DeriMo 2023

FOURTH WORKSHOP ON RESOURCES AND TOOLS FOR DERIVATIONAL MORPHOLOGY

DUBROVNIK, 05-06 OCTOBER 2023

Automatic detection of grammatical aspect of Russian verbs based on their morphological properties

Uliana Petrunina and Hana Filip

Heinrich Heine University Düsseldorf, Institute for Linguistics

October 5, 2023

Heinrich Heine Universität Düsseldorf

Detection of grammatical aspect in Russian

- Differentiation between perfective (PFV) and imperfective (IMPF) verbs:
  - derivational pairs
  - derivational affixes
- Distribution of derivational pairs in a vector-space model

Main Questions

Heinrich Heine
Universität Düsseldorf

Q1

- Does morphological form contribute to distinguishing grammatical aspect in Russian?
  - derivational affixes
    - typically modify the lexical meaning of the verb stem they are added to
    - often modify argument structure of the verb stem

Q2

- How well is grammatical aspect detected by the distributional semantics model?
  - subword (morphology-based) information
  - distributional vector space
  - visualization of clustering

Outline

I. Theoretical background

- Grammatical Aspect
- Derivational pairs
- Prefixes and Suffix -Nu-
- Distributional Semantics: Methods

II. Experiment

- FastText
- Tools
- Derivational data
- Processing
- Vector space of Russian aspect
- Error analysis

III. Conclusions and future work

I. Theoretical background

Grammatical Aspect

Perfective versus Imperfective Distinction

- IMPERFECTIVE: verb *čitat'* 'read'

Ja	*čital*	*knigu*	dva	dnja.
I	read.IMPF.PST	book	two	day

'I have been reading the book for two days.'

- PERFECTIVE: verb ***pro**čitat'* 'read; read over'

Ja	***pro**čital*	*knigu*	za	dva	dnja.
I	read.PFV.PST	book	in	two	day

'I read the book in two days.'

Grammatical Aspect

Semantics: Klein (1994)

- Aspect is the relation between event and topic time
- PFV aspect
  - event time within topic/reference time
- IMPF aspect
  - topic/reference time within event time or
  - overlap with event time

Grammatical Aspect

Perfective-imperfective opposition

- Derivationally related verbs (Dahl, 1985; Filip, 1993/1999; Filip, 2000; Weimer & Seržant, 2017)
  - "derivational pairs"
  - simplex imperfective – derivationally complex perfective
- Aspectual form and affixation (Filip, 2000, 1993/1999, 2003, 2005)
  - derivational prefixes and semelfactive suffix *-nu-*
  - imperfectivizing suffix*
  - affixes as modifiers of eventuality types denoted by verbal predicates

*The suffix has multiple allomorphic realizations such as *-(o)va-*, *-v-*, etc.

Derivational Pairs: Examples

Simplex IMPF

- *kopat'* ‘to dig’

Dnem	Fedor	kopal	zemlju
day	Fedor	dig.IMPF.PST	ground

‘At daytime Fedor dug ground.’

Derived PFV

- ***za-**kopat'* ‘to dig in, bury’

Piraty	**za**kopali	klad
pirates	dig.PFV.PST	treasure

‘Pirates buried treasure.’

- ***pere-**kopat'* ‘to dig again’

Grjadki	my	**pere**kopali
garden.bed	we	dig.PFV.PST

‘We dug again garden beds.’

- *kop-**nu-**t'* to dig up’

Ja	lopatoj	kop**nul**	čto-to	zvjaknulo
I	spade	dig.PFV.PST	something	clink

‘I dug up with the spade, something clinked.’

