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

Uliana Petrunina, PhD student, University of Tromsø
CONTENTS

01. INTRODUCTION
(ambiguity in linguistics)

02. DISTRIBUTION ANALYSIS
(data analysis, corpus data)

03. DISAMBIGUATION
(constraint grammar formalism)

04. CONCLUSIONS
(+ further research)
INTRODUCTION
CONVERSION

- source for part of speech ambiguity (homonymy)
- an affixless means of derivation which is converting a lexeme from one part of speech into another.

Properties:
change in word-class/change in syntactic function (1)
no change in graphical form (1)
change in semantic properties (2)
universal process in many world-languages (3)

E.g.:
(1) Russian *bol'noj*:
   Adjectival function (modifier) *bol'noj- ADJ-M* škoľ nik ‘an ill schoolboy’,
   Nominal function (subject) *bol'noj-N-M* zasnul ‘an ill [person] fell asleep’

(2) Russian PTCP *povyšennyj*: PTCP ‘raised, increased’, ADJ ‘high’

ADJECTIVIZATION

ADJECTIVIZATION

a type of conversion:

• participial lexeme may retain some morphological verbal properties (aspect, tense, voice) but lose syntactic ones (e.g.: argument selection). It results in separation from the verbal paradigm and formation of an independent word with retention of the adjectival declension (Lopatin, Miloslavskij, Šelâkin 1989)

• gradual process of loss of semantic and syntactic properties (shift from verbal to adjectival paradigm) (Say 2006)

Examples:

1. No direktor ne spešil nanosiť rešajuščij-ADJ udar.
   ‘But the director was in no haste to strike a decisive blow’. (RNC)

1. [...] A. Karpov, mučiteľ no rešajuščij-PRES-ACTV-PTCP dlja sebja vyšеoznаčennyе voprosy, vygljadit étakim prekrasnym idealistom bezvozvratno ušedšix vremen, žertвуjuščim kař eroj vo имja idei česti.
   ‘[...] A. Karpov, painfully solving-PRES-ACTV-PTCP for himself the stated questions above, looks like a beautiful idealist of the past times, sacrificing his career for the sake of the idea of honesty’. (RNC)
Rysakov & Klyshinsky 2015:

- Homonyms in NLP amount to $\frac{1}{2}$ of all the words in Russian texts
- Over 50% of words in the syntactically marked Russian National Corpus (RNC, http://www.ruscorpora.ru/instruction-syntax.html) are ambiguous
- Much of this ambiguity results from multiple morphosyntactic interpretations
- Part of speech ambiguity amounts to 24.10%

Webcrawl 2008 corpus, 604718 wordforms

<table>
<thead>
<tr>
<th>Total number of homonyms</th>
<th>Homonymous participles/adjectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.72%</td>
<td>31.47%</td>
</tr>
</tbody>
</table>
OBJECTIVES

**word level**

Observe the effect of **CORPUS FREQUENCY, TRANSITIVITY, VOICE AND TENSE**

- What patterns can be derived from the observations?
- How do they correspond to the general knowledge about participles?

**context level**

Implement and evaluate disambiguation of ambiguous participles using **CONTEXT BASED RULES, CORPUS FREQUENCY**

- What kind of morphosyntactic rules should describe the context?
- How are they supposed to be ordered?
- How can corpus frequencies be implemented for disambiguation?
- How performative is the disambiguation based on corpus frequency?
WORD LEVEL

Adjectival/verbal properties + semantic criteria for adjectivization +

Morphological properties

- Verbal properties: transitivity, voice, tense
- Adjectival properties: case, gender, number

Semantic properties

- Idiomatic augmentation (shift in lexical meaning)

Ty čelovek obespečennyj-ADJ, podberem čego ni tam...
You are a wealthy-ADJ man, we can try to sort it out...(RNC)

Distribution analysis
Morphosyntactic criteria for adjectivization

- no syntactic parallelism
- ability for relativization and transformation into finite clauses
- compatibility with adverbs of measure and degree
- compatibility with comparative/superlative degree
- word order: preference for preposition rather than postposition
- localization in time and space (adverbs of time and place)
- loss of ability to join verb dependents (direct and indirect objects)

Disambiguation using constraint grammar (CG) formalism
02

DISTRIBUTION

ANALYSIS
ANALYSIS + DATA

Distribution 1:
wordforms with participial suffixes:
-ущ/-иющ - ACT PRES
-ащ/-ящ - ACT PRES
-вш - ACT PAST
-ш - ACT PAST

Distribution 2:
wordforms with RNC tags:
manually annotated (transitivity, voice, tense)

