Uliana Petrunina, doctoral fellow, University of Tromsø

Ambiguity of Russian participles on word/context levels and use of frequency for its resolution

The topic of the paper is the ambiguity (homonymy) of participles in Russian, produced by an affixless means of converting a participle into an adjective, e.g.: *neopoznannyj* ***letajuščij*-adjective** *ob"ekt* "unidentified **flying** object" vs. ***letajuščij*-participle** *za oknami pux* "[poplar] fluff **flying** outside the window". It stands out among other types of homonymy in Russian (Appendix A) as it results from *adjectivization*, one type of conversion[1] marked by the gradual process of acquisition of the morphosyntactic adjectival properties and loss of the verbal ones, with an optional modification of semantics (Say 2016). The instances of adjectivized participles appear to be frequent in Russian texts: in my pilot corpus study, I extracted about 55% of ambiguous participial wordforms out of total participial wordforms in the Webcrawl 2008 corpus (Appendix B).

In this paper, I will explore the effect of **corpus frequency**, **transitivity** and **voice** of verbal lemmas on forming participles in general and participles that *become adjectivized* in my corpus experiments. I will also touch upon syntax and semantics in adjectivization and the disambiguation using frequencies from the frequency dictionary for Russian[2] (Sharoff 2002).

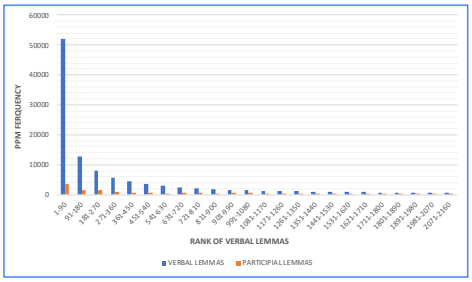

Figure 1: Verbal and participial lemmas sorted by the rank of verbal lemmas frequency (interval 1 -2160)

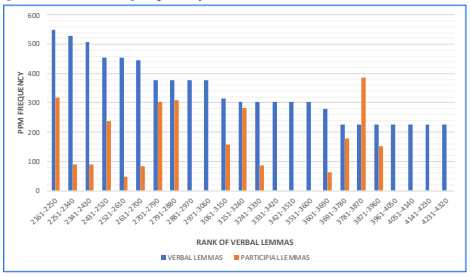

Figure 2: Verbal and participial lemmas sorted by the rank of verbal lemmas frequency (interval 2161 - 4320)

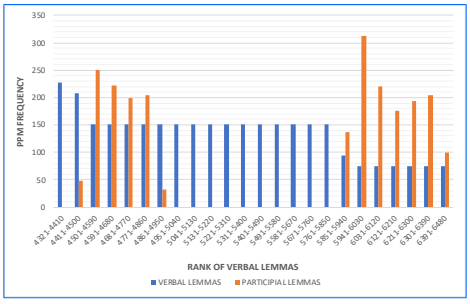

Figure 3: Verbal and participial lemmas sorted by the rank of verbal lemmas frequency (interval 4321 - 4320)

I extracted the occurrences of ambiguous/unambiguous participles and verbal lemmas from the gold standard of the Russian National Corpus (1,190,540 wordforms)[3]. The occurrences of words with participial suffixes (Appendix C) in the corpus were extracted and annotated with the morphological analyzer *analyser-disamb-gt-desc.hfstol*[4] without disambiguation. The annotated wordforms were lemmatized and grouped as: *participial lemmas* (verbal lemmas for ambiguous participles (with more than one reading) and unambiguous participles), *verbal lemmas* (finite/infinite forms only). The verbal lemmas that matched the participial lemmas and the frequencies of wordforms of the verbal and their matching participial lemmas were sorted in ascending order and ranked from 1 to 8900 (verb rank, narrowed down to 6480 as beyond that, the frequencies equaled 1 and 0). The frequencies were transformed into part per million (PPM) values. The distribution of participial and verbal lemmas sorted by the increasing rank of verbal lemmas, split into three intervals (90 lemmas/subinterval), are shown in Fig. 1, 2 and 3 (see the entire graph in Appendix D, Fig. 6).

Fig. 1 includes high-frequency verbal/ participial lemmas. The verbal lemmas are around 25.9 times (per subinterval) more frequent than the participial lemmas. Among the top ten lemmas, *moč'* "be able to do smth" has only 1 present active participle *moguščego "being able"* and *byt'* "be" is 71.7 times more frequent than the participial lemma (with the adjectivized form *buduščij* "future"). Most of these ten verbal lemmas have active participial lemmas only, except *skazat'* "say" and *govorit'* "*speak"* which have both active and passive participial lemmas. The verbal lemmas within the range of 1 – 100 that have 0 participial lemmas are: *vremenit'* "delay", *teret'* "rub", *sprosit'* "ask a question", *našit'* "sew on", *smoč'* "manage". As a native speaker, I regard their aspect and semantics to limit both the frequency of use and formation of participles. The participial lemmas forming adjectivized participles (in active voice) are idiomatized, e.g.: *govorit'* "*speak"*=> ***govorjaščaja*** *familija* "self-explanatory surname", *sledovat'* "follow" => ***sledujuščij*** *god* "next year", *ponimat'* "understand" => ***ponimajuščaja*** *ulybka* "understanding smile" (metonymy).

Fig. 2 indicates that: (a) verbal lemmas (per subinterval) are around 1.2 times more frequent than participial lemmas, at average, (b) 58.3% of all subintervals contain more verbal than participial lemmas, (c) 4.2% of all subintervals contain more participial than verbal lemmas, (d) 37.5% of all subintervals contain verbal lemmas without participial lemmas[5]. The participial lemmas also include adjectivized and idiomatized forms, e.g.: *potrjasat'* "astonish" => ***potrjasajuščix*** *glazax* "gorgeous eyes" (metonymy), *osmyslit'* "apprehend" => ***osmyslennaja*** *politika* "prudent policy". Most verbal lemmas on this interval have both active and passive participial forms.

Fig. 3 reveals four types of patterns: (a) verbal lemmas are about 0.4 times at average more frequent than participial lemmas, (b) 45.8% of all subintervals include verbal lemmas without participial lemmas, (c) 50% of all subintervals contain more participial than verbal lemmas, (d) 4.2% of all subintervals contain more verbal than participial lemmas. Most verbal lemmas have both active and passive participial lemmas. Most passive participles tend to be adjectivized while the active ones tend to remain unambiguous, e.g.: PASSIVE PTCP: *avtomatizirovat'* "automate" => ***avtomatizirovannogo*** *kontrolja* "automated control", vs. ACTIVE PTCP: *černet'* "turn black" => ***černejuščie*** *prižarenii* "tuning black when fried".