Examples from RUSSIAN WEB 2017 (RUTENTEN17) corpus

Prefixes and the Suffix *-nu-*

Extension of lexical meaning

- *za-:* completive or inceptive meaning (i.a.)
- *pere-:* distributive or iterative meaning (i.a.)
- *-nu-*: semelfactive suffix

Change of argument structure

- *kopat' zemlju* 'to dig the ground'
- ***za**kopat' #zemlju* 'to bury #the ground'
  versus ***za**kopat' klad v zemlju* 'to bury a/the treasure into the ground'
- ***pere**kopat' #klad* 'to dig over/again #the treasure'
  versus ***pere**kopat' zemlju* 'to dig over/again the ground'

here is "uninterpretable", "odd", "unacceptable"

Distributional Semantics: Methods

Heinrich Heine Universität Düsseldorf

Linguistic items with similar meanings have similar distributions (Firth, 1957)

- fastText method (Bojanowski et al., 2017)
  - non-contextual (static) word embeddings unique for each word
  - word representations: internal structure of words
- t-SNE: t-Distributed Stochastic Neighbor Embedding (van der Maaten and Hinton 2008)
  - unsupervised non-linear dimensionality reduction technique for exploratory analysis
  - visualization of word embeddings generated by fastText

Distributional Semantics: Methods

Heinrich Heine Universität Düsseldorf

fastText

- Form similarity in addition to context similarity
- Captures lexically similar words
- Internal subword information
  - less-resourced and inflected languages

  (Finnish, Turkish, Russian, etc.)
  - suffixes and prefixes
  - out of vocabulary words
  - infrequent words

t-SNE

- Clustering method
  - clusters of similar points
- Nominal inflection, paradigm cell-filling issues in Finnish and Russian (Nikolaev et al. 2023, Chuang et al., 2023)
  - e.g. model for the conceptualization of Finnish inflected nouns

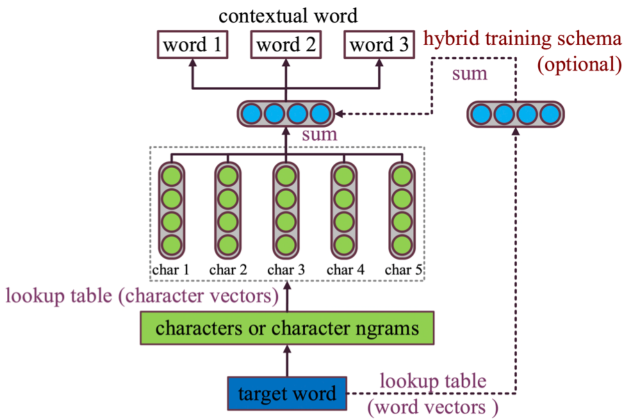
II. Experiment

FastText

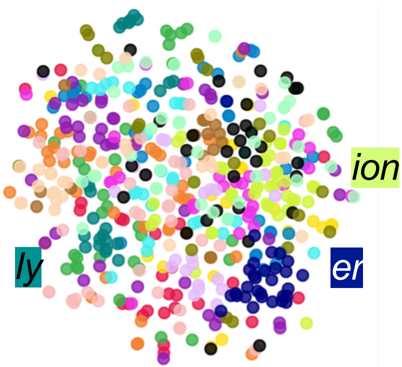
Subword-level model

- Prediction of target word from its form/context
  - word: N-gram characters (subword information)
  - context: sum of all vectors of N-gram characters
  - result: high-dimensional vector space

- T-SNE visualization
  - mapping data from high-dimensional vector space to lower-dimensional space
  - representation of datapoints in a lower-dimensional space (2D plane)
  - visual verification of predicition results via clusters



FastText model (Li et al., 2018: 40)



Cluster visualization of different English affixes by color, FastText model (Li et al., 2018: 44)

Tools

RusVectores pre-trained model (Kutuzov and Kuzmenko, 2017)

- fastText distributional model
  - Continuous Bag of Words (CBOW) architecture
  - Vocabulary size: 195,782 words
  - Vector dimension size: 300
  - Web-corpus Araneum Russicum Maximum 2018: 10 billion words
- *Gensim, sklearn* packages