Statistical analysis

- ✔️ Manually disambiguated sample of Russian National Corpus: 1,190,540 wordforms
- ▶️ List of participial suffixes
- 🔄 Ratio of participial to verbal lemmas
  Part per million (ppm) frequency
  Rank values on the scale from 1 to 7.
- ✗ transtivity (transitive, intransitive)
  voice (active, passive)
  tense (past, present)
  ambiguity (ambiguous, unambiguous)
**DATA PREPARATION FOR DISTRIBUTION 1, 2**

01. Extraction of all wordforms

02. Annotation of the wordforms

03. Lemmatization of participial lemmas
   
   Verbal lemmas

04. Sorting of participial and verbal lemmas by rank

**Dist1:** Extracted ambiguous participial and adjectival forms were lemmatized.

**Dist2:** Extracted verbal lemmas with transitivity, voice, tense tags.

For each verbal lemma find a participial lemma, add ppm frequency of verbal lemma and participial lemma.
Verbal lemmas are 25.9 times (per subinterval) more frequent than the participial lemmas

1 participial form (top 2 verbs)
moč’ “be able to do smth” => moguščega ACT-PRES “being able” (rare)
byt’ ”to be” => adjectivized form buduščij-ACT-PRES “future”

0 participial forms (limitations of verb semantics)
vremenit’ “delay”
teret’ “rub”
sprosit’ “ask a question”
našit’ “sew on”
smoč’ “manage”

change of semantics
gоворит’ “speak” => говорjašča familija “self-explanatory surname”
sledovat’ “follow” => sledujuščij god “next year”
ponimat’ “understand” => ponimajuščaja ularyka “understanding smile”
Verbal lemmas are 1.2 times more frequent than participial lemmas

Adjectivized and idiomatized forms:
- *potrjasat’* “astonish” => *potrjasajuščix glazax* “gorgeous eyes” (metonymy)
- *osmyslit’* “apprehend” => *osmyslennaja politika* “prudent policy”
Verbal lemmas are about 0.4 times at average more frequent than participial lemmas.

`avtomatizirovat` “automate” =>
`avtomatizirovannogo`- PASS-PTCP kontrolja
“automated control”

vs.

`černet` “turn black” =>
`černejuščie`- ACT-PTCP pri žarenii “tuning black when fried”
Transitive verbs form much more participles than intransitive ones

=>

transitive verbs both active and passive forms while intransitive verbs can only have active forms
Ratio for passive voice is higher than ratio for active voice

=>
- highly frequent verbal lemmas have much less participial lemmas than mid- and low-frequency lemmas
- they have more transitive than intransitive participial lemmas,
- transitive lemmas can form active/passive lemmas while intransitive lemmas can only form active ones
LOGISTIC REGRESSION

ppm frequencies => values on the scale from 1 to 7 (rank 3, 4, 5, 6, 7)
ambiguity => ambiguous lemmas (adj_pp, adj_n_pp, pp_n) vs. unambiguous lemmas (pp) (0,1)

*(glm function in R): generalized linear model

RANK:
• rank 4, 5

TENSE:
• past tense
• rank 4, 5
• past tense + rank 4

TRANSITIVITY:
• transitive verbs
• transitive verbs + rank 4, 5

VOICE:
• rank 4, 5
• rank 4 + passive voice
**Table 1. The Zipf scale of word frequency**

<table>
<thead>
<tr>
<th>Zipf value</th>
<th>$f_{pmw}$</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.01</td>
<td>antifungal, bioengineering, farsighted, harelip, proofread</td>
</tr>
<tr>
<td>2</td>
<td>0.1</td>
<td>airstream, doorkeeper, neckwear, outsized, sunshade</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>beanstalk, cornerstone, dumpling, insatiable, perpetrator</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>dirt, fantasy, muffin, offensive, transition, widespread</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>basically, bedroom, drive, issues, period, spot, worse</td>
</tr>
<tr>
<td>6</td>
<td>1000</td>
<td>day, great, other, should, something, work, years</td>
</tr>
<tr>
<td>7</td>
<td>10,000</td>
<td>and, for, have, I, on, the, this, that, you</td>
</tr>
</tbody>
</table>

*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. $f_{pmw} =$ frequency per million words.*

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

Distribution 1:
- Proportion of participial to verbal lemmas increases as the rank of the verbs decreases

Distribution 2:
- Transitive verbal lemmas forms more participles than intransitive ones
- Highly-frequent active verbal lemmas have more adjectivized participles than low-frequency active lemmas
- Most low-frequency passive participles tend to be adjectivized