---

[1] Conversion is an unmarked change of syntactic function and semantics of a word (Schönefeld 2005). In Russian it also includes substantivization and adverbalization.

[2] Available : http://www.artint.ru/projects/frqlist/frqlist-en.php

[3] Available: http://www.ruscorpora.ru/en/corpora-usage.html

[4] *analyser-disamb-gt-desc.hfstol* is the morphological analyzer for Russian, giving a morphological and syntactical analysis of words in a text. It is developed on the basis of Helsinki Finite-State Transducer, available: http://www.ling.helsinki.fi/kieliteknologia/tutkimus/hfst/documentation/index.html

[5] The absence of the corresponding participial lemmas can be explained by their absence in the gold standard.

The observations above show that high-frequency basic verbs (within the range of 1 – 100) appear to form participles reluctantly but continuously. Most frequent verbal lemmas have active participial forms which can be adjectivized, with the additional extension of semantics as in *govorjaščaja familija* "self-explanatory surname". There is less difference between the values of verbs and participles on the second interval, verbal lemmas can have both active and passive participles that can be adjectivized (and idiomatized as well). On the third interval, the difference between the frequency of verbs and participles becomes even and low, and mostly passive participles are adjectivized, without a considerable extension of semantics.

These patterns lead to the question of whether and how internal properties of a given lexeme can be potentially related to its corpus frequency. I followed this thread by exploring transitivity and voice (see tense and aspect in Appendix D, Fig. 7 and 8). This time, I extracted verbal lemmas, their transitivity and voice features from the gold standard, matched them with the participial lemmas, calculated the arithmetical ratio (PPM frequency of verbal lemmas/PPM frequency of participial lemmas) and sorted them by the increasing rank of the verbal lemmas. The distributions of the ratios are shown in Fig. 4 and 5.

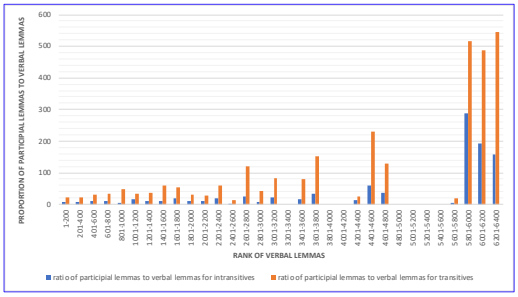

Figure 4: Distribution of the ratio of participial lemmas to verbal lemmas for transitive/intransitive forms

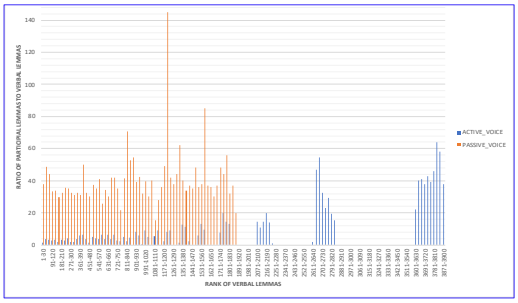

Figure 5: Distributions of the ratio of participial lemmas to verbal lemmas for active/passive voice

Fig. 4 indicates that the ratio of participial lemmas to verbal lemmas for transitives is considerably higher than the one for intransitives, across the whole ranking scale. In other words, transitive verbs from the gold standard form much more participles than intransitive ones. This is due to the fact that transitive verbs can have both active and passive forms while intransitive verbs can only have active forms.

Fig. 5 illustrates that the ratio of participial to verbal lemmas for passive voice is much higher than the ratio for active voice. This can be explained by the following: (a) highly frequent verbal lemmas have much less participial lemmas than mid- and low-frequency lemmas, (b) they have more transitive than intransitive participial lemmas, (c) transitive lemmas can form active/passive lemmas while intransitive lemmas can only form active ones. The question is if transitive lemmas tend to have more passive than active forms only in this corpus? The distributions of the ratio of participial lemmas to verbal lemmas for active/passive voice are superposed. The rank corresponds to both types of distinct lemmas: lemmas for active voice and lemmas for passive voice. The passive distribution ends at the point 1830 while the active voice distribution continues until the point 3900. So, 2070 remaining low-frequency verbal lemmas (that do not overlap with the ratio of passive participles) have active participles only or no participles at all.

I will sum up the observations: 1) highly frequent transitive passive lemmas have less participles per lemma than low-frequency transitive active lemmas, 2) highly-frequent verbal lemmas have a more steady number of participles than mid and low-frequency ones, 3) highly-frequent active verbal lemmas have more adjectivized participles than low-frequency active lemmas, although most low-frequency passive participles tend to be adjectivized. These tendencies give just a representation of how frequency, transitivity and voice can account for production of participles (including adjectivized ones) by verbal lemmas.

To identify the significance of effect of rank, transitivity, tense and voice on ambiguity, I used the *glm* (generalized linear model) function to fit four models (see Appendix E)[6]with the following effects:

Rank model: ambiguity(ambiguous, unambiguous) as a dependent variable explained by rank, ration as independent variables

TransAsp model: ambiguity as a dependent variable explained by transitivity (transitive, intransitive), aspect (perfective, imperfective), rank (1, 2, 3, 4, 5, 6, 7) and ratio as independent variables

Tense model: ambiguity as a dependent variable explained by tense (past, present), aspect, rank and ratio as independent variables

Voice model: ambiguity as a dependent variable explained by voice (passive, active), aspect, rank and ratio as independent variables

The rank represents smoothed frequency of verbal lemmas on the scale of the values from 1 to 7, on which 1-3 are low-frequency values and 4-7, high-frequency values (van Heuven et al. 2014), ratio is the ratio of frequency values of participial to verbal lemmas The models produced the following results:

1. High-frequency ranks 4 and 5 are a statistically significant predictor of ambiguity of verbal lemmas
    a. interaction of rank with ratio, tense and voice is statistically significant in predicting ambiguity
2. Overall transitivity is not significant but **transitive** verbal lemmas and **perfective** aspect are statistically strong predictors
3. Overall tense and voice are not significant but **present tense** has a significant effect as well as the interaction of factors of passive voice and rank 4.