Derivational Data

Resources

- Exploring Emptiness Database (Janda 2007)
- Database of Russian Verbal Aspect
  (OSLIN database; Borik and Janssen, 2012)
- Essex Database of Russian Verbs and their Nominalizations
  (Essex database; Spencer and Zaretskaya 2017)

Derivational Data

Heinrich Heine Universität Düsseldorf

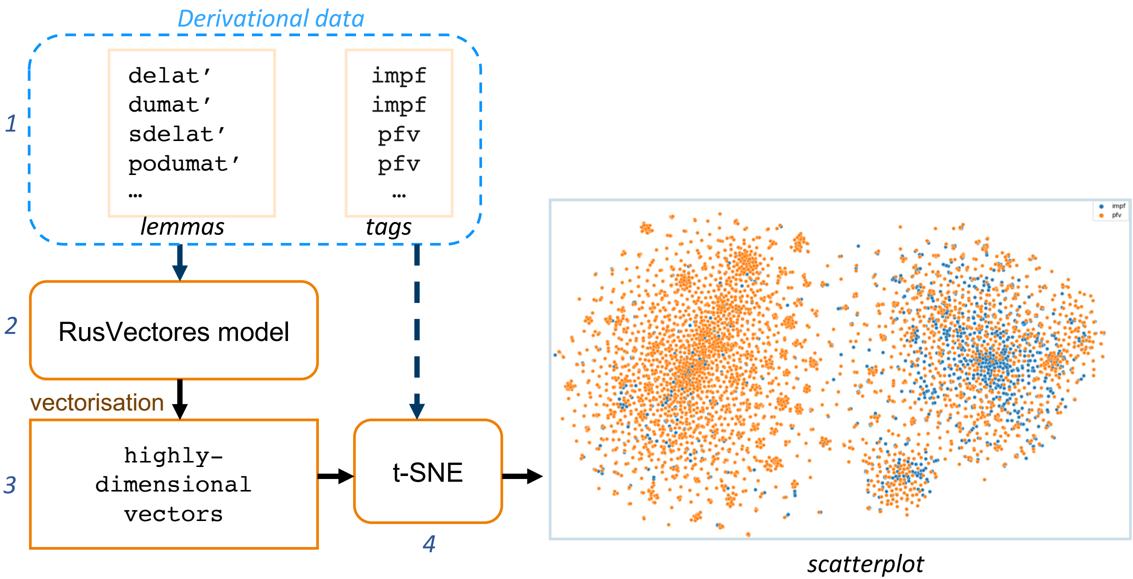
Counts

- 4032 derivational pairs:
  - 3976 prefixes
  - 56 inst. of suffix *-nu-*
- 3986 PFV verbs
- 1766 IMPF verbs

IMPF	PFV	Affix type	Affix
вить	**с**вить	prefix	с
влажнеть	**по**влажнеть	prefix	по
возить	**от**возить	prefix	от
гаркать	гарк**ну**ть	suffix	ну
двигать	дви**ну**ть	suffix	ну+alt
дёргаться	дер**ну**ться	suffix	ну
…	…	…	…