Statistics:
- rank 4 and 5 are significant (high, medium frequency)
- transitive verbs, past tense, passive voice

Mid- and low-frequency transitive verbs tend to form more ambiguous participial forms than intransitive ones. Verbs also tend to form past passive forms that often come to be ambiguous. Transitivity implies the presence of an argument
03

DISAMBIGUATION
Constraint Grammar formalism:

- rule-based approach to morphosyntactic analysis
- part of speech disambiguation depends on context
- weights: lexical unigram probabilities
- sequentiality of constraint grammar parsing:

CG rules using morphosyntactic context:

- REMOVE V IF (0 V OR A + $$NGDAIP)(1 N + $$NGDAIP OR Pron + $$NGDAIP LINK NOT 1 Pr);
  - Context 0: A or V agreeing in the same case
  - Context 1: a noun or a pronoun agreeing in the same case with A followed by anything but a preposition

- SELECT V IF (0 V)(0 A)(1 Ins);
  - Context 0: V or A
  - Context 1: any part of speech in Instrumental case

CG rule using weights:

- SELECT:maxweight (<W=MAX>);
WEIGHTING
Sharoff’s frequency list 2008:
• corpus of modern Russian
• 40 million words
• between 1970 and 2002
• lemmas/wordforms and frequencies for verbs, adjectives, nouns, prepositions, adverbs

SUBLEXICON FOR PARTICIPLES:
lemmas from participial wordforms with their weights

Zipf=log10((rawFreqSmooth/(corpusSizePm + typesNbPm)) + 3.0)
Van Heuven et al. (2014)
where:
rawFreqSmooth=rawFreq + 1: smoothed corpus frequency of a token (Laplace smoothing)
rawFrequency: raw corpus frequency of a token
corpusSize: total number of tokens in a corpus
corpusSizePm: number of tokens per million in a corpus, e.g.:
corpusSizePm=corpusSize/1000000,0
typesNb: total number of types in a corpus
typesNbPm: number of types per million in a corpus types, e.g.:
NbPm=typesNb/1000000,0

for ppm frequency of 0,01, weight = 1,8516799803768287
for ppm frequency of 29,56, weight = 3,271115574417968
Zipf = \log_{10}\left(\frac{\text{rawFreqSmooth}}{(\text{corpusSizePm} + \text{typesNbPm})}\right) + 3.0

weigted lexicon in adjectives.lexc

островской:остро́вск n_b?~ "weight: 3.329793225321471" ; ! Z 3b%[x]%~
шутовской:шуто́вск n_b?~ "weight: 3.184168788029196" ; ! Z 3b%[x]%~
складской:скла́дск n_b?~ "weight: 3.200861521635764" ; ! Z 3b%[x]%~
заводской:заво́дск n_b?~ "weight: 4.078358855309672" ; ! Z 3b%[x]%~
городской:горо́дск n_b?~ "weight: 4.936406937266904" ; ! Z 3b%[x]%~
людской:лю́дск n_b?~ "weight: 4.35422521203718" ; ! Z 3b%[x]%~
мужской:мужск n_b?~ "weight: 4.805525503791406" ; ! Z 3b%[x]%~
поварской:пова́рск n_b?~ "weight: 3.1067543982673835" ; ! Z 3b%[x]%~
мирской:ми́рск n_b?~ "weight: 3.5495102485703462" ; ! Z 3b%[x]%~
APPLICATION OF HELSINKI FINITE STATE TRANSUDUCER (HFST)

preprocessing: case conversion, sentence delimitation specification, etc.

compiling weighted lexicon (. lexc)

running morphological analysis
"<действующий>"
   "действовать" V Impf IV PrsAct Msc Anln Sg Nom <W:4.4485321045>
   "действующий" A Msc Anln Sg Nom <W:4.4408416748>
   "действующий" A Msc Inan Sg Acc <W:4.4408416748>
   "действовать" V Impf IV PrsAct Msc Inan Sg Acc <W:4.4485321045>

running disambiguation

"<действующий>"
   "действовать" V Impf IV PrsAct Msc Anln Sg Nom <W:4.4485321045> SELECT:1342:GCPtcpV7
   ;
   "действующий" A Msc Anln Sg Nom <W:4.4408416748> SELECT:1342:GCPtcpV7
   ;
   "действующий" A Msc Inan Sg Acc <W:4.4408416748> SELECT:1342:GCPtcpV7
   ;
   "действовать" V Impf IV PrsAct Msc Inan Sg Acc <W:4.4485321045> SELECT:1342:GCPtcpV7 REMOVE:905
   ;
CONSTRANINT GRAMMAR RULES