The preliminary analysis indicates that there is a significant relationship between ambiguity and high rank of verbal lemmas, their transitivity, perfective aspect and present tense. Besides, the interaction of high rank (mostly 4) with transitivity, aspect and tense is also statistically significant for ambiguity. The results from this analysis could be used for deeper contextual analysis of ambiguous participles to reveal more patterns of their adjectivization . I can also be used for disambiguation tasks, such as design of disambiguation rules dependent on morphological properties of wordforms and their frequencies.

Apart from these morphological features, verb semantics (e.g.: stativity, duration, motion, etc.) can also account for predisposition of certain verbs towards adjectivization. For instance, this affects their compatibility with certain types of

[6] I also tested the goodness of fit of these models with null models using Likelihood ratio test (LRT), the resulting probabilities were similar to the probabilities produced by the models.

adverbs, e.g.: the adverb of measure *očen* "very" (used with adjectives) can combine with the verb *spešit': [on] očen' spešil* "[he] was in a great hurry"(='acted quickly') but it cannot combine with the verb *xodit'* "go, walk" (Sičinava 2011).

Adjectivization is a gradual process of transition of a lexeme from the verbal to the adjectival paradigm. It may manifest in a context by the absence of syntactic parallelism[7], compatibility with adverbs of measure and degree/comparative and superlative degrees, position before or after a (pro)noun, absence of localization in time and space, loss of ability to join verbal dependents, idiomatic augmentation or the interaction of one or more of these factors with each other (Say 2016). I hypothesize that high frequency of verbal lemmas (combined with their semantics/morphology) can be a constraint for them to form steadily fewer participles than other verbal forms whereas low-frequency can be a constraint for them to form larger number of (adjectivized) participles than verbal forms sporadically.

How can morphosyntactic/frequency features fit in the task of disambiguating texts with adjectivized participles? Constraint Grammar (CG)[8] offers options for capturing both word and context levels (morphosyntax), and for using frequency features. The latter is problematic as it depends on corpus parameters that can influence the reliability of frequencies for disambiguation: a corpus size, types (written/oral), genres, frequency threshold, quality/availability of annotation. I analyze the Russian corpus data with an open source finite state transducer for Russian, *giella-rus*[9] (Tyers & Reynolds 2015) and CG for disambiguation. I have annotated the *giella-rus* lexicon with frequency weights weights[10] for the Russian CG. The selection of the appropriate morphological reading by the weight criterion is done by a rule `SELECT:maxweight (<W=MAX>);` meaning "select the reading with the greatest weight" and applied if all the preceding CG rules fail to disambiguate. This rule lets the analyzer straightforwardly sort results that were not disambiguated by the regular CG rules, e.g.:

```
"<напряженную>" ambf
        "напрячь" V Perf TV PstPss Lxc Fem AnIn Sg Acc <W:4.4885711670>;
        "напряжённый" A Fem AnIn Sg Acc <W:0.0027313232>
```

The weighting rule will thus select the reading `V Perf TV PstPss` because it has the highest weight `4.4885711670`.

The CG rules for disambiguating participles consist of the following types:

- fine-grained REMOVE rules for specific contexts: if there are adjectival and participial readings given a specific context described by a rule, remove adjectival readings or remove verbal readings
- coarse-grained SELECT rules for general contexts: if there are adjectival and participial readings, select adjectival readings or select participial reading
- straightforward 'sort' weighting rule: select the reading with the maximal weight

After implementing weights, adding a sublexicon for lemmas of participial forms with their weights and developing CG rules I carried out an experiment on the sample of texts from the SynTagRus corpus[11]. It is a dependency treebank with morphological/syntactic annotation, humanly corrected. I extracted sentences containing all instances of participles and adjectives from the file *ru_syntagrus-ud-train.conllu*, randomly shuffled the texts and selected 297 sentences containing the instances of participles and adjectives. I annotated these sentences with *analyser-disamb-gt-desc.hfstol* using disambiguation option *vislcg3 - disambiguator.cg3 -t* using several options: (a) CG rules only, (b) CG rules combined with weights, (c) weights only. Out of 297 sentences there were 103 sentences containing 119 instances of ambiguous participles which then received my manual tags and the tags from the SynTagRus. These instances include: 15 occurrences of substantivized participles and 104 occurrences of adjectivized participles. The uneven number and mixed types of ambiguity is due to the fact that I did not predefine their number before processing the SynTagRus corpus. I used the following notations to evaluate the tagging:

ambt: true positives, i.e. wordforms tagged correctly by the analyzer

ambf: false positives, i.e. wordforms tagged incorrectly by the analyzer

ambn: false negatives, i.e. wordforms not disambiguated by the analyzer, completely or partially (with more than 1 tag left after disambiguation)

? (as in *ambt?*, *ambf?*): additional marker for problematic or dubious readings

The results of the annotation and of the evaluation measures are given in Tables 1 and 2.

Table 1

	CG rules	**CG rules + weights**	**weights**	**SynTagRus**
amb	119	119	119	119
ambt (true positive)	76	83	61	82
ambf (false positve)	16	33	54	37
ambn (false negative)	27	3	3	0
ambt + ambf	92	116	115	119

[7] It is the possibility to transform a participial clause into a relative/finite verbal clause, e.g.: *the students reading the journal => the students that are reading the journal/ the student are reading the journal*.

[8] CG is a formalism, introduced by Karlsson (1990), using the *vislcg3* compiler for disambiguating readings. This compiler is a CG compiler implemented by VISL, a research and development project at University of Southern Denmark (SDU) (http://beta.visl.sdu.dk/cg3.html)

[9] Available: http://giellatekno.uit.no/doc/lang/rus/rus.html

[10] In my study, weights are corpus frequencies of verbal, adjectival and nominal lemmas taken from the frequency dictionary for Russian (see p. 1) and logarithmically transformed for smoothing difference between freq. values of lemmas on the scale from 1 (least frequent) to 7 (most frequent) using the equation for smoothing and scaling frequency values of the SUBTLEX-UK database by van Heuven et al. (2014) .