Processing and Prediction

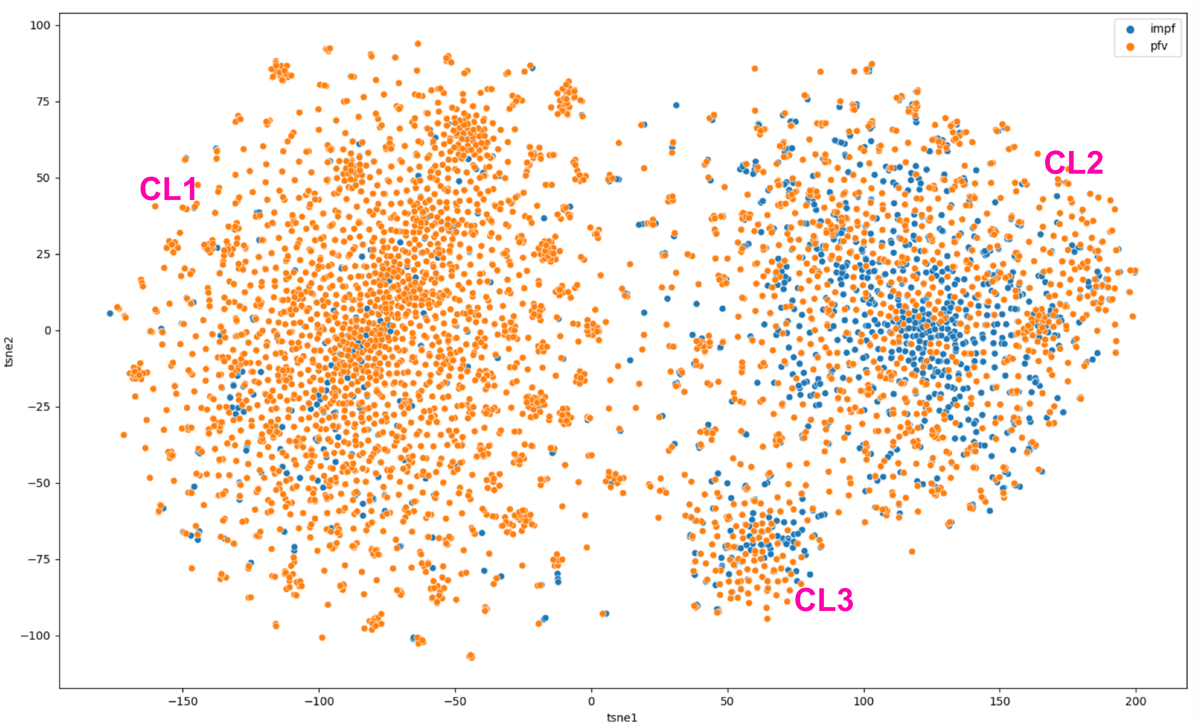
Heinrich Heine Universität Düsseldorf



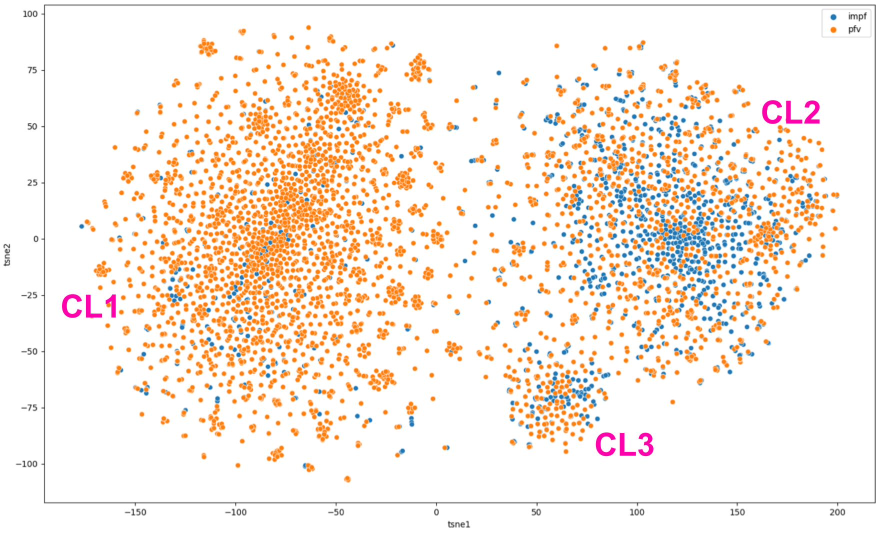
Scattered clusters: Aspect

Heinrich Heine Universität Düsseldorf

Perfective lemmas (CL1), imperfective lemmas (CL2, 3)



