINIČESKIJ / CINIČNYJ.

Relevant features

-Ovsk- and *-sk-*:

- inflectional class II, $p<0.018$
  BANKIR 'banker' - BANKIROVSKIJ / BANKIRSKIJ
- common human or common animate, $p<0.001$
  SULTAN 'sultan' - SULTANOVSKIJ / SULTANSKIJ)
- common abstract, $p<0.000$
  INTERNET 'internet' - INTERNETOVSKIJ / INTERNETSKIJ)

-Ov- and *-Ovsk-* :

- the length of the stem in syllables, $p<0.000$
- common concrete nouns, $p<0.019$
  BOJEC 'fighter' - BOJCOVYJ / BOJCOVSKIJ
- common abstract nouns, $p<0.001$
  ONLAJN 'online' - ONLAJNOVYJ / ONLAJNOVSKIJ.

Relevant features

-ičesk-/-sk-:

- inflectional class II, p<0.050
  VAMPIR 'vampire' - VAMPIRIČESKIJ / VAMPIRSKIJ;
- the length of the stem in syllables, p<0.043.

-esk- and *-n-*:

- common abstract, p<0.003
  MISTIKA 'mysticism' - MISTIČESKIJ / MISTIČNYJ;
- stems end with velars, p<0.010
  ISTERIKA 'hysterics' - ISTERIČESKIJ / ISTERIČNYJ.

-n- and *-sk-*:

- inflectional class II, p<0.028
  INVALID 'disabled person' - INVALIDNYJ / INVALIDSKIJ;
- the length of the stem in syllables, p<0.012
  ZRITEL' 'viewer' - ZRITEL'NYJ / ZRITEL'SKIJ

Conclusion

Most recurrent properties: semantic and morphological factors

- Animacy
  -Ov-/-n-, *-esk-/-n-*, *-Ov-/-Ovsk-* and *-Ovsk-/-sk-* doublets
- Inflectional class
  -ičesk-/-ičn-, *-n-/-sk-*, *-ičesk-/-sk-* and *-Ovsk-/-sk-* suffixes

Less recurrent properties: phonological factor

- the length of stem in syllables
  -n-/-sk-, *-ičesk-/-sk-* and *-Ov-/-Ovsk-* suffixes

Perspectives

- Frequency of doublets
  - Formally/semantically transparent/opaque
    TRUD 'work' - TRUDNYJ 'difficult' / TRUDOVOJ 'work-related'
  - Common/hapax
    PRIZRAK 'ghost' - PRIZRAČNYJ (2724) / PRIZRAKOVYJ (1)
- Type of subcorpus
  - main: ALMAZNYJ 'dimond', NOVOSTNOJ 'news'
  - poetic and oral: ALMAZOVYJ, NOVOSTEVOJ

Thank you for your attention

References I

Ol'ga Pavlovna Antipina. *Sopostavitel'nyj analiz paronimov russkogo i anglijskogo jazykov*. PhD thesis, Bashkir State University, 2012.

Mark Aronoff. Competition and the lexicon. In Elia, Annibale, Iacobini, Claudio, and Voghera, Miriam, editors, *Livelli di Analisi e fenomeni di interfaccia. Atti del XLVII congresso internazionale della Società Linguistica Italiana*, pages 39–52. Bulzoni, 2016.

Natalia Bobkova and Fabio Montermini. Suffix rivalry in russian: what low frequency words tell us. In *Mediterranean Morphology Meetings*, volume 12, pages 1–17, 2019.

Olivier Bonami and Juliette Thuilier. A statistical approach to rivalry in lexeme formation: French *-iser* and *-ifier*. *Word Structure*, 11(2), 2018.

Georgette Dal and Fiammetta Namer. Faut-il brûler les dictionnaires? ou comment les ressources numériques ont révolutionné les recherches en morphologie. In *SHS Web of Conferences*, volume 1, pages 1261–1276. EDP Sciences, 2012.

Nabil Hathout. Une approche topologique de la construction des mots: propositions théoriques et application à la préfixation en anti. *Des unités morphologiques au lexique*, pages 251–318, 2011.

Christine Hénault and Sergueï Sakhno. Çem supermarket-n-yj luçșe supermarket-sk-ogo? slovoobrazovatel'naja sinonimija v russkix ad"ektivnyj neologizmax po dannym interneta. *B. Tošovic, A. Wonisch. Wortbildung und Internet*, 2015.

Vsevolod Kapatsinski. Velar palatalization in russian and artificial grammar: Constraints on models of morphophonology. *Laboratory phonology*, 1(2):361–393, 2010.

Galina Ivanovna Kustova. Prilagatel'nye. *Materialy k korpusnoj grammatike russkogo jazyka. Vyp.3. Časti reči i leksiko-grammatičeskie klassy*, pages 40–107, 2018.

Stéphanie Lignon. –iser and –ifier suffixations in French: Verify data to verize hypotheses? In *Décembrettes 7*, 2010.

References II

Mark Lindsay and Mark Aronoff. Natural selection in self-organizing morphological systems. In *Morphology in Toulouse. Selected Proceedings of Décembrettes 7 (Toulouse 2-3 December 2010)*, pages 133–153. Lincom Europa, 2013.

Marc Plénat. Enquête sur divers effets des contraintes dissimilatives en français. *M. Roché, G. Boyé, N. Hathout, S. Lignon & M. Plénat, Des unités morphologiques au lexique. Paris: Hermès-Lavoisier*, pages 145–190, 2011.

Michel Roché. Quel traitement unifié pour les dérivations en-isme et en-iste?, 2011.

Andrea D Sims. Slavic morphology: Recent approaches to classic problems, illustrated with Russian. *Journal of Slavic Linguistics*, 25(2):489–524, 2017.

Juliette Thuilier. *Contraintes préférentielles et ordre des mots en français*. PhD thesis, Université Paris-Diderot-Paris VII, 2012.

Alan Timberlake. *A reference grammar of Russian*. Cambridge University Press, 2004.

Charles Edward Townsend. *Russian word-formation*. Slavica Publishers, 1975.

Natal'ja Švedova. *Russkaja grammatika*, volume 1. Moskva: Nauka, 1980.

Elena Andreevna Zemskaya. *Jazyk kak dejatel'nost'. Morfema, slovo, reč*. Moskva: Flinta, 2015.