[11] Available: https://github.com/UniversalDependencies/UD_Russian-SynTagRus

Table 2

Options	precision	recall	f-score	ambiguity solved (accuracy) %
CG rules	0,83	0,74	0,78	63,87
CG rules + weights	0,72	0,97	0,82	69,75
weights	0,53	0,95	0,68	51,26
SynTagRus	0,69	1,00	0,82	68,91

Table 1 shows the counts for instances manually annotated with *ambt*, *ambf*, *ambn* notations. Weights (3rd column) indicates that using weights only allowed to disambiguate almost all instances (except for 3 tokens), 61 wordforms were disambiguated correctly, less than half 54 wordforms, incorrectly. Table 2 illustrates the scores of the evaluations measures for each option of disambiguation. Disambiguation with **CG rules** only shows the highest precision (0.83) but the lowest recall () which means that CG rules produced the lowest number of incorrect readings but also the lowest number of correct readings. The **CG rules + weights** option showed the second highest recall (0.97), the highest f-score (0.82) and accuracy (69.75%) which means that it produced the highest number of correct analyses. The **weights** option shows the second highest recall but the lowest precision, f-score and accuracy which indicates that it disambiguated most instances of ambiguous participles but did it incorrectly in less than half of the cases. This may imply that using weights alone can be risky but combining together with CG rules increases the number of true positives and false positives and reduces the number of false negatives The **SynTagRus** option showed the highest recall and the second highest f-score and accuracy as it disambiguated all instances of the ambiguous participles with the correct analysis for most of them.

The major source for incorrectly analyzed instances (for **CG** and **CG + weight** options) is due to:

1. adjectives being identified as participles, e.g.: *igrat'_V matč_N protiv_PREP* ***dejstvujuščego_ADJ*** *čempiona_N* "to play a match against **the current_ADJ** champion"
2. nouns being identified as participles, e.g.: *sud'ba_N ostal'nyx_ADJ soten_NUM tysjač_NUM i_CC millionov_NUM* ***pogibšix_N*** "the fate of the rest hundreds and millions of **casualties_N**"
3. participles being identified as adjectives, e.g.: *on_PRON* ***vynužden_PTCP byl_V*** 'he **was obliged_PTCP**'

The major source for correctly analyzed instances of participial forms is due to their recognition in constructions such as:

a. verbal forms + prepositions, e.g.: *i_CC* ***orientirovan_PTCP na_PREP*** *praktičeskoe_ADJ spasenie_N usadeb_N* "and **oriented_PTCP towards_PREP** the practical rescue of farms"
b. present/past passive verbal form + Instrumental case, e.g.: *postavok_N,* ***reguliruemyx_PTCP*** *proizvoditeljami_N+Ins* "supplies **handled_PTCP** by the manufacturers"
c. the copula *byt'* 'be' + verbal form, e.g.: ***bylo_Vbyt' prinjato_PTCP*** *odobritel'no_ADV* "[suggestion] **was_Vbyt' taken_PTCP** favorably"

The selection of adjectival readings was not consistent, mostly because of the CG rules appeared not to recognize the agreement between an adjectival form and nominal/pronominal forms they modify. Otherwise, adjectival forms were recognized:

a. in the position after another adjective, e.g.: *žitelej_N* ***bližajšix_ADJ naselennyx_ADJ*** *punktov_N* "inhabitants from **nearest_ADJ [inhabited] _ADJ** settlements"
b. by selection of the maximal weight, e.g.: ***okružajušče*** *sredy_N* "**[surrounding]** environment" wherein ***okružajuščej _ADJ*** has the weight 4.5574188232 and ***okružajuščej_PTCP*** has the weight 3.7054901123, so ***okružajuščej_ADJ*** is selected

I will also highlight the performance of weights in disambiguation. Their accuracy rate is 51.26% which implies that in about half of all the cases the disambiguation will be correct. The most successful instances of weight disambiguation concern:

- selection of participial forms, e.g.: *vraždebnost'_N,* ***suščestvujuščaja_PTCP*** *meždu_Prep učastnikami_N+INS* 'hostility **existing** among the participants'
- selection of adjectives which frequency outnumbers considerably the frequency of verbal lemmas or lemmas for participial lexicalized participles, e.g.: ***zadannoj_ADJ*** *formy_N* "predefined shape", ***okružajuščuju_ADJ*** *sredu_N* "[surrounding] environment", ***sledujuščej_ADJ*** *oseni_N* "**next_ADJ** autumn", ***obsluživajuščij_ADJ*** *personal_N* "[operational] staff"
- copular constructions, e.g.: ***byl_Vbyt'*** *široko_ADV* ***rasprostranen_PTCP*** "**was_Vbyt'** widely **spread_PTCP"**, **byl_V*byt'* napravlen_PTCP** na_PREP "**was_Vbyt' directed_PTCP** at"

The most successful rules identified copular constructions, prepositional phrases and verbal forms followed by a part of speech in Instrumental cases. The least successful rules were the rules for removing verbal readings: they tended to sort out reading that were supposed to be participial instead of adjectival one. The use of weights increased the number of correctly disambiguated wordforms because of the dominant frequency of adjectival wordforms which were not recognized by CG rules without weights.

The experiment does not indicate if the CG rules + weights model will outperform the SynTagRus over the large set of data. It shows that the combination of weights and CG rules is beneficial for disambiguation as it decreases the number of false negatives and increases the number of true positives.

References

- Ahmanova, O. 1974. *Slovar' omonimov russkogo jazika*. Izdatel'stvo Sovetskaja entsiklopedija, Moskva. 448 p.
- Belousov, V. N., Kovtunova, I. I., Kručinina, I. N. 1989. *Kratkaja russkaja grammatika*. 639 p.
- Derbyshire, W.W. 1967. Verbal homonymy in the Russian language. In: *Canadian Slavonic Papers / Revue Canadienne des Slavistes*, Vol. 9, No. 1. Pp. 131-139.
- Karlsson, F. 1990. Constraint grammar as a framework for parsing running text. In: *Karlgren, Hans (ed.), Proceedings of 13th International Conference on Computational Linguistics*, volume 3. Pp. 168-173.
- Klepousniotou, E. 2002. The Processing of Lexical Ambiguity: Homonymy and Polysemy in the Mental Lexicon. In: *Brain and Language 81*. Pp. 205 – 223.
- Koskela, A., Murphy, M. 2006. Polysemy and homonymy. In: *Brown, Keith, Anderson, Anne H, Bauer, Laurie, Berns, Margie, Hirst, Graeme and Miller, Jim (eds.) Encyclopedia of language and linguistics (2nd ed)*. Elsevier. Pp. 742-744.
- Manova, S. 2011. *Understanding Morphological Rules: With Special Emphasis on Conversion and Subtraction in Bulgarian, Russian and Serbo-Croatian*. Studies in Morphology. Springer Netherlands. 239 p.
- Say, S. 2016. Pričastie (v pečati). In: *Materialy k Korpusnoj grammatike russkogo jazyka*. Retrieved March 20, 2018, from: http://rusgram.ru/Причастие#52
- Schönefeld, D. 2005. Zero-derivation–functional change–metonymy. In: *Approaches to conversion/zero-derivation*. Pp. 131–159.
- Sharoff, S. 2002. Meaning as use: exploitation of aligned corpora for the contrastive study of lexical semantics. Proc. of Language Resources and Evaluation Conference (LREC02). May, 2002, Las Palmas, Spain.
- Sičinava, D. V. 2011. *Narečie*. Retrieved March 20, 2018, from: http://rusgram.ru/Наречие
- Tyers, F. M., Reynolds, R. 2015. A preliminary constraint grammar for Russian. In: *Proceedings of the Workshop on "Constraint Grammar - methods, tools and applications" at NODALIDA 2015, May 11-13, 2015, Institute of the Lithuanian Language, Vilnius, Lithuania*. Pp. 39 – 46.
- van Heuven, W. J. B., Mandera, P., Keuleers, E .and Brysbaert, M. 2014. Subtlex-uk: A new and im-proved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, 67(6):1176–1190. PMID: 24417251.
- Vinogradov, V. V. 1960. Ob omonimii i smezhnyh javlenijah. In: *Voprosy jazykoznanija , IX No. 5. Izdatel'stvo akademii nauk SSSR*. Pp. 3 - 17.
- Zalizniak, A. A. 1977. *Grammatičeskij slovar' russkogo jazyka*. 880 p.