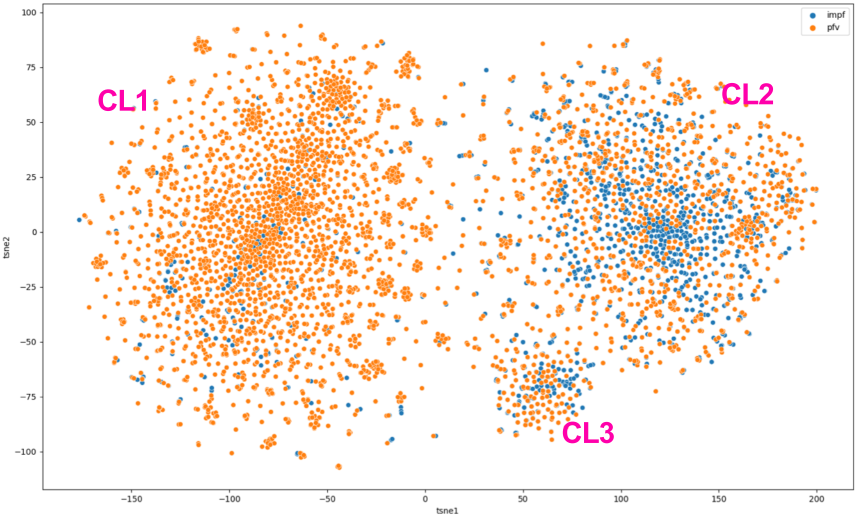
Observation: CL1 and CL2



- Clear-cut separation of PFV and IMPF verbs
  - significant similarity by morphological form

- PFV verbs that co-occur with IMPF verbs in CL2
  - e.g. ***za**molčat'* ‘to get silent’, ***po**buret'* ‘to turn brown’, *mig**nut'*** ‘to wink’

Observation: CL1, 2, 3



Lack of similarity (CL2, 3)

- by morphological form
- by verb semantics
  - (verb classes; Levin, 1993)

Diverse lexical semantic classes

- manner of speaking (*prokvakat'* 'to croak', CL1)
- measure (*sosčitat'* 'to count'; CL1)
- gestures involving body parts (*mignut'* 'to wink [once]'; CL2),
- contact by impact (*užalit'* 'to sting'; CL2)
- creation/transformation (e.g., *vygravirovat'* 'to engrave'; CL3)

Overlap of semantic classes

- psychological state (*zainteresovat'* 'to interest', *pozavidovat'* 'to envy'; CL1, 2)
- psychological state/inchoative (*obozlit'sja* 'to get angry', *obradovat'sja* 'to become glad'; CL1, 2)
- change of state (*prixvornut'* 'to get [a bit] sick', *poburet'* 'to turn brown', *otremontirovat'* 'to repair'; CL1, 2, 3)

Error Analysis

Heinrich Heine Universität Düsseldorf

Corpus frequency effect

- Higher frequency counts
  - higher similarity scores (Asudani et al., 2023)
  - increased model's bias to one category over another (Caliskan et al., 2022, Brunet et al., 2019)
- Araneum Russicum Maximum 2019
- Zipf transformation measure (Van Heuven et al., 2014):
  - low frequency verbs (rank 1, 2, 3)

    raw freq 4 > 1.94 > rank 2 > low freq
  - high frequency verbs (rank 4, 5, 6, 7)

    raw freq 46899 > 5.92 > rank 6 > high freq
- t-SNE clusters of frequency ranks

Biaspectual verbs, CL3

- Borrowed biaspectual *-ova-* verbs
  - function as PFV or IMPF verbs dependent on context
  - *kristallizir**ova**t'* 'to crystallize', *modelir**ova**t'* 'to model', *transkribir**ova**t'* 'to transcribe'