Specific context rules
REMOVE: SCPtcpV2 V IF (0 V) (0 A)(-1 MeasureDegreeAdv)(-1 Ne LINK -1 MeasureDegreeAdv)(NOT 1 Pron)(NOT 1 N)(NOT 1 Det)(NOT 1 Pron + Poss);

General context rules
SELECT: GCPtcpV1 V IF (0 V)(0 A)(1 Ins);

Weigting rules
SELECT: maxweight (<W=MAX>);

Specific context rules
причем процесс этот настолько-MeasureDegreeAdv продуман-V и автоматизирован, что даже не требует привозить на Гран При лишних людей.

General context rules
«— спрашивает Дмитриев, все более и более удивленный-V и испуганный-V словами-N-Ins солдата.

Weight rules
"<связаны>" ambt
"связать" V Perf TV PstPss Lxc MFN Pl Pred <W:5.1529998779> SELECT:1162:maxweight
; "связать" V Perf TV PstPss MFN Pl Pred <W:3.6455535889> REMOVE:1157
; "связанный" A MFN Pl Pred Fac <W:5.1062469482> SELECT:1162:maxweight
## EVALUATION

124 ambiguous forms from SynTagRus

<table>
<thead>
<tr>
<th>source</th>
<th>precision</th>
<th>recall</th>
<th>f-score</th>
<th>ambiguity solved (accuracy) %</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG rules</td>
<td>0,79</td>
<td>0,71</td>
<td>0,75</td>
<td>59,68</td>
</tr>
<tr>
<td>CG rules + weights</td>
<td>0,69</td>
<td>0,97</td>
<td>0,80</td>
<td>66,94</td>
</tr>
<tr>
<td>weights</td>
<td>0,52</td>
<td>0,97</td>
<td>0,68</td>
<td>51,61</td>
</tr>
<tr>
<td>SynTagRus</td>
<td>0,69</td>
<td>1,00</td>
<td>0,82</td>
<td>66,13</td>
</tr>
</tbody>
</table>

**CG rules**: highest precision (0,79) but the lowest recall (0,71)

**CG rules + weights**: the highest recall (0,97), the second highest f-score, the highest accuracy

**weights**: the lowest precision, the highest recall (0,97), the lowest f-score and accuracy

**SynTagRus**: the highest recall and f-score
CG rules that worked in most cases:

• remove adjectival reading if a wordform is followed by a prepositional phrase
• remove adjectival reading if a wordform has an argument in instrumental or dative case
• remove verbal reading if a wordform agrees in case with its argument
• weighting rule for adjectives
• adjectives being identified as participles:
  igrat‘-V matč-N protiv-PREP dejstvujuščego-ADJ čempiona-N
  “to play a match against the current-ADJ champion”

• nouns being identified as participles:
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”

• participles being identified as adjectives:
on_PRON vnužden-PTCP byl-V
  ‘he was-Vbyt‘ obliged-PTCP’
CORRECTLY ANALYZED INSTANCES

• verbal forms + prepositions:
  i-CC orientirovan-PTCP na-PREP praktičeskoe-ADJ spasenie-N usadeb-N
  “and oriented-PTCP towards-PREP the practical rescue of farms”

• present/past passive verbal form + Instrumental case:
  postavok-N, reguliruemyx-PTCP proizvoditeljami-N-Ins
  “supplies handled_PTCP by the manufacturers”

• the copula byt' ‘be’ + verbal form:
  bylo-Vbyt' prinjato-PTCP odobritel'no-ADV
  “[suggestion] was-Vbyt' taken-PTCP favorably”
SUCCESSFUL INSTANCES OF WEIGHT RULES

- selection of participial forms:
  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 adjectivized participles:
  zadannoj-ADJ formy-N
  “predefined shape”
  okružajuščuju-ADJ sredu-N
  “[surrounding] environment”

- copular constructions:
  byl-Vbyt' široko-ADV rasprostranen-PTCP
  “was-Vbyt' widely spread-PTCP”
CONCLUSIONS
CONCLUSION

DISTRIBUTION ANALYSIS
• high-frequent verbs form less participles than low frequency verbs
• proportion of passive and transitive participial lemmas is higher than present and intransitive ones
• relationship of ambiguity and rank 4, 5, past tense, passive and transitive lemmas is significant