APPENDIX A

Table 1: Types of morphological ambiguity including the descriptions and examples listed in Ahmanova (1974), Vinogradov (1960), Tyers and Reynolds (2015), Derbyshire (1967)

morphosyntactically incongruent (PARTIAL HOMONYMY)	***in between morphosyntactically congruent and incongruent ambiguity***	***morphosyntactically congruent (FULL HOMONYMY)***
HOMONYMS: **verbs with nouns**: *moč* (“be able” and “power”), *dulo* ("blow" and "muzzle"); **imperatives with nouns**: *glad*' from *gladit*' (“press” and “glossy surface”) **verbs with adverbs:** počti from *počtit*' (“honour” and “almost”) **imperative verbs with numeral**: *pjat*' from *pjatit*' (“move backward” and “five”), *tri* from *teret*' (“wipe” and “three”) **past tense forms with the short forms of adjectives:** *smel* from *smet’* (“dare”) and *smelyj* (“daring”) **imperatives with adjectives:** *moj* from *mit’* (“wash”) and “my” **HOMOGRAPHS:** **Past tense forms agreeing with feminine/masculine/neuter nouns**: *nachálo* ("beginning") and *náchalo* ("it started") *žíla* ("vein") and *žilá* ("(she) lived") **Verb in the 1st person singular with masculine nouns in the dative singular or feminine nouns in the accusative singular**: béregu ("bank") and beregú ("I keep") *prístan*' ("harbor") and *pristán’* ("Stick to") **Feminine accusative adjectives in the singular and the first person of singular of verbs:** celúju ("I kiss") and céluju ("the whole")	**Verbs do not agree in infinitive/imperative forms, tense or person.** **HOMOGRAPHS:** **verbs agreeing in several forms of the present tense but not in the infinitive:** *krojú* from *krojit*' ("cut") and *króju* from *kryt*' ("cover") **HOMONYMS:** **Verbs agreeing in present/future conjugation but not in the infinitive form:** *Dobreju* from *dobrit’* (“finish shaving”) and *dobret’* (“become kinder”) **Verbs agreeing in the 1st person singular:** leču from lečit' ("treat") and letet' ("fly") *melju* from *melit*' ("rub with chalk" or "make small") and molot' ("mill") **Clash of imperatives:** *vej* from *vejat’* ("fan") and *vit*': ("twist") **Verbs agreeing in infinitives but not in present/future conjugation:** **Suffix alternation –a > -aj** *žat*': from ("reap" and "squeeze") klepat' from ("rivet" and "slander") *zret*': "reapen" and "behold"	**HOMONYMS:** **Verbs agreeing with verbs: vowel/consonant alternation, no aspect match:** Suffix ***-yvat’—ivat***' + consonant alternation(**s> š**) or vowel alternation (**o > a**) in a verb stem: *domešivat*' (imperfective)from *domešat*' “finish mixing” and *domesit’* “finish kneading” (perfective) *zasalivat*' from *zasalit*' (soil) and *zasolit’* (salt down) (imperfective) zakapyvat’ (from zakopat’ (dig) and zakapat’ (begin to drip)) **Homonymous prefixes:** **No aspect match:** *sxodit’* ("to go to" (perfective) and "to go down" (imperfective)) **Aspect match (in these examples, the meanings of the verbs are remotely related, so there is a question of whether it is relevant to view them as homonymous words)** homonymous morphemes (represented by such prefixes as *c-, po-, ob-, na-, pere-*) can develop homonymy (Vinogradov 1974): **na-:** *nakolot*' with quantitative and spacial meaning (“chop wood/carve patten” and “pin up a badge”) *nastroit*' (“tune up chords” and “build up (houses)”) **pro-:** *prosmotret*' (“watch until the end”, “look through” and “overlook”) **za-:** *zažit’* (“heal a wound” and “begin to live”) **HOMOGRAPHS:** **Nouns agreeing with nouns:** *zamók* ("lock") and *zámok* ("castle") *muká* ("flour") and *múka* ("suffering") **Verbs coinciding graphically in infinitive forms and throughout the entire conjugation:** **aspect match:** *zapáxnut*' ("smell") and *zapaxnút*' ("wrap") (perfective) **no aspect match:** *srezát*' *(“to cut off”, imperfective)* and *srézat*' (“cut”, perfective) spešít’ (“to hurry”, imperfective) and spéšit’ (“dismount”, perfective) zasypát' ("fall asleep", imperfective) and zasýpat' ("fill in", perfective) *napadát*' ("attack", imperfective) and *napádat*' ("fall down", perfective) **POLYSEMANTIC WORDS:** *boltat*' ("stir" and "chatter") *ostrit*' ("sharpen" and "crack jokes")

APPENDIX B

Table 3: Results of the News Crawl: articles from 2008 analysis

	Zalizniak's dictionary (type frequency)	**News Crawl: articles from 2008 (token frequency)**
total number of participles	63540 (100%)	9019 (100%)
participle and adjective ambiguity	767 (1.2%)	4937 (54.73%)
participle and noun ambiguity	76 (0.1%)	1144 (12.6%)

Source for token frequency:

Name of Subpart	**file name**	**file size**	**words**
News Crawl: articles from 2008	*news.2008.ru.shuffled*	*7.5M*	*604,718*

Corpus is available at: http://www.statmt.org/wmt15/translation-task.html#download

Source for type frequency:

The normalizing generator (*generator-mt-apertium-norm.hfstol*, version 19.09.16), which produced wordforms and their morphological readings based on material from Zalizniak's dictionary (1977).