Scattered clusters: Frequency rank

Low-frequency PFV verbs, high-frequency PFV verbs

Error Analysis

Frequency rank bias

- High-frequency PFV verbs: IMPF cluster
  - *e.g. **raz**budit'* ‘to wake up’
- Low-frequency PFV verbs: PFV cluster
  - *e.g. **za**torcevat'* ‘to pave with wood blocks’

Error Analysis

Context similarity

- Biaspectual borrowed verbs, which are integrated into the Russian verb system with the suffix *-ova-*
  - e.g. *orientirovat'* ‘to orient(ate)’, ‘to guide’, ‘to aim’, ‘to walk s.o. through’, *kooperirovat'sja* ‘to cooperate, to partner with’
- They co-occur with derived prefixed PFV counterparts in CL3
  - e.g., ***s****orientirovat'*, ***s****kooperirovat'sja*

Heinrich Heine Universität Düsseldorf

III. Conclusions and future work

Conclusions

Correct prediction of Russian aspect by RusVectores model

- Morphological structure: significant criterion for determing the grammatical aspect of the verb
  - Lexical semantic classes: unsignificant criterion
- CL2: fastText model's bias to high-frequency PFV verbs
- CL3: context similarity for borrowed biaspectual *-ova-* verbs and their PFV derivates

Future work

- Cosine similarity scores using different word embedding model (e.g. contextual ELMo model)
- Identify best predictors of aspect for derivational pairs, e.g.
  - Random Forests
  - Generalized logistic regression
- Prediction of sematic relations between derived perfectives and their imperfective bases (see e.g. Bonami and Naranjo, 2023)

References

- Bonami, O., & Naranjo, M. G. (2023). Distributional evidence for derivational paradigms. *The semantics of derivational morphology: Theory, methods, evidence*, 219- 258.Peters, M. E., Neumann, M., Zettlemoyer, L., & Yih, W. T. (2018). Dissecting contextual word embeddings: Architecture and representation. arXiv preprint arXiv:1808.08949.
- Borik, O., & Janssen, M. (2012). A database of russian verbal aspect. In *Proceedings of the conference Russian Verb*, St. Petersburg, Russia.
- Brunet, Marc-Etienne, Alkalay-Houlihan, Colleen, Anderson, Ashton and Zemel , Richard. 2019. Understanding the origins of bias in word embeddings. In *International Conference on Machine Learning*. PMLR, 803–811.
- Caliskan, A., Ajay, P. P., Charlesworth, T., Wolfe, R., & Banaji, M. R. (2022, July). Gender bias in word embeddings: a comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 156-170).
- Dahl, Ö. (1985). *Tense and Aspect Systems*. Oxford: Blackwell.
- Drozd, A., Gladkova, A., & Matsuoka, S. (2015). Discovering Aspectual Classes of Russian Verbs in Untagged Large Corpora. In *2015 IEEE International Conference on Data Science and Data Intensive Systems*, (pp. 61–68).
- Filip, H. (2000). The Quantization Puzzle. *Events as grammatical objects, from the combined perspectives of lexical semantics, logical semantics and syntax*, *39*, 39–91.
- Filip, H. (2003). Prefixes and the Delimitation of Events. *Journal of Slavic linguistics*, *11*(1), 55–101.
- Filip, H. (2005). On accumulating and having it all: Perfectivity, prefixes and bare arguments. *Perspectives on Aspect. Studies in Theoretical Psycholinguistics*, *32*, 125–148.
- Filip, H. (1993/1999). *Aspect, situation types and nominal reference*. Ph.D. Thesis, University of California at Berkeley. New York/London: Garland