CG DISAMBIGUATION
• Context-based rules capture context but may be erraneous in selecting appropriate contexts
• Weights do straightforward sorting which decreases the number of false negatives.
• Combining weigting rules with context rules containing information on voice, tense, transitivity may increase the level of resolved ambiguity
• Adding tags with semantic content is going to be the next important step in disambiguation and theoretical analysis of ajectivization
• Adding semantic properties of verbs will highlight more on their predisposition for forming ambiguous participles.
BIBLIOGRAPHY

THANK YOU!

uliana.petrunina@uit.no
Criteria for adjectivization. Disambiguation using VSG3 constraint grammar formalism

- **no syntactic parallelism**
  - ability for relativization and transformation into finite clauses

- **compatibility with adverbs of measure and degree**

- **compatibility with comparative/superlative degree**

- **word order: preference for preposition rather than postposition**

- **localization in time and space (adverbs of time and place)**

- **loss of ability to join verb dependents (direct and indirect objects)**

Na ètom ego blestjaščee-ADJ vystuplenie bylo okončeno...
At this point, his brilliant performance was over... (RNC)

*Na ètom ego vystuplenie, kotoroe blestelo-V-PAST, bylo okončeno... At this point, his performance that was shining was over... transformation is impossible

Tam v sovetskie vremena byla očen’ prognoziuromajja i u kogo-to daže normal’naja žizn’!
People there had a very predictable — and some of them even normal — life in Soviet times! (RNC)

Samye prodavaemymye produkty pitanija — èto produkty social’noj značimosti.‘
The best selling foods are products of social importance [...] (business-ideal.ru).

V vojere rezviloš tri medvedja, raskormlenyhx i otdoxnuvšix.
‘there were three well-fed and rested bears romping in the open-air cage’(Mel’čuk, 1988)[p.384] (preposition)

Molodye ljudi, kurjašcie tabak, vopreki obšceprinjatomu mneniju imejut do-
voľ no privlekateľ nyj vid [...] (postposition)

Young people smoking tobacco, contrary to popular belief, look quite attractive [...] (RNC).

Osobuju sklonnost’ izbegat’ samostojatel’nyj rešenij projavljajut v japonskom delovom mire
ljudi, toľko čto povyšennye v range.
A particular inclination to avoid independent decisions is shown in the Japanese business world by the people just forwarded in rank (RNC).

Potrjasajuščij voobraženie šedevr arhitektury stali nazyvať “Sadom desjati tysjačasov”.
The masterpiece of architecture, amazing imagination, became known as “The garden of ten thousand gardens” (RNC).
DISCOVERING AMBIGUITY IN CORPUS DATA

News Crawl
• Articles of 2008
• "Shared Task: Machine Translation"
• 604718 wordforms
• file news.2008.ru.shuffled
• File size 7.5 M

The total number of generated wordforms: 9257801
# AMOUNT OF AMBIGUITY

<table>
<thead>
<tr>
<th></th>
<th>Zalizniak (type frequency)</th>
<th>Webcrawl 2008 corpus (token frequency)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of participles</td>
<td>63540 (100%)</td>
<td>9019 (100%)</td>
</tr>
<tr>
<td>Ambiguity of participles and adjectives</td>
<td>767 (1,2%)</td>
<td>4937 (54,73%)</td>
</tr>
<tr>
<td>Ambiguity of participles and nouns</td>
<td>76 (0,1%)</td>
<td>1144 (12,6%)</td>
</tr>
<tr>
<td>Total number of homonyms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homonymous participles/adjectives</td>
<td>2,72%</td>
<td>31,47%</td>
</tr>
</tbody>
</table>
5: ambig intr > unambig intrans
4, 3: trans ambig > intrans ambig
DISTRIBUTION 3: + AMBIGUITY + VOICE

Distribution of ratio of participial to verbal lemmas across rank, voice, ambiguity

3, 4: unambig > ambig
pass ambig > act ambig
Ratio of participial to verbal lemmas

7, 6: ambig >= unambig
4, 3: pres ambig > pres ambig
TYPES OF PARTICIPLES

• **Unambiguous participles**
  (3). *prosjaščij o pomošči čelovek*
  the man *asking* for help.

• **Adjectivized participles**
  (4). [...] tak čto lučše pokleit’ *oboi mojuščiesja* na netkanoj baze s pokrytiem iz vinila.
  So, it is better to stick the *wallpaper [that is] washable* on the unwoven base made of vynil (RNC).

• **Deverbal adjectives ((un)ambiguous with participles)**
  (5). I srazu uvidel belye dikie gory i wysokij *koljučij les*.
  And, immediately, I saw the white wild mountains and the tall *bristly woods* (RNC).