APPENDIX C

Source: Belousov et al. (1989)

Suffixes for full forms of participles			
active		**passive**	
present	**past**	**present**	**past**
-ущ-/-ющ- несущий поющий	*-вш-* писавший даривший	*-ем-/-ом-* изучаемый ведомый	*-нн-* званный избранный измотанный
-ащ-/-ящ- лежащий строящий	*-ш-* забредший отсветший умерший	*-им-* слышимый гонимый	*-енн-* увлечённый ушибленный ношенный
			-т- кинутый завернутый

Suffixes for short forms of participles	
present **passive**	**past** **passive**
-ен-/-н- обижен-а-о-ы нарисован-а-о-ы **-т-** бит-а-о-ы взят-а-о-ы	**-ем-/-ом-/-им-** читаем-а-о-ы ведом-а-о-ы терпим-а-о-ы

Forms of inflection in declension:

passive participles:
ый|ого|ому|ым|ом|ое|ого|ому|ая|ой|ую|ою|ые|ых|ым|ыми|

active participles:
ий|его|ему|ий|им|ем|ее|ая|ей|ую|ею|ие|их|ими

List of suffixes:
full form:
ущ
ющ
ащ
ящ
вш
ш
ем
ом
им
нн
енн
т

short forms:
ен
н
т
ем
ом
им

APPENDIX D

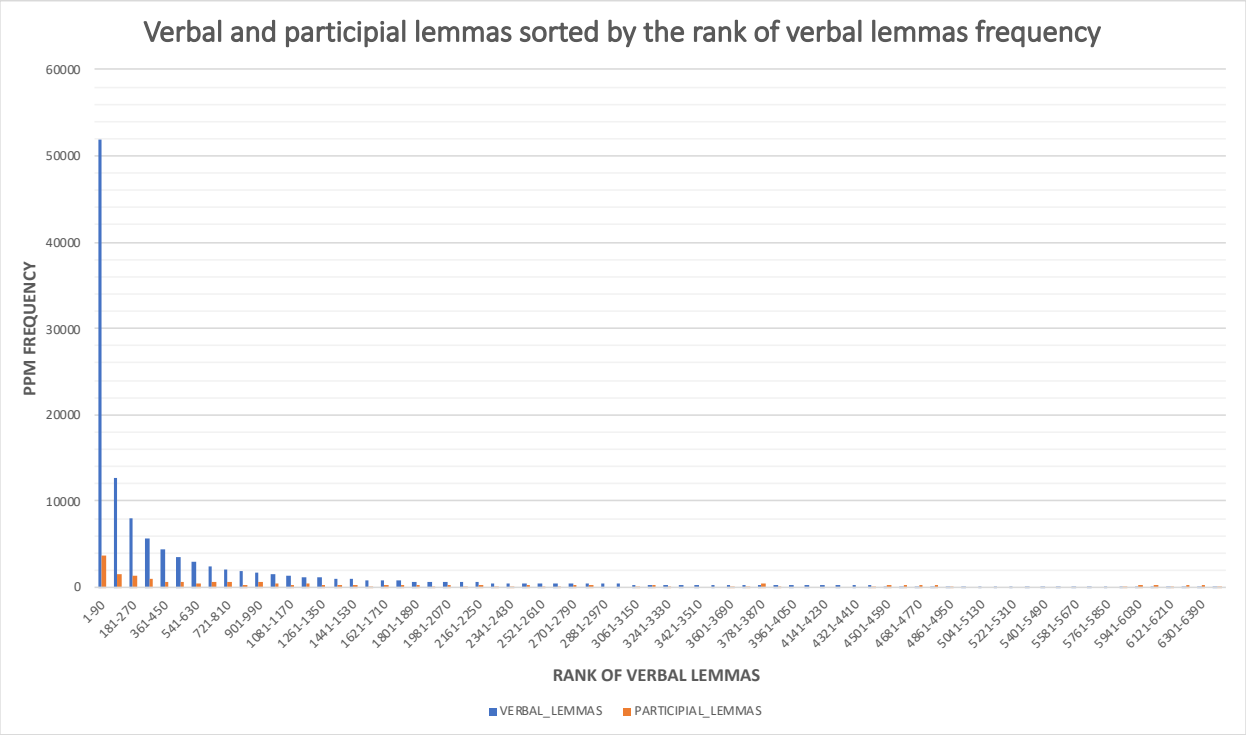

Figure 6: Distribution of verbal and participial lemmas sorted by the rank of verbal lemmas frequency