References

- Firth, J. R. (1957). A Synopsis of Linguistic Theory, 1930-55. *Studies in Linguistic Analysis*, *Special Volume of the Philological Society*, 1–31.
- Klein, Wolfgang. (1994). *Time in Language*. Routledge, London.
- Kustova, G. I., Lashevskaja, O. N., Paducheva, E. V., & Rakhilina, E. V. (2009). Verb Taxonomy: from Theoretical Lexical Semantics to Practice of Corpus Tagging. *Studies in Cognitive Corpus Linguistics*, (pp. 41–56).
- Kutuzov, A., & Kuzmenko, E. (2017, April). Building web-interfaces for vector semantic models with the webvectors toolkit. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 99-103).
- Janda, Laura A. (2007). Aspectual clusters of Russian verbs. *Studies in Language 2007*. Volume 31 (3). S. 607-648.
- Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.
- Li, B., Drozd, A., Liu, T., & Du, X. (2018, June). Subword-level composition functions for learning word embeddings. In *Proceedings of the second workshop on subword/character level models* (pp. 38-48).
- Spencer, Andrew and Zaretskaya, Marina (2017). *Russian nominalizations database*. [Data Collection]. Colchester, Essex: UK Data Archive. 10.5255/UKDA-SN-852633
- van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-uk: A New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190.
- Wiemer, B., & Seržant, I. A. (2017). Diachrony and typology of Slavic aspect: What does morphology tell us. *Unity and diversity in grammaticalization scenarios*, *16*, 239.

Appendix

Corpus Frequency Transformation

- Frequency rank of a lemma in a corpus
  - log-transformation of raw frequencies counts
  - scaling of log-transformed frequency values
  - frequency rank of lemmas (low frequency, high frequency; van Heuven et al., 2014)

blestet' ‘shine’ (IMPF verbal lemma)
corpus frequency: 756

vypisat' ‘write down’ (PFV verbal lemma)
corpus frequency: 1019

logarithmic transformation

4.66510116

4.253130176

Frequncy Rank

Van Heuven et al. SUBTLEX-UK (2014: 1180)

Table 1. *The Zipf scale of word frequency*

Zipf value	*fpmw*	*Examples*
1	0.01	antifungal, bioengineering, farsighted, harelip, proofread
2	0.1	airstream, doorkeeper, neckwear, outsized, sunshade
3	1	beanstalk, cornerstone, dumpling, insatiable, perpetrator
4	10	dirt, fantasy, muffin, offensive, transition, widespread
5	100	basically, bedroom, drive, issues, period, spot, worse
6	1000	day, great, other, should, something, work, years
7	10,000	and, for, have, I, on, the, this, that, you

Note: The Zipf scale is a word frequency scale going from 1 to 7. Words with Zipf values of 3 or lower are low-frequency words; words with Zipf values of 4 and higher are high-frequency words. Examples are based on the SUBTLEX-UK word frequencies. fpmw = frequency per million words.

1, 2, 3 – low-frequency
4, 5, 6, 7 – high-frequency

Zipf Transformation

hhu
Heinrich Heine
Universität Düsseldorf

```
Zipf = log10((rawFreqSmooth/(corpusSizePm + typesNbPm)) + 3.0
```

Zipf smoothing   Laplace smoothing   scaling for ipm values

Lemma	Raw Freq	Laplace smoothing	Zipf Smoothing	Scaling (+3)	Rank
бацнуть	16	17	-0,5250134	2	Low freq
выездить	64	65	0,05745064	3	Low freq
выдрессировать	1521	1522	1,42694083	4	High Freq
вызубрить	2099	2100	1,56674106	5	High Freq
выкупить	99544	99545	3,24179872	6	High Freq
…	…	…	…	…	…

Corpus examples

- PERFECTIVE: verb ***pere**čitat'* ‘reread’

Ja perečital ètu frazu raz desjat' .

I **read.**pfv.pst this phrase time ten

‘I **reread** this phrase ten times.’

- PERFECTIVE: verb ***na**čitat'* ‘read (aloud)’

Igor' načital 3 knigi Verdona.

Igor' ***read**$_{PFV.PST}$* three book Verdon

‘Igor' *has read/read three Verdon’s books.’*

Examples from RUSSIAN WEB 2017 (RUTENTEN17) corpus