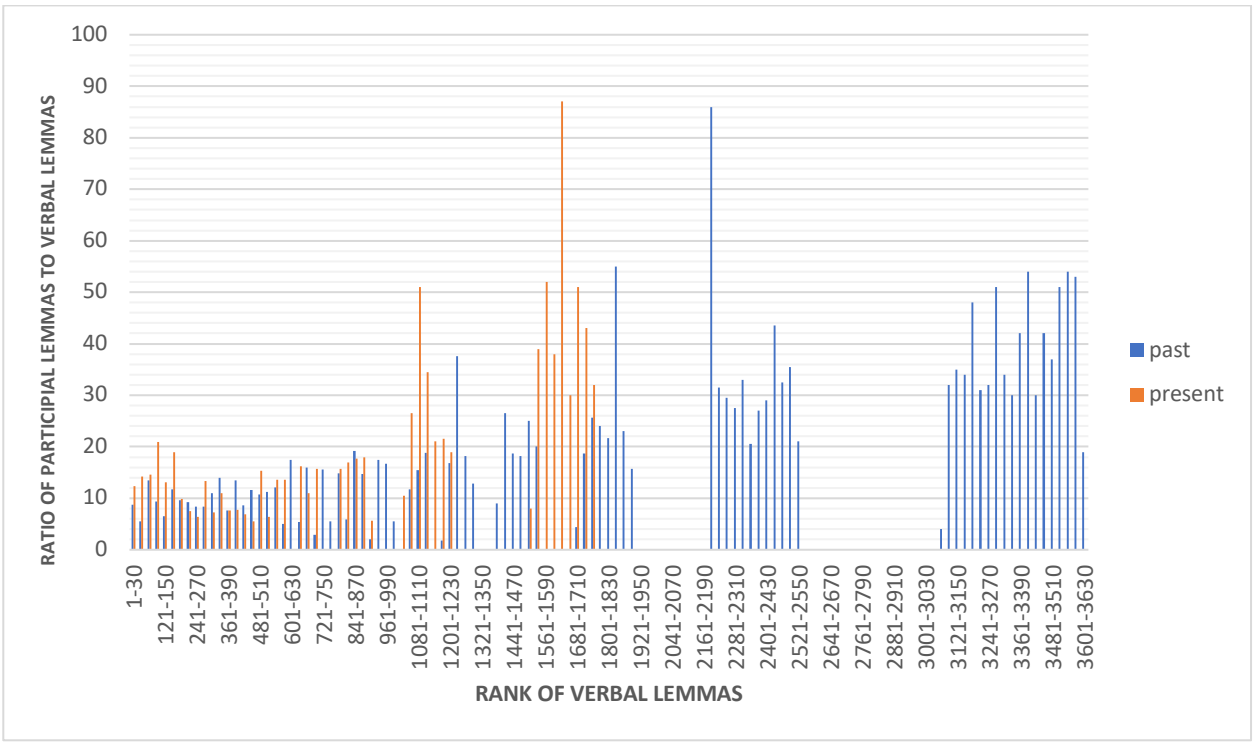

Figure 7: Ratio of past and present participial lemmas to verbal lemmas distributed over the rank of verbal lemmas

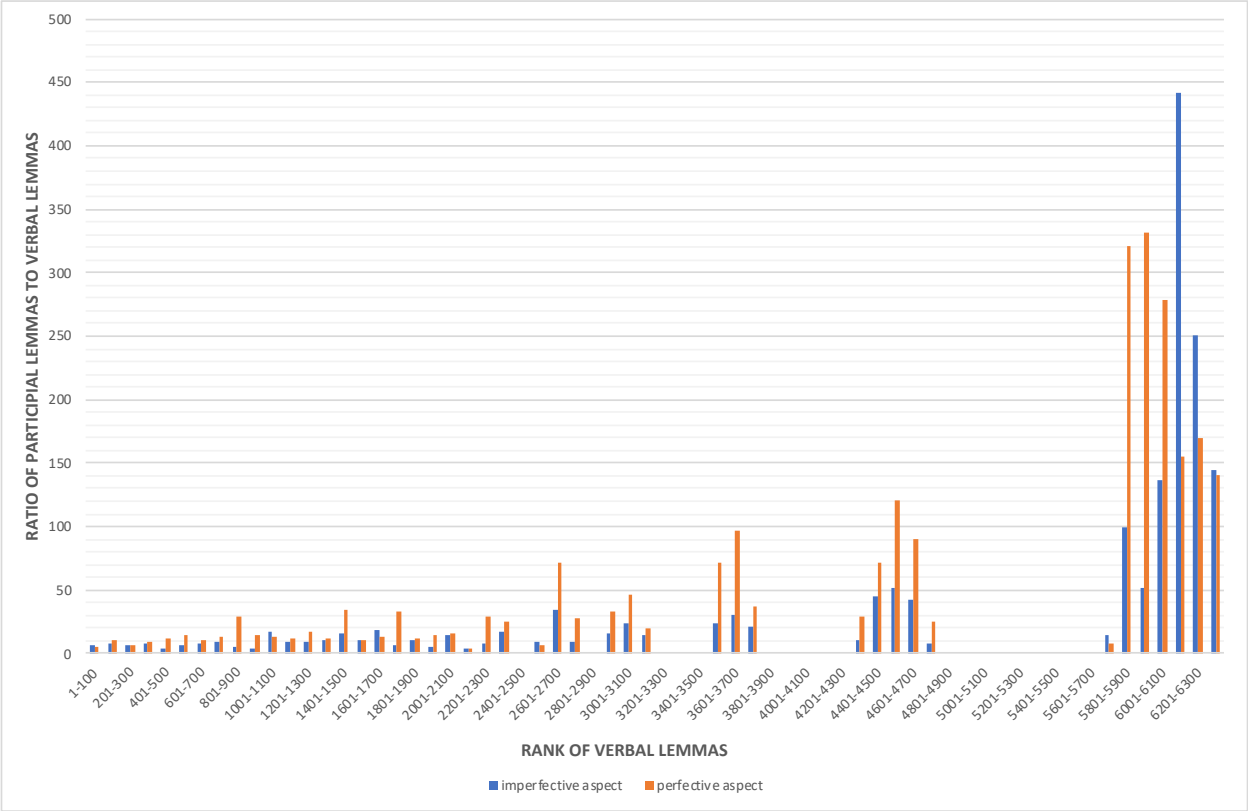

Figure 8: Ratio of imperfective and perfective participial lemmas to verbal lemmas distributed over the rank of verbal lemmas

APPENDIX E

RANK

```
> fit.rank = glm(ambig.rank$ambiguity ~ ambig.rank$rank + ambig.rank$ratio + ambig.rank$rank * ambig.rank$ratio,family=binomial())
> summary(fit.rank)

Call:
glm(formula = ambig.rank$ambiguity ~ ambig.rank$rank + ambig.rank$ratio +
    ambig.rank$rank * ambig.rank$ratio, family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0576  -0.3393  -0.2214  -0.2168   2.7354

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                             -3.739e+00  9.807e-02 -38.122  < 2e-16 ***
ambig.rank$rank4                         8.857e-01  1.189e-01   7.448 9.51e-14 ***
ambig.rank$rank5                         1.337e+00  1.872e-01   7.144 9.04e-13 ***
ambig.rank$rank6                         1.887e+00  5.210e-01   3.622 0.000292 ***
ambig.rank$rank7                        -9.374e+00  3.319e+02  -0.028 0.977466
ambig.rank$ratio                         4.247e-02  1.651e-02   2.573 0.010096 *
ambig.rank$rank4:ambig.rank$ratio        2.006e-01  4.425e-02   4.532 5.85e-06 ***
ambig.rank$rank5:ambig.rank$ratio        2.081e+00  3.845e-01   5.413 6.20e-08 ***
ambig.rank$rank6:ambig.rank$ratio        3.666e+00  2.206e+00   1.662 0.096452 .
ambig.rank$rank7:ambig.rank$ratio        2.123e+03  3.880e+04   0.055 0.956360
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3703.3  on 9584  degrees of freedom
Residual deviance: 3442.1  on 9575  degrees of freedom
AIC: 3462.1

Number of Fisher Scoring iterations: 11
```

TRANSITIVITY

```
> fit.trans = glm(ambig.trans$ambiguity ~ ambig.trans$transitivity + ambig.trans$aspect + ambig.trans$ratio + ambig.trans$rank + ambig.trans$transitivity*ambig.trans$rank + ambig.trans$aspect*ambig.trans$rank, family=binomial())
> summary(fit.trans)

Call:
glm(formula = ambig.trans$ambiguity ~ ambig.trans$transitivity +
    ambig.trans$aspect + ambig.trans$ratio + ambig.trans$rank +
    ambig.trans$transitivity * ambig.trans$rank + ambig.trans$aspect *
    ambig.trans$rank, family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.4298  -0.5515  -0.4406  -0.4020   2.2956

Coefficients: (2 not defined because of singularities)
                                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                                        -1.61494    0.06878 -23.481  < 2e-16 ***
ambig.trans$transitivitytran                       -0.66867    0.09118  -7.333 2.25e-13 ***
ambig.trans$aspectpf                               -0.19132    0.08864  -2.158   0.0309 *
ambig.trans$ratio                                   0.13875    0.01783   7.784 7.03e-15 ***
ambig.trans$rank4                                  -0.83284    0.11962  -6.962 3.35e-12 ***
ambig.trans$rank5                                   0.18480    0.19739   0.936   0.3491
ambig.trans$rank6                                 -12.42691  225.88007  -0.055   0.9561
ambig.trans$rank7                                  16.17942  882.74338   0.018   0.9854
ambig.trans$transitivitytran:ambig.trans$rank4      0.84401    0.14734   5.728 1.02e-08 ***
ambig.trans$transitivitytran:ambig.trans$rank5      0.57777    0.26208   2.205   0.0275 *
ambig.trans$transitivitytran:ambig.trans$rank6     14.00743  225.88091   0.062   0.9506
ambig.trans$transitivitytran:ambig.trans$rank7           NA         NA      NA       NA
ambig.trans$aspectpf:ambig.trans$rank4              0.07867    0.14629   0.538   0.5907
ambig.trans$aspectpf:ambig.trans$rank5             -0.50101    0.27752  -1.805   0.0710 .
ambig.trans$aspectpf:ambig.trans$rank6            -12.99887  249.80457  -0.052   0.9585
ambig.trans$aspectpf:ambig.trans$rank7                   NA         NA      NA       NA
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6427.8  on 9172  degrees of freedom
Residual deviance: 6227.8  on 9159  degrees of freedom
AIC: 6255.8

Number of Fisher Scoring iterations: 13
```

TENSE

```
> fit.tense = glm(ambig.tense$ambiguity ~ ambig.tense$tense + ambig.tense$ratio + ambig.tense$rank + ambig.tense$tense
+ ambig.tense$tense*ambig.tense$rank,  family=binomial())
> summary(fit.tense)

Call:
glm(formula = ambig.tense$ambiguity ~ ambig.tense$tense + ambig.tense$ratio +
    ambig.tense$rank + ambig.tense$tense + ambig.tense$tense *
    ambig.tense$rank, family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6229  -0.5336  -0.5026  -0.4690   2.2068

Coefficients: (1 not defined because of singularities)
                                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                                  -1.93987    0.07631 -25.422  < 2e-16 ***
ambig.tense$tensepraet                       -0.21200    0.09420  -2.251  0.02442 *
ambig.tense$ratio                             0.29362    0.03667   8.007 1.18e-15 ***
ambig.tense$rank4                            -0.40348    0.12873  -3.134  0.00172 **
ambig.tense$rank5                             0.78624    0.20210   3.890  0.00010 ***
ambig.tense$rank6                             0.82777    0.67100   1.234  0.21733
ambig.tense$rank7                            15.71049  535.41117   0.029  0.97659
ambig.tense$tensepraet:ambig.tense$rank4      0.60795    0.15538   3.913 9.13e-05 ***
ambig.tense$tensepraet:ambig.tense$rank5     -0.02900    0.27043  -0.107  0.91459
ambig.tense$tensepraet:ambig.tense$rank6    -12.27944  202.32648  -0.061  0.95161
ambig.tense$tensepraet:ambig.tense$rank7          NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6124.5  on 7956  degrees of freedom
Residual deviance: 5976.8  on 7947  degrees of freedom
AIC: 5996.8

Number of Fisher Scoring iterations: 12
```

VOICE

```
> fit.voice = glm(ambig.voice$ambiguity ~ ambig.voice$ratio + ambig.voice$rank + ambig.voice$voice +
ambig.voice$voice*ambig.voice$rank,  family=binomial())
> summary(fit.voice)

Call:
glm(formula = ambig.voice$ambiguity ~ ambig.voice$ratio + ambig.voice$rank +
    ambig.voice$voice + ambig.voice$voice * ambig.voice$rank,
    family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9804  -0.5563  -0.5339  -0.4457   2.1718

Coefficients: (2 not defined because of singularities)
                                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                                  -1.87628    0.05177 -36.241  < 2e-16 ***
ambig.voice$ratio                             0.16753    0.02780   6.026 1.68e-09 ***
```

```
ambig.voice$rank4                          -0.38279    0.08732  -4.384 1.17e-05 ***
ambig.voice$rank5                           0.53423    0.15409   3.467 0.000526 ***
ambig.voice$rank6                           0.48286    0.56140   0.860 0.389736
ambig.voice$rank7                          13.43938  196.96769   0.068 0.945601
ambig.voice$voicepass                      -0.07904    0.10158  -0.778 0.436501
ambig.voice$rank4:ambig.voice$voicepass     0.83940    0.15000   5.596 2.19e-08 ***
ambig.voice$rank5:ambig.voice$voicepass    -0.10626    0.42570  -0.250 0.802880
ambig.voice$rank6:ambig.voice$voicepass         NA         NA      NA       NA
ambig.voice$rank7:ambig.voice$voicepass         NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6270.4  on 7800  degrees of freedom
Residual deviance: 6135.7  on 7792  degrees of freedom
AIC: 6153.7

Number of Fisher Scoring iterations: 10
```